SENTIMENT ANALYSIS BASED ON ONLINE WOMEN'S CLOTHING REVIEWS AND RATINGS USING MACHINE LEARNING APPROACHES

BY

MARZANA RAHMAN
ID: 191-15-12198

AND

RAIHANA RAHMAN
ID: 191-15-12200

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

ISRAT JAHAN
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

MR. MD. AZIZUL HAKIM
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project/internship titled **"Sentiment Analysis Based on Online Women's Clothing Reviews and Ratings Using Machine Learning Approaches"**, submitted by Marzana Rahman, Raihana Rahman, ID No: 191-15-12198, 191-15-12200 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *23 January,2023*.

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Zahid Hasan
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Fahad Faisal
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Ahmed Wasif Reza
Associate Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Israt Jahan, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

____

Israt Jahan
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

____

Mr. Md. Azizul Hakim
Sr. Lecturer
Department of CSE
Daffodil International University

Submitted by:

Marzana

____

Marzana Rahman
ID: 191-15-12198
Department of CSE
Daffodil International University

Raihana

____

Raihana Rahman
ID: 191-15-12200
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratitude to Almighty God for His divine blessing, which made it possible for us to successfully complete the final year project.

We are grateful and wish to express our profound indebtedness to **Israt Jahan, Lecturer,** Department of CSE, Daffodil International University, Dhaka. deep knowledge and the keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartfelt gratitude to **Professor Dr. Touhid Bhuiyan**, Professor and Head of the Department of CSE, for his kind help in finishing our project and to other faculty members and the staff of the CSE department at Daffodil International University.

We would like to thank our entire class at Daffodil International University, who participated in this discussion while completing the course work.

Finally, we must acknowledge with respect the constant support and patience of our parents.

ABSTRACT

Analyzing how customers to act is essential for e-commerce marketing strategies because e-commerce can significantly boost economic growth. In natural language processing (NLP), techniques like sentiment analysis are used to determine whether data is positive, negative, or neutral. Using the user reviews in our database, we can build a machine-learning model that gives the right sentiment for each product. In addition to helping customers understand the product better, an accurate sentiment research also helps the business gain better market feedback. In this study, we perform sentiment analysis on a data set from online reviews of women's clothes downloaded from Kaggle. Three well-known machine learning algorithms were used to tackle the issue: logistic regression, Naive Bayes classifiers, and Support Vector Machine (SVM). The best results came from the LR algorithm, which had the best AUC value and accuracy.

TABLE OF CONTENTS

CONTENTS **PAGE**

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Short Form	Full Form
NLP	Natural Language Processing
ML	Machine Learning
TF-IDF	Term frequency-inverse document frequency
SVM	Support Vector Machine
NB	Naive Bayes
LR	Logistic Regression
NLTK	Natural Language Toolkit

CHAPTER 1
INTRODUCTION

1.1 Introduction

Currently, businesses are opting for internet strategies to expand their consumer base, and product reviews are essential for businesses to comprehend their customers' demands and customize their products accordingly. The review and analysis of data help achieve this goal. Text classification is one of the most severe problems in the field. By categorizing the reviews, we can perform today's most crucial task: sentiment. Sentiment analysis (a computational method that uses statistics and natural language processing to categorize the opinions expressed in the text) can assist businesses in enhancing their product or service and providing a better customer experience. In this article, we will try to comprehend consumers' feelings based on the evaluations they leave on e-commerce websites. Our goal is to facilitate the creation of a robust and reliable recommendation system for customers, which will ultimately increase the profitability of respective companies.

1.2 Motivation

Sentiment analysis may be used to analyze the overall feeling of a textual item. To provide just one example to clarify how it works, let's say a customer intends to go shopping for her own. She searched online. Numerous comments have been shown. One says the material is comfortable, and the size is fine for me, but some parts are gruesome; others say it makes you look at least ten years older when you wear it, and still, others say you look the most remarkable person on the street when you wear it outside. All the review texts might be overwhelming for a first-time buyer. People may lose interest in the product if they have to read too many reviews, and they may decide against purchasing it altogether.

On the other hand, sentiment analysis can give direct suggestions to customers, telling them whether the product is good, bad, or average. But fake reviews can confuse buyers and lower the product's overall Rating, leading to fewer sales. Therefore, it is essential to do a sentiment analysis of customer feedback. This study aims to improve how well online reviews of women's clothing are put into groups.

1.3 Rationale of the Study

The purpose of this project is to assess whether the product is recommended or not by analyzing sentiment. As an input text, we have utilized a review. We utilized various machine learning methods to make more precise predictions. Text categorization is based on categorizing documents into one or more groups. Sentiment analysis is one of the most valuable ways to sort texts. It determines the user's perspective on a product, topic, or service. User reviews play an essential part in the e-commerce business. The product's quality can be measured based on consumer reviews. A new customer can determine whether or not to purchase a product based on previous reviews.

1.4 Research Questions

The following query will come up as we work on our dataset. We have tried to solve this question by preprocessing and analyzing our dataset. Because a cleaned dataset will give reasonable accuracy and help us to solve this question.

1. Which machine learning method could make it easier to sort the feelings expressed in online reviews of women's clothing?
2. What is automated sentiment extraction and analysis?
3. What would it solve?
4. How can sentiments be automatically extracted and analyzed?
5. How can sentiment analysis be used in predicting financial markets?
6. How can I make my sentiment analysis better?

Solving Questions

1. For sentiment analysis categorization, Naive Bayes is a relatively straightforward category of probabilistic algorithms that determines the likelihood that a particular word or phrase should be categorized as positive or negative. but, in our paper, we have used Logistic regression classifier from which we get better accuracy than other algorithms.

2. Machine learning (ML) techniques are used in automated sentiment analysis. In these circumstances, ML algorithm is trained in this case to classify sentiment based on both the words and their order. The quality of the training data set and the algorithm are critical to the success of this approach. Using qualitative feedback, sentiment analysis assists brands in learning more about customer

perception. Businesses can learn how their customers truly feel about their products, services, marketing campaigns, and more by utilizing an automated system to analyze text-based conversations. In our study, sentiment analysis determines whether or not a product is recommended. We work on sentiment analysis, which can provide customers with direct recommendations, such as whether the product is good, bad, or average. Purchasing products will save time and money if the company has a positive rating rather than a negative one.

3. In our research paper, sentiment analysis determines whether the product is recommended or not. We work on Sentiment analysis which can give direct suggestions to customers, telling them whether the product is good, bad, or average. It will save time and money to purchase products if the business has a positive rating rather than a negative one.

4. There are numerous sources from which relevant sentiment analysis data can be gathered for sentiment analysis. News, public information, social media, customer reviews, customer service, and many more are examples. The most common method for calculating a sentiment score is to use a dictionary of negative, neutral, or positive words. The text is then examined to determine how many negative and positive words are present. This can provide us with a good sense of the overall tone of the text. The working principle followed some sequence:
    - First collecting dataset
    - Cleaned dataset
    - Data analysis, whether the data is balanced or imbalanced.
    - Preprocessing the dataset perfectly
    - Feature extraction
    - Data visualization
    - Applying algorithms for classification and showing the polarity based on class and rating gives accurate results to help customers buy good products quickly.

5. Investor sentiment and market prices are linked in such a way that when these opinions include positive sentiments, business stock prices tend to rise. Examining investor sentiment using sentiment analysis methods can thus provide useful insights into the stock market's future.

6. The steps below must be followed in order to train a custom sentiment analysis model:
   - Gather unstructured raw data for sentiment analysis.
   - Text preparation.
   - Text that is encoded numerically.
   - Selecting the proper ML algorithm.
   - ML model training.
   - Prediction.

1.5 Expected Output

From Table 4.2.2, We obtain the achieved accuracy after classifying a dataset. We found the SVM and logistic regression accuracy to be 78.7% and 78.9%, respectively. For Naive Bayes, we got the result of 77.0% accuracy. On a total of 23486 records, we applied all the classifications. Both logistic regression and SVM have a similar precision value of 92.3%. Logistic regression has the highest F1, recall, and AUC scores. The AUC score of SVM is quite similar to that of logistic regression.

1.6 Project Management and Finance

"Sentiment Analysis of Customer Reviews Using Machine Learning and NLP Based on Online Women's Clothing Reviews and Ratings" is a research-based project. Here, we found a way for e-commerce websites to analyze the customer's sentiments based on their review. We have used various techniques and algorithms to find the optimal result. We didn't have to do our project on a larger scale while we worked on finding the best way to determine how customers felt based on what they wrote in reviews. We obtained the necessary datasets from Kaggle. So, we didn't have to start our work from scratch, which saved us much time. And as our project is software-based, and we ran our project on our machine, we didn't have to carry any extra cost.

1.7 Report Layout

We have included six chapters in total in this study. This six-chapter overview is provided below:

- In the first chapter, we gave a general overview of the paper by listing the introduction, motivation, rationale of the research, project management and financing, report structure, and projected results.
- We examined the project's history in the second chapter, including preliminary work, related works, comparative studies, and challenges.
- Data collection, proposed method, statistical description, and implementation Requirements are all shown in depth in Chapter Three.
- Chapter four discusses the experimental analysis, setup, and results.
- Chapter five examined the impact on society, the Environment, and ethical aspects in detail.
- Chapter six examined the study's summary, conclusion, and future scope.

CHAPTER 2
BACKGROUND

2.1 Preliminaries/Terminologies

The main goal of this study is to ascertain whether consumers—especially those who are assumed to be women—would recommend the product they bought based on the details in their reviews. In particular, it should emphasize that this project's goal is to ignore the other variables and use only the "Review Text" variable. There are 23486 rows of data in total, including 11 column variables. A written comment and extra client data are present in each row. Additionally, each row contains the variables and corresponds to customer feedback. In the review text and body, references to the corporation have been changed to "retailer" because this is actual commercial data.

2.2 Related Works

The Internet has made online shopping increasingly popular. People want to share their product experiences and recommendations online after purchases. More researchers are looking at how people feel about these evaluations to make better classifications or predictions. People want to discuss their product experiences on the network after purchasing and then make recommendations to other buyers. In the meantime, to perform sentiment analysis and provide more precise classifications or predictions, more researchers are concentrating on these evaluations. In this section, we review earlier research articles. Several recent studies have revealed how online reviews affect decisions, especially about what to buy. We'll discuss prior research publications in this area:

Peter Turney's study [1]on classifying reviews used an unsupervised method, which is done by figuring out the word's meaning. For this, only phrases with adjectives or adverbs are taken into account. A phrase can make you feel good or bad, depending on how it is connected to. The review is either positive or negative based on how it makes you feel. The research is based on phrases, and with 75% accuracy, it can classify reviews.[2] In this work, the author first tries to evaluate the non-text review functions, such as age, costume class, and so on. Then, he tries to figure out how they relate to the suggested product. He tries to build a two-way recursive neural network with long-term memory to decide, based on the review, whether or not to buy the product.

[3]Agar ap, A. F., and P. M. Grafilon (2018) suggested using a recurrent neural network (RNN) with long-short term memory (LSTM) and sentiment analysis to figure out if it is recommended or not. F1 scores of 0.88 and 0.93 were observed for recommendation and sentiment classification, respectively.[4] The task is to determine from the review text whether a piece of writing is good, bad, or neutral. Liu B (2012) Based on the scope of the text, it can break down sentiment polarity into three levels: the document, sentence, and entity/aspect levels.[5] Khan et al. (2019) also made a method that involved getting data, cleaning it up, and pulling out features. Consumer sentiment was classified using SVM, NB, and Decision Tree (DT) algorithms. Researchers, service providers, and policymakers all benefited from the framework. When evaluating the framework with the airline's data set, the findings showed that SVM was as accurate as 90.3%, which is much more accurate than other technologies.[6] Gamon et al. (2005) devised a way to use customer reviews for topic mining and sentiment analysis. The main goal was to use clustering techniques like k-means, entropy-based, and n-gram feature vectors to find the keywords in the phrase that best represented the customer's feelings. This method allows customers to quickly and easily find what they're looking for in a mountain of content.[7] Adekitan and Noma-Osaghae have used machine learning algorithms in their work to make predictions about how well university students will do.[8] Multiclass categorization was used by Bo Pang et al. This technique finds the overall sentiment of a document, not a single sentence or phrase. This study considers product star ratings on e-commerce websites. The reviews are first put into groups based on their ratings, then the whole document is put into groups using the multiclass method.[9] Jiani Zhang and his colleagues devised a new way to analyze sentiment. They used a hierarchical network to classify how people felt about different things. The goal of this module is to figure out both the category of emotion and its polarity. The text input into the network was processed using recurrent neural networks. [10]Twitter collects tweets from eight western and eastern countries. Twitter is rich with opinions and perspectives from diverse countries, languages, and perceptions.

2.3 Comparative Analysis and Summary

In[11], they worked on sentiment analysis through opinion extraction. In their study, they used machine learning algorithms to sort the reviews into those that were positive and those that were negative. They use linear SVC, logistic regression, and decision tree algorithms to test the accuracy, and among the three classifications, logistic regression showed the highest accuracy of 87%. We discovered that logistic regression had the highest accuracy, at 79%.[12] They illustrate the taxonomy of sentiment analysis techniques. They looked at data from online shopping sites and found that the SVM method is the most accurate (98.2%) compared to the Naive Bayes and maximal entropy methods. In[13], they collect the dataset from Amazon, which contains electronic accessories. They used SVM and Naive Bayes to sort the best product reviews and found that Naive Bayes was the most accurate, with a 98.17% accuracy rate. We worked on 23486 and found the best accuracy of logistic regression, which is 79%. In this paper[14], they collect customer feedback data from the eBay app store, which has 50,000 reviews. Their study shows that SVM performs the best, whereas our research shows the best result of logistic regression on 23486 datasets. [15]They compiled a dataset of over 35 million Amazon.com product reviews. Linear SVC shows the best accuracy of 88.11%, and we got the best accuracy of logistic regression of 79% on the 23468 datasets. The following Table 2.3.1 shows the comparison of related works.

Table 2.3.1 Compare with related works

Paper Name	**Classification Techniques**	**Accuracy**
Our Works		
Sentiment analysis based on online women's clothing reviews and ratings using machine learning approaches	Logistic Regression Naïve Bayes SVM	79% 75% 75%
Other Works		
Sentiment Analysis of Product Reviews: Opinion Extraction and Display [11]	Linear SVC Logistic Regression Decision Tree	86% 87% 73%
Sentiment analysis of product reviews [12]	SVM Naïve Bayes Maximal Entropy	98.2% 94.1% 95.6%
Sentiment Analysis on Product Reviews Using Machine Learning Techniques [13]	SVM Naïve Bayes	97.7% 98.17%
Sentiment Analysis of Product-Based Reviews Using Machine Learning Approaches [15]	Linear SVC	88.11%

2.4 Scope of the Problem

- In our study, it was simple to distinguish between real and fake reviews, allowing us to provide users with accurate results.
- This paper uses review to assess, apply, and analyze the most successful sentiment analysis methodologies.
- We can easily detect positive, negative, or neutral reviews from textual data from any online shop.

2.5 Challenges

We obtained our dataset from Kaggle. This dataset's data was unbalanced. As a result, we must rebalance our dataset using the SMOTE-ENN model. This study will use technologies for natural language processing (NLP) to find broad trends in what customers write. This project aims to determine whether customers would recommend the product based on the details in their review text. One of the more difficult parts of this project is using text mining algorithms to find helpful information in the "Review Text" variable. We also need to turn text files into numeric feature vectors to run machine learning algorithms.

- Collect a suitable dataset for our project.
- While selecting the classification method, we faced some problems.
- When we implemented those classification methods in Python, we encountered problems because running models took a long time.
- Using text mining methods to extract meaningful information from the "Review Text" variable is one of the difficulties of this project.
- For machine learning algorithms to work, they must turn text data into numeric feature vectors.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

Our research topic is “Sentiment analysis of customer reviews using machine learning based on women's clothing reviews and ratings."

Instruments:

In this project, we have built models that provide better results in terms of accuracy for sentiment analysis on women's clothing reviews. The following are the devices, tools, and programs that we used:

Hardware Requirements:

- Core i5/i7 processor
- Ram size 16 GB.
- Usable Hard Disk Space

Software Requirements:

- python version (3.10.4)
- Google Collaboratory Distribution
- NLTK Toolkit
- Operating system: Windows 10 (64-bit).
- Internet connection.

3.2 Data Collection Procedure/Dataset Utilized

Kaggle ("https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews") provides the data used in this analysis. Its feedback is collected from women's clothing e-commerce site visitors. This set of data consists of 23486 rows and 11 columns in total. Each row that reflects a customer review has the following data:

- **Clothing ID:** Integer Categorical variable that refers to the specific piece being reviewed.
- **Age:** Indicates the reviewer's age as a positive integer variable
- **Title:** The review's title is a string variable.
- **Review Text:** String variable for the substance of the review**.**

- **Rating:** Positive Ordinal Integer variable for the customer's product rating, ranging from 1 Worst to 5 Best.
- **Recommended IND:** A binary variable that indicates whether or not the consumer recommends the product, where 1 indicates that the client recommends the product and 0 indicates that the customer does not recommend the product.
- **Positive Feedback Count:** Positive Integer showing how many customers found this review positive.
- **Division Name:** The high-level product division's categorical name.
  Department Name: Categorical name of the product department name.
- **Class Name:** Categorical name of the product class name.

3.3 Statistical Analysis

In the description and statistical sections, we found the values of mean, count, and standard deviation. The mean value refers to the average of all the integers included data set and is determined by the following formula.

Mean= $\frac{sum\ of\ all\ data\ values}{number\ of\ data\ values}$

Symbolically,

$$\bar{\boldsymbol{x}} = \frac{\boldsymbol{\Sigma x}}{\boldsymbol{n}} \quad ;$$

Where,

$\bar{x}$= mean of the set of x values

Σx =sum of all the x values

n =number of x values

Standard Deviation:

It is a method for measuring the deviation of observations from the mean. Sigma (σ) represents this quantity, which is the square root of the variance. The standard deviation is expressed in the same unit as the dataset's values; therefore, it indicates how far the observations in the dataset deviate from the mean.

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum(\boldsymbol{x_i} - \boldsymbol{\mu})^2}{\boldsymbol{N}}}$$

σ =population standard deviation

N =the size of the population

x_i =each value from the population

μ =the population mean

Figure 3.1 describes the mean, count std, min, max-age value, Rating recommended, and positive feedback column.

```
df.describe()
```

	Age	Rating	Recommended	PositiveFeedback
count	22641.000000	22641.000000	22641.000000	22641.000000
mean	43.280376	4.183561	0.818868	2.630582
std	12.326980	1.115762	0.385136	5.786164
min	18.000000	1.000000	0.000000	0.000000
25%	34.000000	4.000000	1.000000	0.000000
50%	41.000000	5.000000	1.000000	1.000000
75%	52.000000	5.000000	1.000000	3.000000
max	99.000000	5.000000	1.000000	122.000000

Figure 3.1 Data Description

Figure 3.2 shows that the count has a positive polarity of 21293.0, a mean value of 0.262335, and a standard deviation of 0.150223.

	sentiment	0	1
Age	count	1.322000e+03	21303.000000
	mean	4.308850e+01	43.296062
	std	1.197634e+01	12.350075
	min	2.000000e+01	18.000000
	25%	3.425000e+01	34.000000
	50%	4.100000e+01	41.000000
	75%	5.100000e+01	52.000000
	max	9.400000e+01	99.000000
Rating	count	1.322000e+03	21303.000000
	mean	2.919062e+00	4.261419
	std	1.349107e+00	1.050932
	min	1.000000e+00	1.000000
	25%	2.000000e+00	4.000000
	50%	3.000000e+00	5.000000
	75%	4.000000e+00	5.000000
	max	5.000000e+00	5.000000
sitiveFeedback	count	1.322000e+03	21303.000000
	mean	3.177761e+00	2.598273
	std	6.810800e+00	5.716785

Figure 3.2 Statistical Analysis

3.4 Proposed Methodology

The first step of sentiment analysis is to take the information given and pull out the relevant data. When the end goal is attained, the process is complete, will follow the next steps during the process:

Step 1: In this step, we import all the libraries which are needed to perform the dataset

Step 2: In this step, we load the entire dataset we downloaded from Kaggle.

Step 3: In this step, we do the exploratory data analysis such as dropping unnecessary columns like "Unnamed," "clothing ID," and "Title," and remove null values from the review text and checking for duplicate columns that exist in the dataset and remove all the duplicate columns for the cleaning process.

Step 4: In this step, we preprocess the review text removing punctuation, lemma and stop words, whitespaces, digits and special symbols to extract the clean review text.

Step 5: In this step, we normalize the word in the corpus in the cleaned review text, and then visualize it. we will classify the sentiments into Create three categories—positive, negative, and neutral—and use word clouds to represent the data for each one in individual categories.

Step 6: In this step, we perform feature extraction and rebalance the data before creating the vectorized training and test data, then split the data into train and test.

Step 7: In this step, we build a model to perform three-classification for prediction accuracy.

Figure 3.3 illustrates the process of our proposed work.

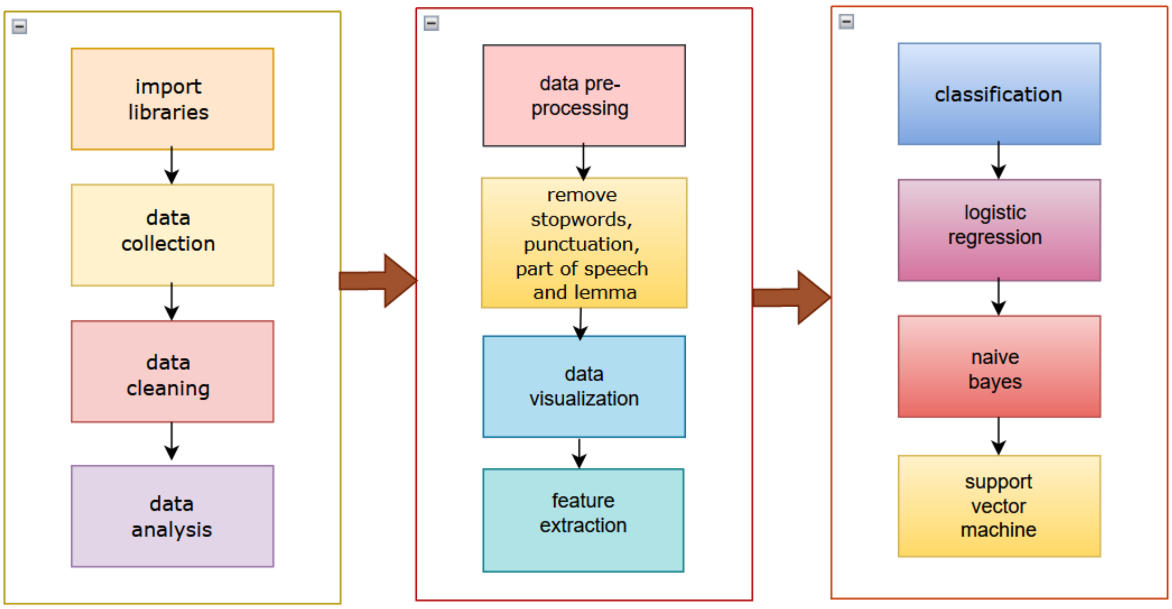


Figure 1.3 Proposed work

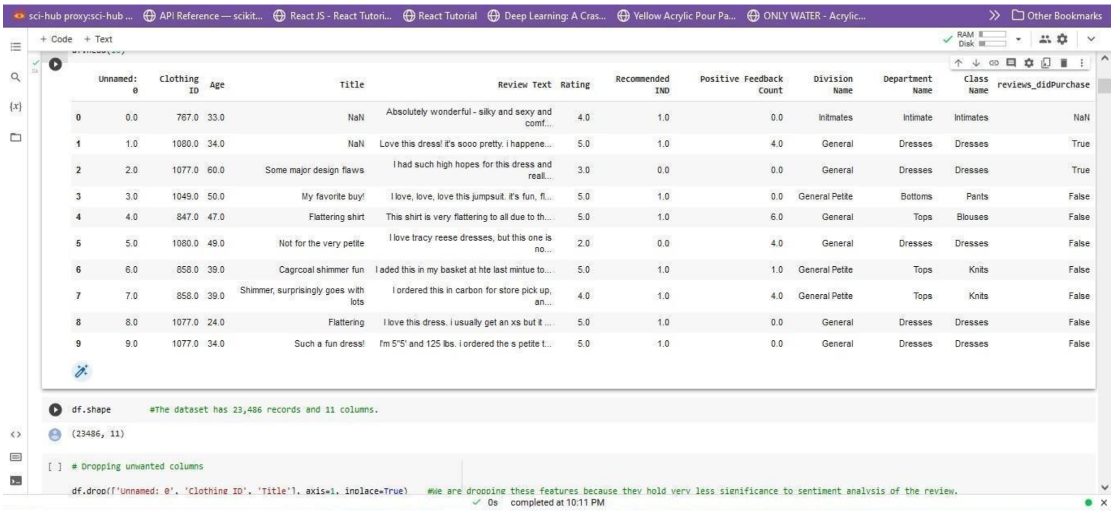
3.5 Implementation Requirements

This section describes the preprocessing, data collection, missing values, unexpected feature removes, data encoding and data exploration.

3.5.1 Data collection and preprocessing

Data collection:

Obtained The information via Kaggle. The CSV-formatted information comprises consumer reviews for women's clothes. The following Figure 3.4 depicts the raw data.



	Unnamed: 0	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name	reviews_didPurchase
0	0.0	767.0	33.0	NaN	Absolutely wonderful - silky and sexy and comf...	4.0	1.0	0.0	Initmates	Intimate	Intimates	NaN
1	1.0	1080.0	34.0	NaN	Love this dress! it's sooo pretty. i happene...	5.0	1.0	4.0	General	Dresses	Dresses	True
2	2.0	1077.0	60.0	Some major design flaws	I had such high hopes for this dress and reall...	3.0	0.0	0.0	General	Dresses	Dresses	True
3	3.0	1049.0	50.0	My favorite buy!	I love, love, love this jumpsuit. it's fun, fl...	5.0	1.0	0.0	General Petite	Bottoms	Pants	False
4	4.0	847.0	47.0	Flattering shirt	This shirt is very flattering to all due to th...	5.0	1.0	6.0	General	Tops	Blouses	False
5	5.0	1080.0	49.0	Not for the very petite	I love tracy reese dresses, but this one is no...	2.0	0.0	4.0	General	Dresses	Dresses	False
6	6.0	858.0	39.0	Cagrcoal shimmer fun	I aded this in my basket at hte last mintue to...	5.0	1.0	1.0	General Petite	Tops	Knits	False
7	7.0	858.0	39.0	Shimmer, surprisingly goes with lots	I ordered this in carbon for store pick up, an...	4.0	1.0	4.0	General Petite	Tops	Knits	False
8	8.0	1077.0	24.0	Flattering	I love this dress. i usually get an xs but it ...	5.0	1.0	0.0	General	Dresses	Dresses	False
9	9.0	1077.0	34.0	Such a fun dress!	I'm 5"5' and 125 lbs. i ordered the s petite t...	5.0	1.0	0.0	General	Dresses	Dresses	False

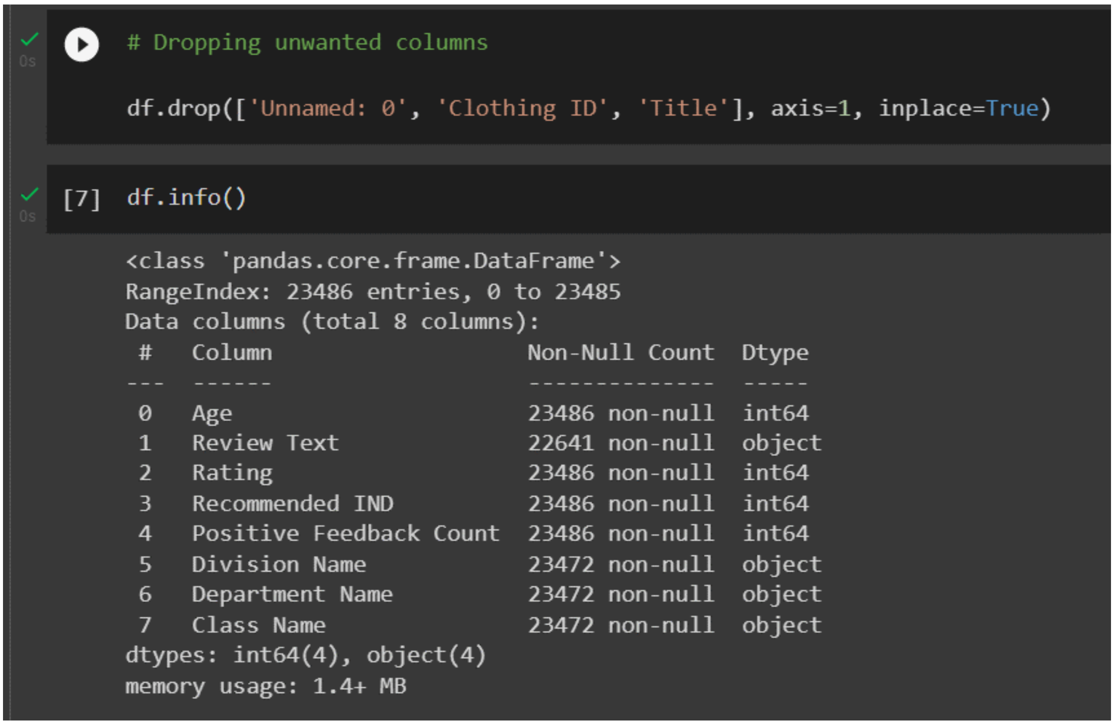
```
df.shape        #The dataset has 23,486 records and 11 columns.
(23486, 11)

# Dropping unwanted columns
df.drop(['Unnamed: 0', 'Clothing ID', 'Title'], axis=1, inplace=True)    #We are dropping these features because they hold very less significance to sentiment analysis of the review.
```

Figure 3.4 Raw data

Remove all the unexpected features:

The raw data is depicted in Figure 3.4. In this dataset, the unexpected features are named "unnamed," “clothing ID," and “Title." We are dropping these features because they hold very little significance for the sentiment analysis of the review. In the same way, special characters like n, stop words, numbers, punctuation, and return lists of words should be taken out of review texts. The following Figure 3.5 shows that the unexpected features have been removed.



```
# Dropping unwanted columns

df.drop(['Unnamed: 0', 'Clothing ID', 'Title'], axis=1, inplace=True)

[7] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23486 entries, 0 to 23485
Data columns (total 8 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Age                      23486 non-null  int64
 1   Review Text              22641 non-null  object
 2   Rating                   23486 non-null  int64
 3   Recommended IND          23486 non-null  int64
 4   Positive Feedback Count  23486 non-null  int64
 5   Division Name            23472 non-null  object
 6   Department Name          23472 non-null  object
 7   Class Name               23472 non-null  object
dtypes: int64(4), object(4)
memory usage: 1.4+ MB
```

Figure 3.5 Unexpected feature removed

Missing Values:

Missing values can affect the performance of a model. It is critical to check for missing values prior to building the model. Missing values, such as null, can be replaced with 0 or deleted. There are more missing values in the review text. Due to the fact that the Class Name, Division Name, and Department Name variables only contain 14 missing values combined, the missing observations for these three variables are being removed. Meanwhile, we should verify the data type. There are four integer characteristics and four object characteristics. To prepare for data exploration, convert Review Text variables to strings. Figure 3.6 depicts the missing value of raw data.

```
df.isnull().sum()

Age                        0
Review Text              845
Rating                     0
Recommended IND            0
Positive Feedback Count    0
Division Name             14
Department Name           14
Class Name                14
dtype: int64
```

Figure 3.6 Missing values

Data Encoding:

Data encoding seeks to quantify factors that cannot quantify. The dataset must be encoded because models cannot distinguish between recommended and not-recommended data. Column recommended IND 0 represents not recommended, and column recommended IND 1 represents recommended. In this endeavor, the term "sentiment" is used for the Rating. A score of 3 or more is considered neutral, a score of two or less is considered negative, and a score of 4 or more is a positive assessment of the situation. To prepare for data exploration, encode the class, Division, and department names. The following Figure 3.7 shows which products customers recommended based on Rating.

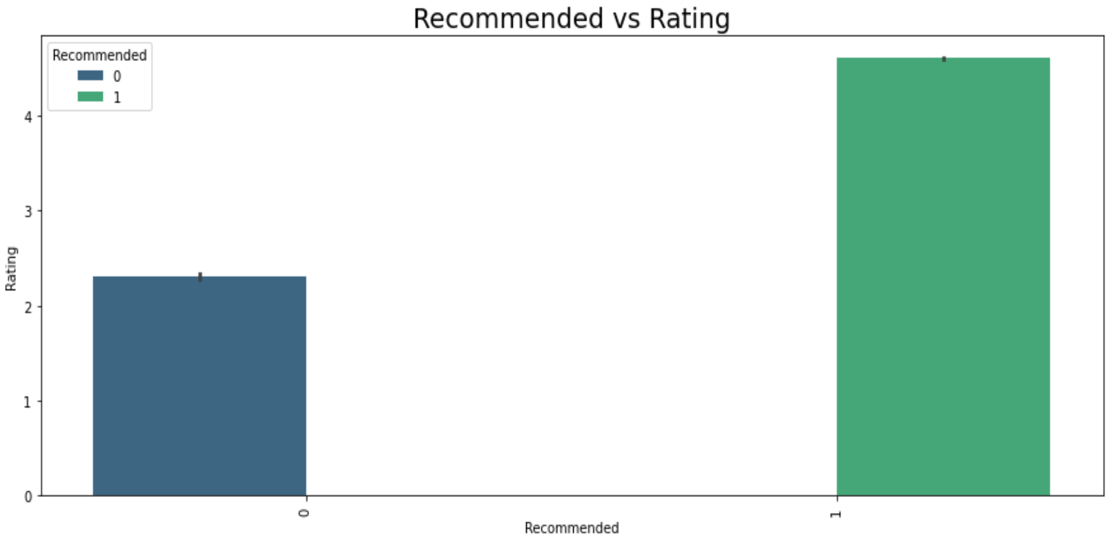


Figure 3.7 Recommended

The above Figure, 1 means "recommended." 0 means not recommended. Women recommend a product if they rate it at or above 3.

The distribution of the sentiment polarity of reviews based on the recommendation is displayed here in Figure 3.8.

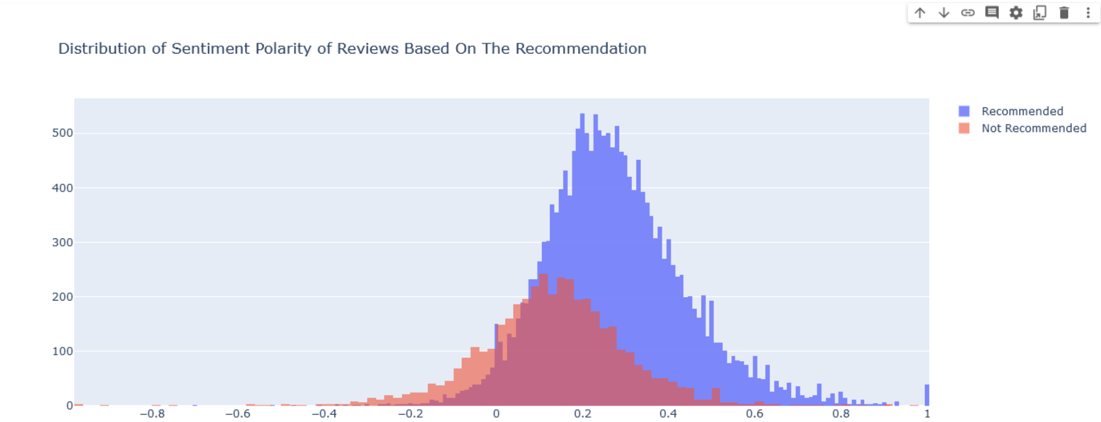


Figure 3.8 Distribution of Sentiment Polarity of Reviews

Figure 3.9 displays the polarity distribution of Rating-based reviews.

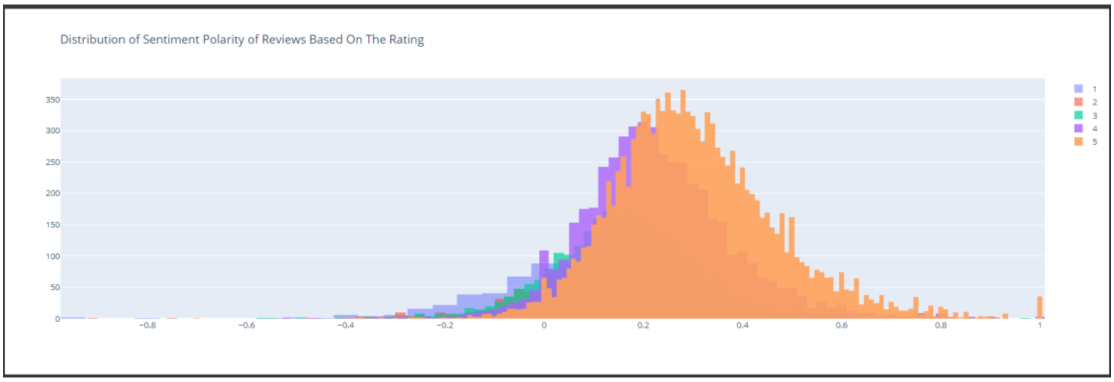


Figure 3.9 Distribution of Sentiment Polarity in Reviews Based on the Rating

3.5.1.5 Data Exploration

Google Colab, an open-source edition, is used for data exploration and preprocessing. NumPy is a Python library for manipulating arrays. It also has routines for dealing with linear algebra and matrices. Pandas is a Python library for manipulating data collections. It includes tools for data analysis, cleansing, exploration, and manipulation. Seaborn is a package that plots graphs using Matplotlib. Matplotlib is a tool for creating graphs, histograms, and bar charts. The following Figure 3.10 shows the data types of the dataset, where both integer and object have a value of 4.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23486 entries, 0 to 23485
Data columns (total 8 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Age                      23486 non-null  int64
 1   Review Text              22641 non-null  object
 2   Rating                   23486 non-null  int64
 3   Recommended IND          23486 non-null  int64
 4   Positive Feedback Count  23486 non-null  int64
 5   Division Name            23472 non-null  object
 6   Department Name          23472 non-null  object
 7   Class Name               23472 non-null  object
dtypes: int64(4), object(4)
memory usage: 1.4+ MB
```

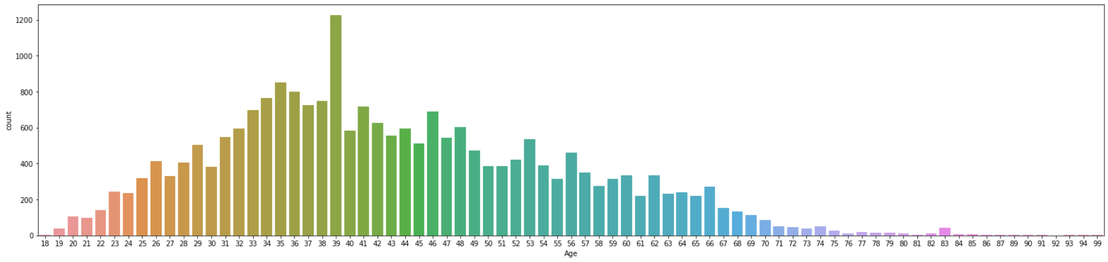
Figure 3.10 Data Types

Performed the data exploration to acquire a better knowledge of the data set. The review texts are string types. Using Word Cloud to investigate this. We added Word Cloud to the cleaned review data to find helpful information. Color, fit, dress, love, top, size, wear, great, will be ordered, and wear was the most frequently used words, as shown in Figure 3.11. It was simple to determine that people between the ages of 25 and 56 are more likely to post online reviews, particularly those over the age of 39, whose number of reviews exceeded 1200 which is shown in Figure 3.12.

Most clothes belong to the General Division, Tops Department, and are dresses or knits. In the division part, "general" got the best score, "intimates" got the worst score, and "dress" got the second-best score in the department part. The number of Tops reviews was greater than 10,000, the highest in the Department. Compared to Tops, Trend had the lowest number of reviews, with just under 2000. From the standpoint of the class name, the number of dresses was double that of blouses. Casual bottoms, chemises, and trends were the least expensive. Most respondents in the Division's Intimates section recommend it, while only a few don't. With almost the same number of "recommend" and "don't recommend" votes, the general group's trends were different from those of the "intimates" group. Except for dresses, the number of recommended items in the Department was more significant than those not recommended. The recommendations in each class and the Department followed the same pattern. Only those who did not recommend jeans outnumbered those who did. It was evident from the graph that item 1078 gained the most popularity. The following Figure 3.11 shows the Word Cloud

Figure 3.11 Word Cloud

The following shapes contain the data exploration in figure 3.12



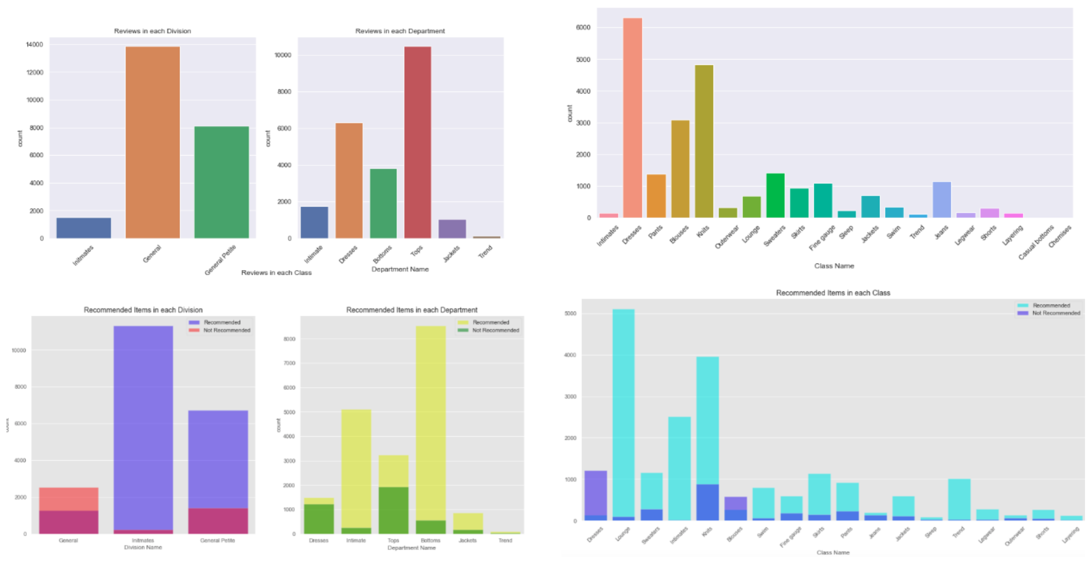


Figure 3.12 Data exploration

Figure 3.13 shows us the positive feedback according to age. There's no unique relationship between Age and Positive Feedback, excluding some outliers.

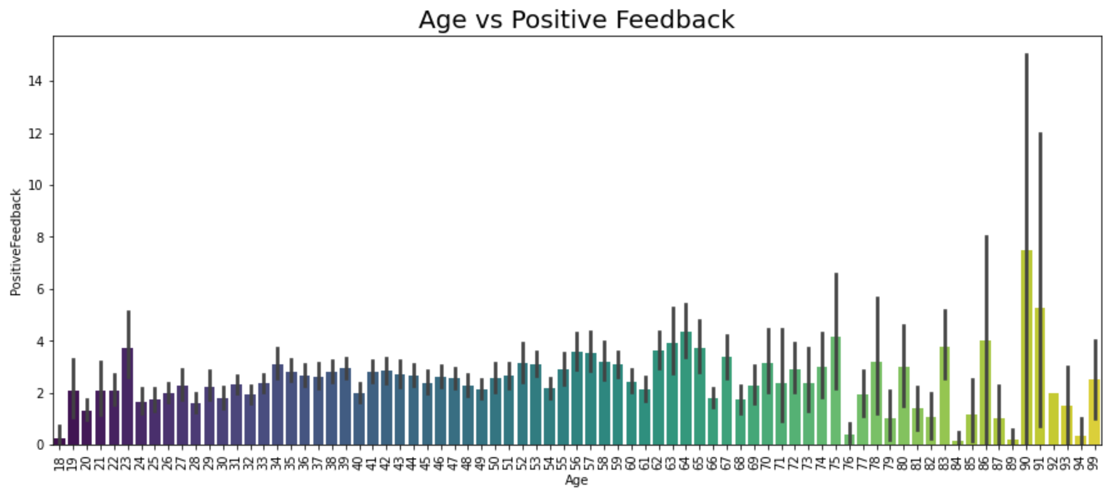


Figure 3.13 Age vs. Positive Feedback

3.5.1.6 Sentiment Polarity

One of the most important and frequently used functions in text analytics and NLP is sentiment analysis, which determines whether a word, phrase, or document is positive, negative, or neutral. The most important aspect of sentiment analysis is analyzing a body of text to understand its point of view. We typically quantify this emotion with a positive or negative number known as polarity. The polarity score's sign is frequently used to determine whether the dominant emotion is positive, neutral, or negative. A polarity score of -1 to -0.5 usually indicates negative sentiment. Polarity values greater than -0.5 and less than +0.5 usually indicate neutral sentiment. Positive sentiment is typically indicated by a polarity score of +0.5 to 1.

Figure 3.14 depicts the review-based positive polarity.

Reviews with positive polarity

```
[35] ex = df.loc[df.polarity == 1,['Review']].sample(3).values
     for i in ex:
         print(i[0])

So comfortable-so versatile-so perfect
Perfect for lunch with the girls, pta, or saturdays with the family. this skirt is a perfect addition to your fall wardrobe.
This is a great blouse for all shapes!   you can wear it under a suit or with jeans!
```

Figure 3.14 Positive polarity

The following Figure 3.15 shows the review-based negative polarity.

Reviews with negative polarity

```
[36] ex = df.loc[df.polarity < 0,['Review']].sample(3).values
     for i in ex:
         print(i[0])

I ordered a small and i'm usually always a small. i could not zip it up. i'm am returning it for a large, and this is concerning because what if it is big in a weird way
I like the pointelle feature in the front. i end up wearing a lot of black, but this at least gives me the option to do a bit of color underneath if i so chose. i've wor
My husband never has anything negative to say about my clothes until i wore these pants! they are not flattering and kept stretching out until i could pull them off with
```

Figure 3.15 Negative Polarity

The following Figure 3.16 shows the review-based neutral polarity.

Reviews with neutral polarity

```
[37] ex = df.loc[df.polarity == 0.5,['Review']].sample(3).values
     for i in ex:
         print(i[0])

I treated myself to get this at full-price at the start of summer. my size is xl and it fits me like the model photos. i have gotten such good use out of this blouse. it
Love this top! cloth & stone makes excellent quality clothes that are durable and have unique details. this top can go with many different looks and i can't wait to take

i bought the medallion crops to go with this top and it is an awesome combo, i would definitely check it out!
I love this shirt. i would recommend an undershirt.
```

Figure 3.16 Neutral Polarity

The following Figure 3.17 depicts the Polarity graph based on the Rating.

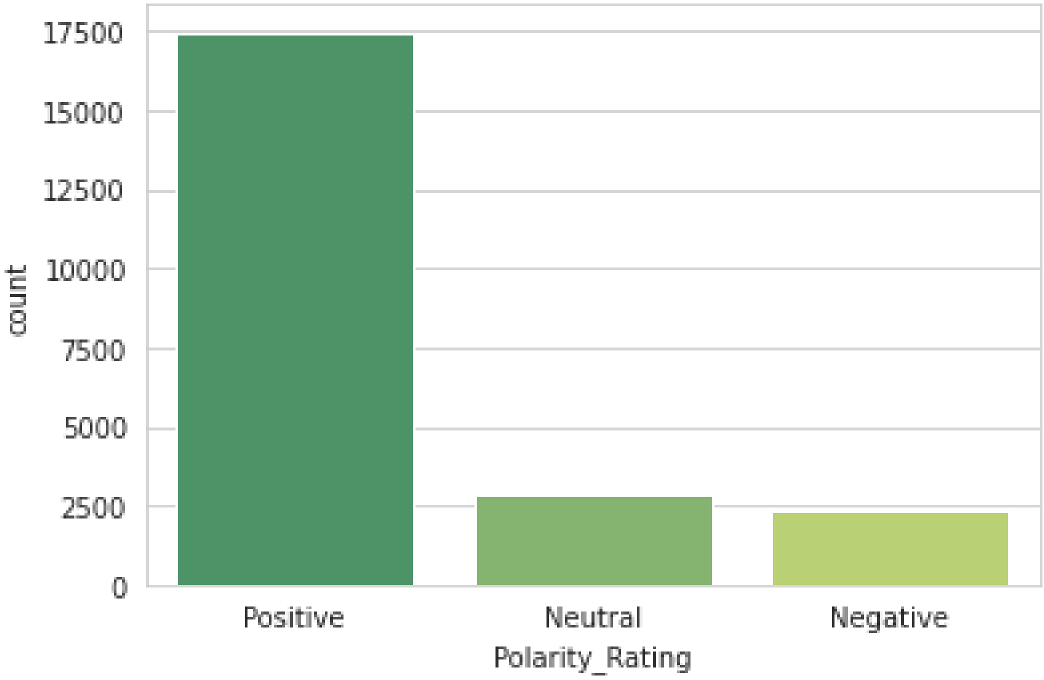


Figure 3.17 Polarity Graph

3.5.1.7 Feature Selection

There are a lot of characteristics in the dataset, but not all of them can be used to make the model. Feature selection might pick the most important traits and combine the ones that look similar. In our investigation, the heat map aids in feature selection. Figure 3.18 depicts the correlation matrix of Age by Positive Feedback Count.

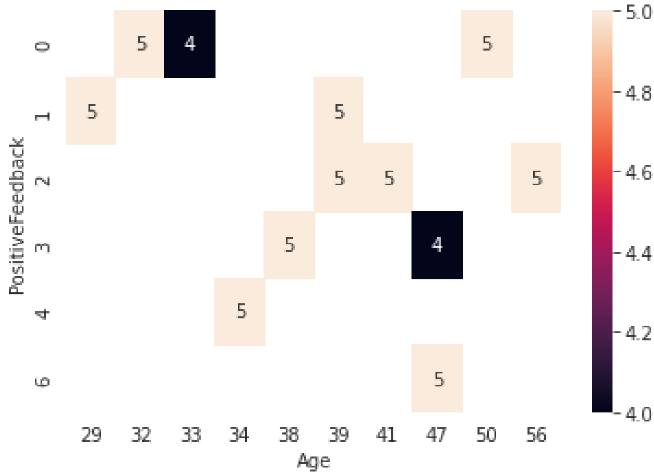


Figure 3.18 Age by Positive Feedback Count.

The following figure 3.19 shows the word frequency distribution in review texts per rating, department, and recommendation.

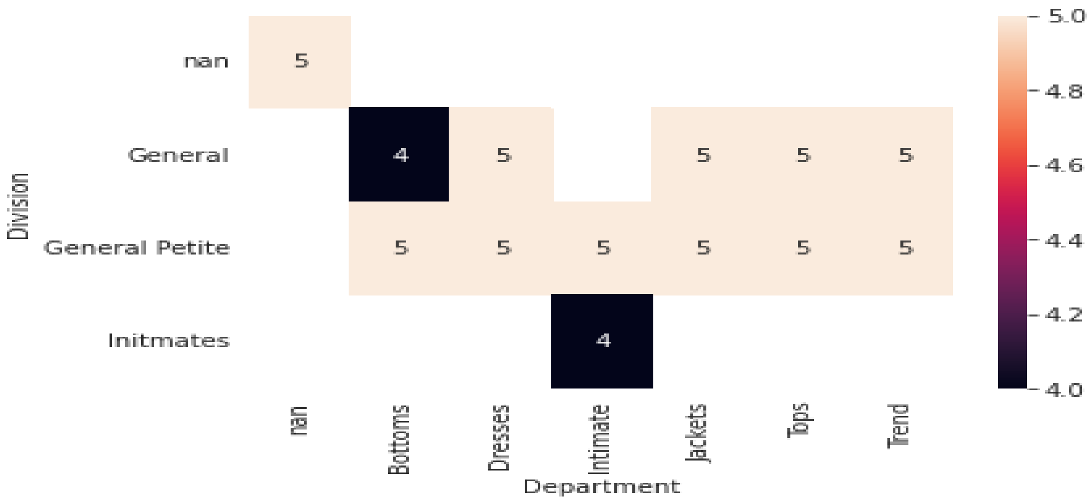


Figure 3.19 Word frequency distribution in review texts per rating, department, and recommendation.

3.5.1.8 Feature Extraction

Since most sentiment analysis methods employ or rely on machine learning techniques, text or document characteristics are represented as feature vectors. The characteristics listed below are used in sentiment analysis –

TF_IDF:

To vectorize the data, use the TF-IDF vectorizer method to get the features from the textual data. IDF is an abbreviation for inverse term frequency, and TF is for term frequency.

Term Frequency (TF) = (number of times the term appears in the document / total number of terms in the document)

The formula for calculating the inverse term frequency is IDF = log*(number of documents in the corpus/number of documents in the corpus that contain the term).

TF-IDF = TF * IDF

We use SMOTE oversampling techniques to rebalance the data before modeling because it is class-imbalanced. The following Figure 3.20 shows using TF-IDF for feature extraction.

Feature Extraction

```
[75] ##using TF-IDF vectorizer using the parameters to get 100  features.
     vectorizer = TfidfVectorizer(stop_words='english', max_features=100, max_df=0.9, min_df=7, binary=True,
                                  ngram_range=(1,2))
     X_train_tfidf =vectorizer.fit_transform(df['Cleaned_Review'])

     y= df['Recommended']
```

```
[43] print(vectorizer.get_feature_names_out())

     ['beautiful' 'better' 'big' 'bit' 'black' 'blue' 'body' 'bought' 'casual'
      'color' 'colors' 'comfortable' 'cut' 'cute' 'definitely' 'design' 'dress'
      'fabric' 'fall' 'feel' 'fit' 'fits' 'flattering' 'going' 'good'
      'gorgeous' 'got' 'great' 'high' 'jeans' 'large' 'lbs' 'length' 'light'
      'like' 'little' 'long' 'look' 'looked' 'looking' 'looks' 'loose' 'love'
      'loved' 'lovely' 'make' 'makes' 'material' 'medium' 'model' 'nice'
      'online' 'ordered' 'pants' 'perfect' 'perfectly' 'person' 'petite'
      'pretty' 'price' 'purchased' 'quality' 'really' 'recommend' 'retailer'
      'right' 'runs' 'sale' 'saw' 'shape' 'shirt' 'short' 'size' 'skirt'
      'sleeves' 'small' 'soft' 'store' 'style' 'summer' 'super' 'sweater'
      'think' 'thought' 'tight' 'time' 'tried' 'true' 'true size' 'try'
      'usually' 'waist' 'wanted' 'way' 'wear' 'wearing' 'white' 'work' 'worn'
      'xs']
```

Figure 3.20 Feature extraction

POS (Part of speech) Tagger:

Each word in a phrase has a syntactic function that governs how it is utilized. Other names for the syntactic roles include "parts of speech." In English, there are eight grammatical parts: the verb, the noun, the pronoun, the adjective, the adverb, the preposition, the conjunction, and the interjection. Part-of-speech (POS) taggers have been created in natural language processing to categorize words based on their parts of speech. A POS tagger is beneficial for sentiment analysis for two reasons: 1) Nouns and pronouns do not frequently include emotion. It is possible to filter out such terms using a POS tagger; 2) A POS tagger may also use to differentiate words that could use in various parts of speech. For example, as a verb, "improved" may convey a different degree of emotion than an adjective. The following Figure 3.21 shows the graph of the POS tagger.

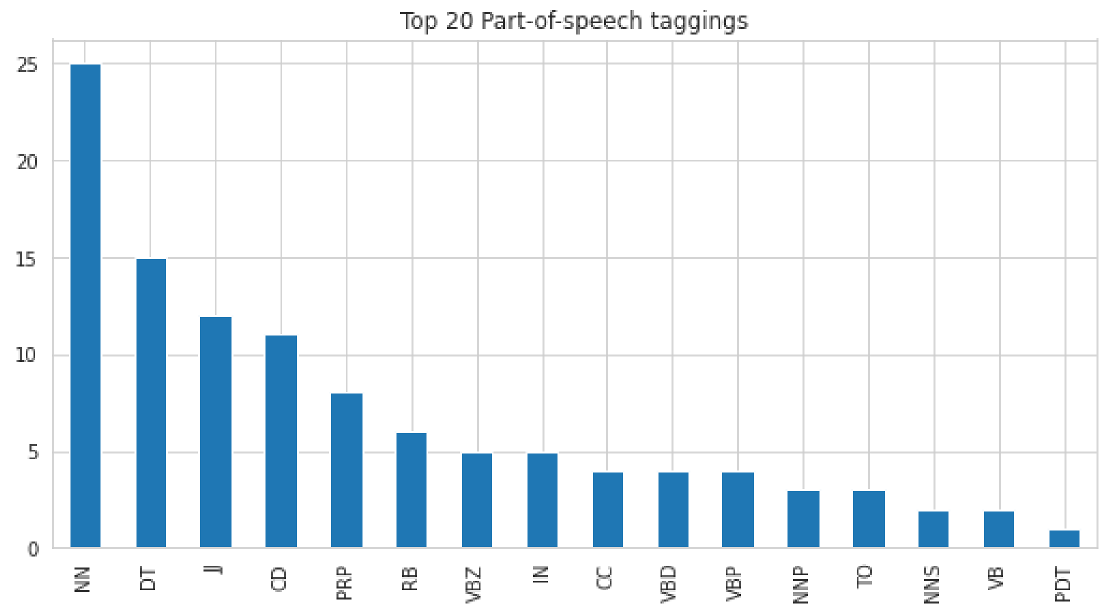


Figure 3.21 POS Tagger

Figure 3.22 depicts rebalancing the data using SMOTE-ENN under Edit-Nearest-Neighbours.

```
# Rebalance the Data with SMOTE-ENN
X_train, y_train=SMOTEENN(enn=EditedNearestNeighbours(sampling_strategy='majority')).fit_resample(X_train,y_train)

y_train.value_counts()
```

```
1    13015
0    12934
Name: Recommended, dtype: int64
```

Figure 3.22 Rebalanced data

Design specification:

The workflow chart for the research is shown in Figure 3.23. There are four stages involved. First, data from Kaggle is collected, followed by data preprocessing, which includes data exploration, missing values, special characters, data encoding, and feature selection. Next, several classifications are constructed to train the dataset. The models use the accuracy, precision, recall, f1-score, AUC, and ROC curves as evaluation measures. The following figure 3.23 shows the Workflow chart.

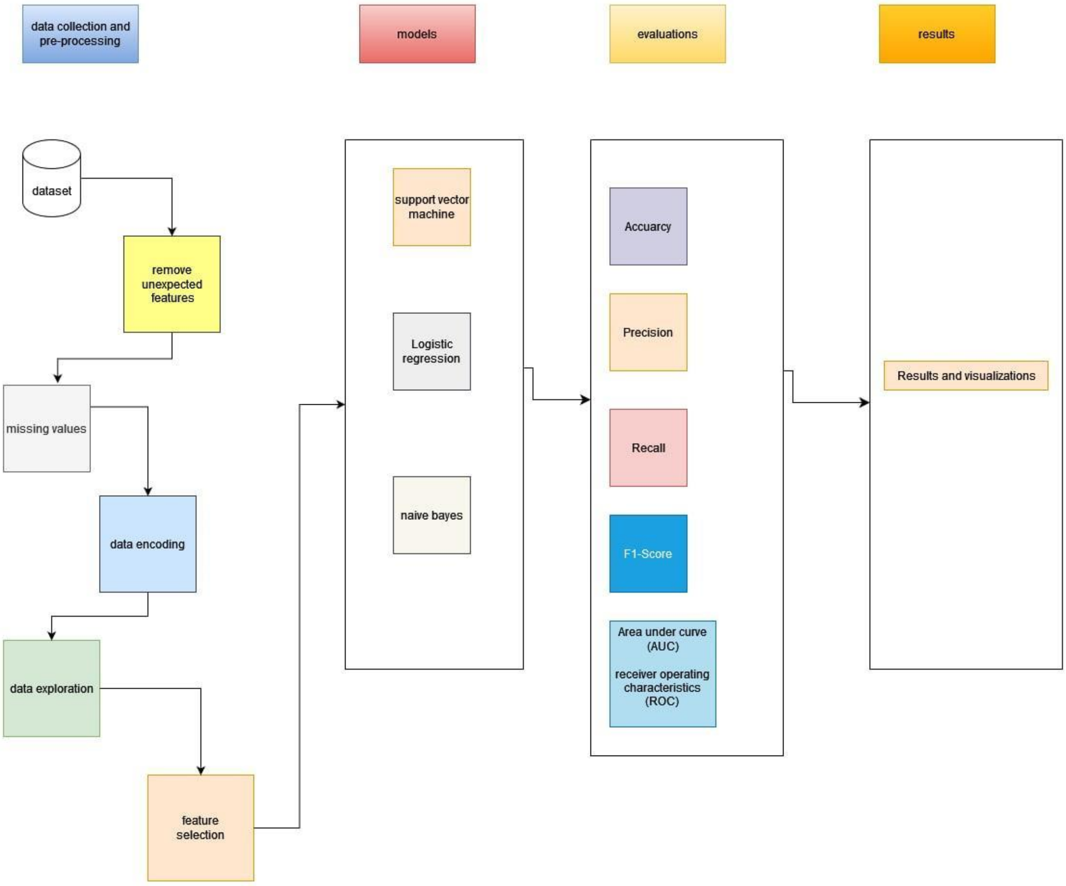


Figure 3.23 Workflow chart

3.5.2 Models:

The heat map allowed us to define the features used in the models: Rating, age, positive feedback, reviews, polarity, and sentiment. 70% of the new data set was training data, and 30% was test data. We used the train-test split to divide the data set.

Support Vector Machine:

The Support Vector Classifier is a fast classifier that uses the linear kernel function to do classification. Linear SVC is a rapid-implementation classifier that performs classification using the linear kernel characteristic. Linear support vector classifiers help shape the data we give and "fine-tune" the selection to narrow down that type of suspect or knowledge. With more options for loss features and praise, many examples will be available. By using SVM, the accuracy shows that 78.7%

The following figure 3.24 shows the accuracy of the support vector machine.

SVM

```
# Initiate and train the model
svm = LinearSVC().fit(X_train, y_train)

# Make prediction
svm_pred = svm.predict(X_test)

# Show model performance
print("Accuracy:{:.3f}". format(accuracy_score(y_test, svm_pred)))
print("Precision:{:.3f}". format(precision_score(y_test, svm_pred)))
print("Recall:{:.3f}". format(recall_score(y_test, svm_pred)))
print("F1 Score:{:.3f}". format(f1_score(y_test, svm_pred)))
print("ROC AUC Score:{:.3f}". format(roc_auc_score(y_test, svm_pred)))
```

```
Accuracy:0.787
Precision:0.923
Recall:0.806
F1 Score:0.860
ROC AUC Score:0.756
```

Figure 3.24 Support vector machine

Figure 3.25 depicts the confusion matrix of support vector machine.

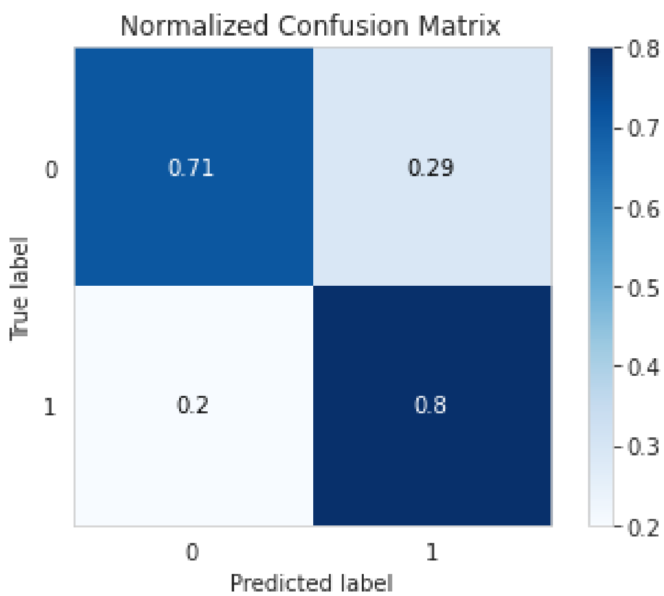


Figure 3.25 Confusion matrix of the support vector machine

3.5.2.2 Logistic Regression

Using supervised learning, the logistic regression classifier determines the probability of a response variable. The performance of the logistic regression classifier is 79%. The following figure 3.26 illustrates logistic regression.

```
lr = LogisticRegression().fit(X_train, y_train)

# Make prediction
lr_pred = lr.predict(X_test)

# Show the model performance
print("Accuracy:{:.3f}". format(accuracy_score(y_test, lr_pred)))
print("Precision:{:.3f}". format(precision_score(y_test, lr_pred)))
print("Recall:{:.3f}". format(recall_score(y_test, lr_pred)))
print("F1 Score:{:.3f}". format(f1_score(y_test, lr_pred)))
print("ROC AUC Score:{:.3f}". format(roc_auc_score(y_test, lr_pred)))
```

```
Accuracy:0.789
Precision:0.923
Recall:0.809
F1 Score:0.862
ROC AUC Score:0.757
```

Figure 3.26 Logistic Regression

Figure 3.27 depicts the confusion matrix of logistic regression.

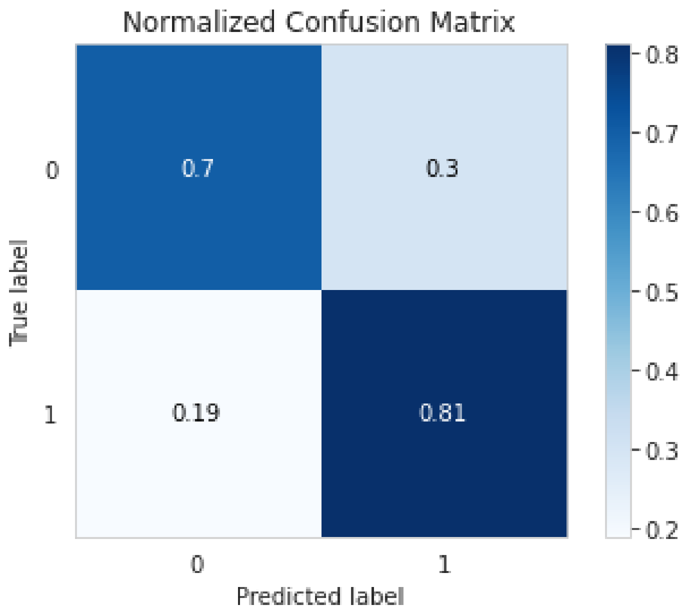


Figure 3.27 Confusion matrix of Logistic Regression

3.5.2.3 Naive Bayes

The theorem of Bayes is the basis for the Naive Bayes method. This theorem says that the characteristic condition must be independent. It is expected that X's and its properties are conditionally independent of how the class says they should be. The Naive Bayes result is shown in Figure 3.28. The rate of accuracy reached 77.0%.

The following Figure 3.28 shows the performance of the naive Bayes classifier.

Naive Bayes

```
[ ] # Initiate and train the model
    nb = MultinomialNB().fit(X_train, y_train)

    # Make Prediction
    nb_pred = nb.predict(X_test)

    # Show model performance
    print("Accuracy:{:.3f}". format(accuracy_score(y_test, nb_pred)))
    print("Precision:{:.3f}". format(precision_score(y_test, nb_pred)))
    print("Recall:{:.3f}". format(recall_score(y_test, nb_pred)))
    print("F1 Score:{:.3f}". format(f1_score(y_test, nb_pred)))
    print("ROC AUC Score:{:.3f}". format(roc_auc_score(y_test, nb_pred)))

    Accuracy:0.770
    Precision:0.921
    Recall:0.785
    F1 Score:0.847
    ROC AUC Score:0.746
```

Figure 3.28 Naïve Bayes

The following Figure 3.29 shows the confusion matrix of naive Bayes.

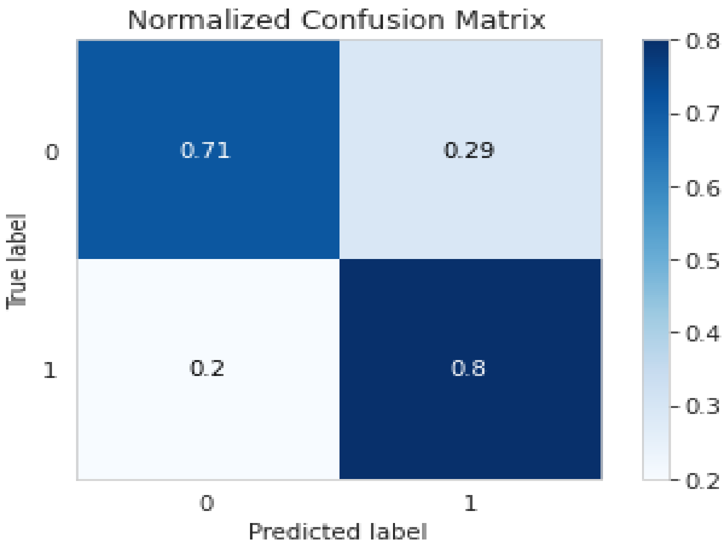


Figure 3.29 Confusion matrix of Naive Bayes

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

Did all the preprocessing and data exploration in Google Collaboratory (collab), which lets anyone write and run Python code in the browser and is especially good for machine learning. Packages such as Pandas, NumPy, Matplotlib, Seaborn, and Word Cloud were installed and used. Pandas contain a plethora of functions and methods for analyzing data sets. NumPy is a Python open-source numerical computing extension that can store and process large matrices. Matplotlib creates visualizations like plots, histograms, bar plots, and other graphs and plots. The express module, usually imported as px, has functions that let you make whole figures at once. Inverse Frequency Document The frequency of records is referred to as the TF-IDF. It can sum up how important a word is to a group of words in a text called a corpus. We utilized sent_tokenize and word_tokenize as tokenizers, then WordNetLemmatizer for lemmatization. We applied the SMOTE-ENN approach to balance the imbalanced dataset. Out of the three classification models we used in this study to predict the data, LR worked the best. The accuracy score, precision score, f1 score, roc-AUC score, and confusion matrix were used to show how well the measures worked.

4.2 Experimental Results and Analysis

The performance of an algorithm will be measured using precision, recall, F1-score, and AUC. Table 4.2.1 presents the confusion matrix.

Table 4.2.1 Confusion Matrix

Algorithm	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Positive (FP)
Actual Negative	False Negative (FN)	True Negative (TN)

Accuracy is measured by how many cases were correctly labeled out of the total number of cases being looked at.

Precision: Precision denotes positive predictive value, and the equation is

TP/(TP+FP)

Recall/Sensitivity: Recall denotes the actual positive rate, and the equation is

TP/(TP+FN)

Accuracy: Accuracy denotes all correct predictions of total datasets, and the equation is

(TN+TP)/(TN+TP+FN+FP)

F-score: F-score is the harmonic mean of precision and recall, the F1 score is determined by recall and precision and the equation is-

(2*Precision*Recall) / (Precision + Recall)

ROC curve: A receiver operating characteristic curve (ROC curve) is a graph that represents the performance of a classification model throughout all classification levels. In a ROC curve, a higher X-axis value means more false positives than real negatives. A more excellent Y-axis value suggests more true positives than false negatives. So, the value of the threshold is based on how well it can balance false positives and false negatives. This graph depicts two parameters:

1. FP/(FP+TN) = FPR
2. TPR =TP/(TP+FN).

The "area under the curve" is represented by the numeric value known as the AUC. The value of the representation is readily obvious. The area between the receiver operating characteristic (ROC) curve and the coordinate axis is what the ROC-AUC graph illustrates. The model has a more significant effect when the TPR is high, and the FPR is low. If we want the model to have a better performance overall, the ROC curve should be positioned so that it is closer to the top left corner. Accuracy, precision, recall, F1-score, and area under the curve are all presented in Table 4.2.2.

Table 4.2.2 Algorithm and metrics

Algorithm	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	79%	92%	81%	86%	75%
Naive Bayes	75%	92%	76%	83%	74%
Support Vector Machine	75%	92%	80%	86%	74%

The following figure 4.1 shows the ROC-AUC curve of the three models. The curve of LR and SVM seems similar.

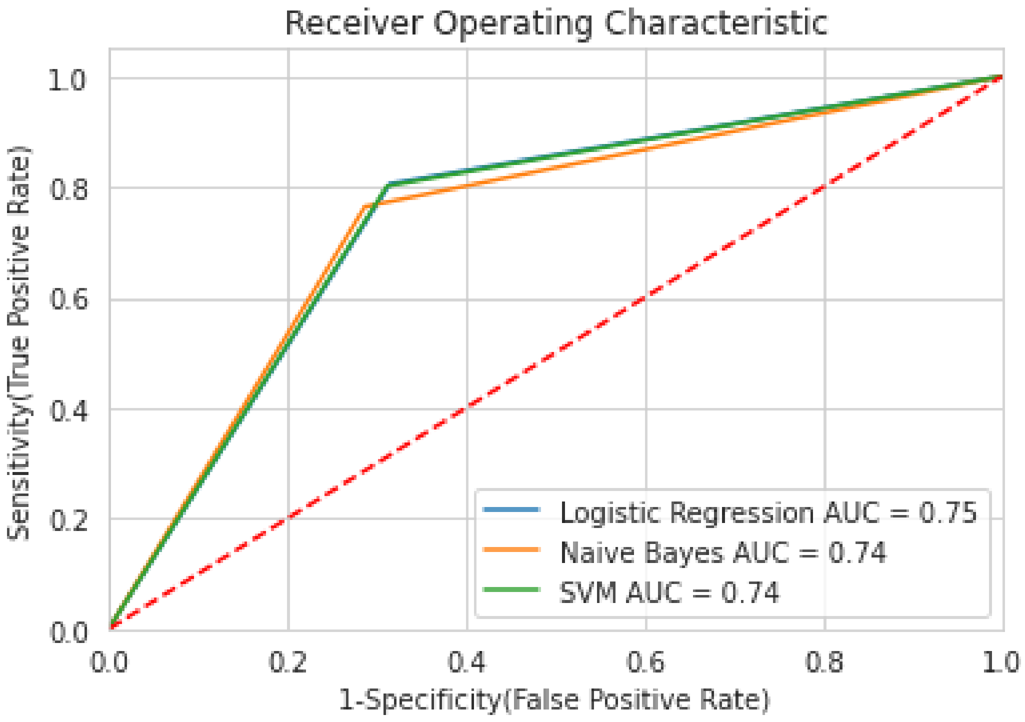


Figure 4.1 ROC-AUC curve

The following figure 4.2 shows the accuracy, precision, recall, f1 and Roc-Auc score of the three models.

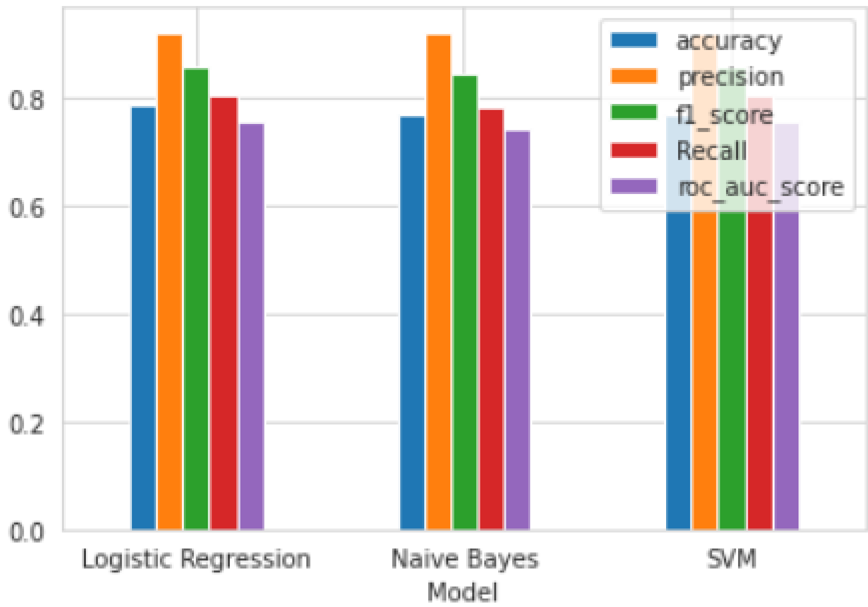


Figure 4.2 Accuracy score

4.2.1 Support Vector Machine

It can be seen from Table 4.2.2 that the accuracy of the SVM is 75%, which is a very respectable score. The precision value is 92%, which is similar to logistic regression and naïve bayes. ROC-AUC and F1 scores are 74% and 86%, respectively.

4.2.2 Logistic Regression

The precision value for logistic regression and SVM is 92% in Table 4.2.2, and the f1 score, recall, and AUC scores are 86%, 81%, and 75%, respectively. The AUC shows the highest value compared with the other two classifiers. It performs better than SVM. It has an accuracy of 79%.

4.2.3 Naïve Bayes

The accuracy of the Naive Bayes method is 75%, which is comparable to that of the SVM and LR. The precision of Naive Bayes is 92%, and the F1 score is 83%. The recall and AUC scores are 76% and 74%, respectively.

4.3 Discussion

Experimental analysis shows that LR has the highest accuracy, F1 score, and recall compared with other classification models. LR had the highest precision and AUC. From the above study and research, we figured out that the LR approach may make it easier to classify how consumers feel. In reality, customers can see, based on the classification results, whether the product is what they want. In contrast, businesses can see which products have low demand and need to be repositioned in the market to meet customers' needs better and boost profits.

CHAPTER 5
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Consumers are benefited in the decision-making process by reading reviews and ratings online. Customers are 63% more inclined to trust a brand that has online reviews than one that does not have any evaluations at all. Customers perceive a greater level of risk when there are no reviews available, which results in them being less willing to make a purchase. Establishing trust is the most important thing for customers, and online testimonials and reviews are a great way to do so. This is especially important for online purchases, because customers can't ask questions about a product before they buy it. Buyers look for information about customer satisfaction from users in the same industry, business size, and function when they read reviews. Focusing on groups they can relate to will make the information more beneficial when buying things or making decisions. It applies to both positive and negative reviews. In our research paper, we have shown that positive reviews will help customers buy good products. Most customers buy products according to the ratings, and our project gives accurate ratings compared to the review text.

5.2 Impact on the Environment

Today, investing in the software tools that matter to the owner's entire organization takes more than just learning about a product. Customers want to fully understand what it means to be involved with the business. What do you value as a company? What is the buying process looks like? What is your experience with customer service, support, and success? Of course, the business sales representative can ask all of these questions. However, to get the most valuable and trustworthy insights, they reach out to peers and product users for their insights and star ratings. If the customer gives positive feedback instead of negative feedback, the product quality will impact the business, which is good.

5.3 Ethical Aspects

We used a dataset on product reviews from the Internet for our study. It is open-source information. Discovered no unethical problems. We don't collect any information about our customers without their permission. Instead, we help them by analyzing their messages with our algorithms. We build a model to help new customers find a better product by analyzing customers' previous texts and reviews. Our research project can be advantageous to both the buyer and the seller.

5.4 Sustainability Plan

In our research paper, we are working on sentiment analysis of women's clothing reviews. We have shown that there are many more positive responses than negative ones. We must ensure that our research leads to good results, not bad or neutral ones. If the rate of positive feedback is higher than the rate of negative feedback, product quality will improve, and sales will grow. Both of these things have a significant effect on business. Also, when new customers visit the product, they will notice the Rating first, and if the Rating is high according to the reviews, they will be convinced and buy the product. In the future, we will be required to add a recommendation system to the project in order to assist customers in finding similar products.

CHAPTER 6
SUMMARY, CONCLUSION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

In this research, we have utilized device-learning algorithms to improve the accuracy of predicting positive, negative, and neutral ratings for clothing and to provide the user with a more accurate prediction of clothing reviews. In the below section, we are able to:

- Data collection
- Data Cleaning
- Data prepossessing
- Apply NLTK (Natural Language Toolkit)
    - Tokenize
    - Remove Punctuation
    - Word cloud and POS tagger
    - Lemmatization
- Feature Selection
- Apply machine learning algorithms for classification.

6.2 Conclusions

Sentiment analysis is a field in which we understand the customer's feelings, thoughts, and emotions about specific items. This research addresses the classification of feelings according to their polarity as a challenge. In this project, distinct online product reviews and ratings from e-commerce websites are recorded and utilized as the basis for applying a classifier to the reviews and ratings. Customers may quickly discern the polarity of the reviews made on an e-commerce website if the sentiment generated by the site is displayed most effectively. This is how we receive the most relevant product reviews. We used NLTK to analyze sentiment by using the Amazon women's clothing reviews as a data set. Evaluated the classifiers by comparing their accuracy in different cases of experimentation. Regarding accuracy, logistic regression is better than SVM and Naive Bayes, and the algorithms can classify things correctly more than 90% of the time.

6.3 Implications for Further Study

In the future, more factors, like location, job, and salary, will be used to categorize customer feedback better. The predicted growth encourages customer satisfaction and faith in online shopping. We could add a recommendation system in the future, tell the difference between fake and honest reviews, and give accurate results to help consumers choose the best product.

REFERENCES

[1] P. D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” Dec. 2002.

[2] A. K. Prajapati, A. Pandey, and M. K. Jain, “Sentimental Analysis on E-commerce Women’s Clothing,” *International Journal for Research in Engineering Application & Management (IJREAM)*, vol. 06, p. 5, 2020, doi: 10.35291/2454-9150.2020.0555.

[3] A. F. Agarap, “Statistical Analysis on E-Commerce Reviews, with Sentiment Classification using Bidirectional Recurrent Neural Network (RNN),” May 2018.

[4] B. Liu, “Sentiment Analysis and Opinion Mining,” *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[5] D. M. Khan, T. A. Rao, and F. Shahzad, “The Classification of Customers’ Sentiment using Data Mining Approaches,” *Global Social Sciences Review*, vol. IV, no. IV, pp. 146–156, Dec. 2019, doi: 10.31703/gssr.2019(IV-IV).19.

[6] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, “Pulse: Mining Customer Opinions from Free Text,” 2005, pp. 121–132. doi: 10.1007/11552253_12.

[7] A. I. Adekitan and O. Salau, “The impact of engineering students’ performance in the first three years on their graduation result using educational data mining,” *Heliyon*, vol. 5, no. 2, p. e01250, Feb. 2019, doi: 10.1016/J.HELIYON. 2019.E01250.

[8] B. Pang, ¡£¢, L. Lee, ¡¤¢, and ¡£¥, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,” 2005. [Online]. Available: http://www.sover.net/

[9] J. Cheng, S. Zhao, J. Zhang, I. King, X. Zhang, and H. Wang, “Aspect-level Sentiment Classification with HEAT (HiErarchical ATtention) Network,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Nov. 2017, pp. 97–106. doi: 10.1145/3132847.3133037.

[10] S. Mansour, “Social media analysis of user’s responses to terrorism using sentiment analysis and text mining,” in *Procedia Computer Science*, 2018, vol. 140, pp. 95–103. doi: 10.1016/j.procs.2018.10.297.

[11] C. M. Suryavanshi, “Sentiment Analysis of Product Reviews: Opinion Extraction and Display,” 2021.

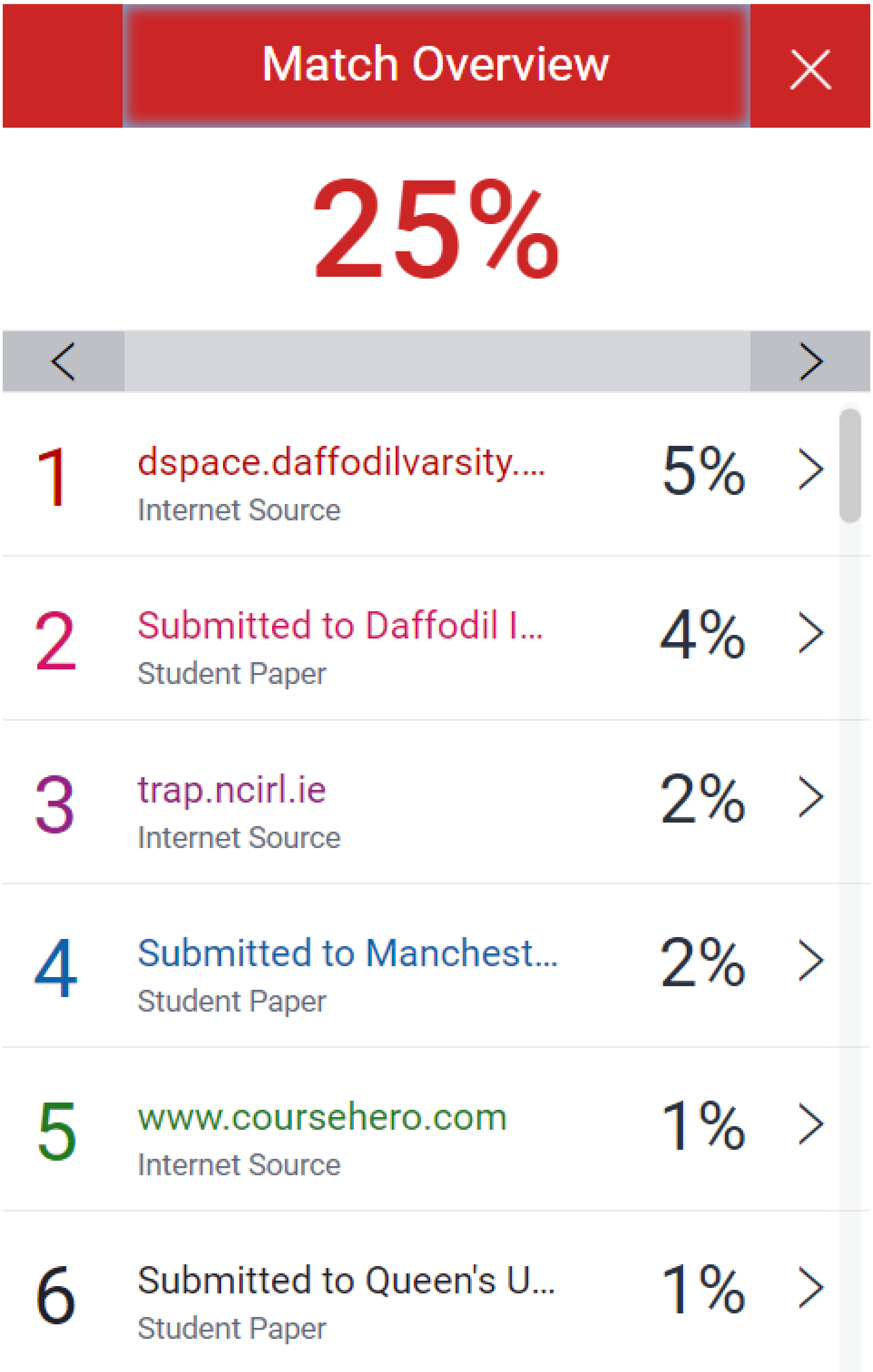
[12] T. K. Shivaprasad and J. Shetty, “Sentiment analysis of product reviews: A review,” in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Mar. 2017, pp. 298–301. doi: 10.1109/ICICCT.2017.7975207.

[13] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh, "Sentiment Analysis on Product Reviews Using Machine Learning Techniques," 2019, pp. 639–647. doi: 10.1007/978981-13-0617--4_61.

[14] S. Moghaddam, "Beyond Sentiment Analysis: Mining Defects and Improvements from Customer Feedback," 2015, pp. 400–410. doi: 10.1007/978-3-319-16354-3_44.

[15] A. Dhara, "Sentiment Analysis of Product-Based Reviews Using Machine Learning Approaches," 2017.

PLAGIARISM


Match Overview
25%
1
dspace.daffodilvarsity....
Internet Source
5%
2
Submitted to Daffodil I...
Student Paper
4%
3
trap.ncirl.ie
Internet Source
2%
4
Submitted to Manchest...
Student Paper
2%
5
www.coursehero.com
Internet Source
1%
6
Submitted to Queen's U...
Student Paper
1%