Sentiment Analysis to Improve Product Marketing Strategies

BY

Oalid Hassan Orpan

ID: 201-15-13940

FINAL YEAR DESIGN PROJECT REPORT

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

Supervised By

Md. Sazzadur Ahamed
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by

Mr. Dewan Mamun Raza
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

13 JULY 2024

APPROVAL

This Research titled **"Sentiment Analysis to Improve Product Marketing Strategies."** submitted by **Oalid Hassan Orpan** to the Department of Computer Science and Engineering, Faculty of Science and Information Technology, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science & Engineering and approved as to its style and contents. This Presentation has been held on 13 July, 2024.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori (SRH) — **Board Chairman**
Professor &Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zahid Hasan (ZH) — **Internal Examiner 1**
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Md Mohammad Masum Bakaul (MB) — **Internal Examiner 2**
Sr. Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zulfiker Mahmud (ZM) — **External Examiner**
Professor
Department of Computer Science and Engineering
Jagannath University

©Daffodil International University

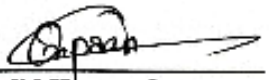
DECLARATION

We hereby declare that this research has been done by me under the supervision of **Md. Sazzadur Ahamed, Assistant Professor, Department of Computer Science and Engineering**, Faculty of Science and Information Technology, Daffodil International University. I also declare that neither this research nor any part of this research has been submitted elsewhere for the award of any degree.

Supervised by:

Md. Sazzadur Ahamed
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Oalid Hassan Orpan
ID: 201-15-13940
Department of CSE
Daffodil International University

©Daffodil International University

ACKNOWLEDGEMENT

At First, I would like to express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year thesis successfully.

I really grateful and wish my profound my indebtedness to **Supervisor Md. Sazzadur Ahamed,** Professor, Department of CSE, Faculty of Science and Information Technology, Daffodil International University, Dhaka. His Deep Knowledge & keen interest with supportive instructions helped us in the field of deep learning-based research, finally I completed my work on "**Sentiment Analysis to Improve Product Marketing Strategies.**". Their endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Md. Sazzadur Ahamed**, Assistant Professor, Department of Computer Science and Engineering, Faculty of Science and Information Technology, DIU, for his valuable support and advice to finish my project and also heartiest thanks to other faculty member and the staff of department of CSE, Daffodil International University.

At last, again I want to thank all the good wishers, friends, family, seniors for all the help and inspirations. This research is a result of hard work and all those inspirations and assistance.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

©Daffodil International University

ABSTRACT

In the whole world, the platform of social networking have become a vital part for expressing feelings because of the swift growth in the digital age. Most of the people convey their feelings by using textual content, pictures, audio and video. Customer reviews and ratings give important insights on numerous e-commerce platforms such as Daraz,, Flipkart, Amazon is famous e-commerce websites to buy online products. The reason is it enables you to have a look on several of additional customer reviews for the product you are considering. As people share their opinion about a product by rating, commenting other people seek to understand the positive, negative experience. This understanding can be done by sentiment analysis. Sentiment Analysis processes text, comments, rating, reviews and categorizes them into positive, negative opinions. These will be helpful for the consumers and also the sellers to purchase any products by seeing the ratings and reviews and sellers can improve the products too.

The sentiment analysis can be done by using Machine Learning models, NLP. The system carries pre-processing operations- stemming, lemmatization, tokenization, visualization, sentiment mapping to extract meaningful information from dataset. It assigns a rank based on the classification of the data as positive or negative.

©Daffodil International University

TABLE OF CONTENTS

©Daffodil International University

©Daffodil International University

©Daffodil International University

LIST OF FIGURES

©Daffodil International University

©Daffodil International University

LIST OF TABLES

©Daffodil International University

CHAPTER 1
Introduction

1.1 Introduction

Due to internet social media has taken over the entire world and have shrunk the world's dimensions. People can communicate through those platforms and exchange their opinions by using those social media. These opinions reflect the mood and sentiment of the people helping to understand their behavioral characteristics. Sentiment Analysis also referred to as opinion mining along with Natural Language Processing (NLP). Sentiment Analysis is being used for monitoring public opinion and aid decision making. Hence, sentiment analysis has a strong foundation due to the abundance of online data. This online data has several flaws that can potentially hinder the sentiment analysis process. The common flaw is that since people can freely post their own opinion, the quality of their opinions can not always be taken.

However, It is essential to analyze and interpret the opinions, reviews, ratings, feedbacks from the text based data. Sentiment Analysis, the machine learning method helps to identify sentiments which enables the entrepreneurs, consumers, sellers to gather insights from various online platform such as social media, surveys, e-commerce website reviews etc. For that vendor utilize the social platforms, e-commerce platforms to disseminate information about the products and efficiently gather client feedback. Active feedback from consumers is important for clients and sellers.

1.2 Objectives

Use sentiment analysis to find complaints and issues raised in reviews to increase customer happiness and service quality. This method aids in solving persistent issues. Examine the sentiment and tone of consumer feedback to develop sympathetic reaction plans. To track and improve customer support teams' performance, keep an eye on sentiment trends. Provide solutions by tailoring support messages to the feelings and particular issues raised in reviews. Recognize attitudes to take proactive measures to address problems before they get worse. Leverage reviews to create compelling marketing content that effectively highlights the benefits of your products and the satisfaction levels of your customers. Use consumer feedback to optimize marketing plans and campaigns. Control brand perception by keeping an eye on feelings, responding to criticism, and improving the brand's image. Measure the success of marketing campaigns by assessing

©Daffodil International University

reviews and post-campaign opinions to determine the influence of the campaign. To find areas for development or advantages for market positioning, conduct analysis by contrasting the sentiments expressed in product reviews with those of rivals. Determine the true product aficionados for influencer partnerships by using their sentiments to identify brand ambassadors. Analyze consumer comments to group them based on their opinions while taking into account their needs and preferences. Create marketing materials and promotions that speak to specific customer groups based on their opinions and personal input. Use sentiment analysis to determine which features or attributes are important to different client segments so that you may improve features and drive product innovation.

1.3 Motivation

Understanding consumer mood is essential for developing successful marketing tactics in today's cutthroat industry. Businesses can obtain profound insights into customer experiences and preferences by using sentiment analysis, which provides a potent tool for interpreting and viewpoints represented in product reviews. Sentiment analysis allows businesses to move beyond surface-level metrics and delve into the emotions and attitudes of their customers. By understanding how customers feel about various aspects of a product, marketers can identify what truly resonates with their audience. This deep understanding helps in crafting messages that speak directly to customer needs and desires, creating a more authentic and engaging marketing narrative.

Sentiment analysis offers measurable information about consumer sentiment, facilitating data-driven decision-making. Instead of relying solely on conjecture, marketing plans can be improved by using real client feedback. By doing this, marketing initiatives are guaranteed to be in line with consumer tastes and expectations, which boosts campaign effectiveness and returns on investment (ROI). Sentiment analysis feedback can direct innovation and product development. Making well-informed decisions about new product lines or feature additions requires an understanding of what customers enjoy and hate about a certain product. In addition to ensuring that marketing efforts are concentrated on promoting the most appealing aspects, this alignment with customer desires increases the quality of the product.

©Daffodil International University

1.4 Expected Outcome

When sentiment research is applied to enhance product marketing strategies, a number of favorable results are anticipated. Gaining more insight into customer insights is one of the main results. Through the examination of customer reviews and feedback, companies can ascertain the genuine feelings and viewpoints of their clientele, so facilitating the development of more authentic and persuasive marketing communications.

An increase in client satisfaction is another expected result. Businesses may swiftly discover and resolve consumer discontent issues by utilizing sentiment research. In addition to enhancing the good or service, this quick response increases customer loyalty and converts happy consumers into brand ambassadors.

1.5 Layout of the Report

The thesis's purpose, goals, expected outcomes and also this section are all included in Chapter 1. This particular section also covers the report's general structure.

All of the earlier research in this field is covered in Chapter 2. Later in the subsequent chapter, they provide an example of the scope that results from their limitation of this subject of research. The last topic addressed dealt with the primary difficulties or obstructions to this investigation. This chapter includes parts on pertinent studies and study summaries in addition to discussing the difficulties encountered throughout project development.

The theoretical appraisal of the attempt of this study is explained in Chapter 3. More details on the statistical methods specifically used to the investigation's mathematics component are provided in this chapter. Additionally, this chapter includes examples of procedural ways to use machine learning techniques. The next chapter describes the process for obtaining datasets and the data preparation system. Confusion matrix analysis is also used in the latter section of this subdivision to assess the model and offer the accuracy tag for the classifier. Implementation analysis must be included in order to ensure genuine accuracy when using machine learning techniques. This section covers the research topic and tools, efficiency of operation, data collection method, data processing, suggested model, instruction mode, and execution requirements that must be met in

©Daffodil International University

order to move this project forward. Every machine learning strategy and classification used in this study is explained in great depth for each technique.

Chapter 4 presents the implementation, train model. A few informative photos are included in this chapter to aid in the project's implementation. A summary of ML methods marks the end of this chapter. Additionally, describe the online reviews and ratings used in Amazon products.

Chapters 5 provided the experimental result and discussion, classification and descriptive analysis of all the used models, result’s pictures, confusion matric pictures and Chapter 6 describes its impact on society, environment and life with sustainability plan.

Lastly, Chapter 7 overview of the research, information on planned actions, and a conclusion. To show that the construction report complies with all standards, this chapter offers a verified example. Effect on Sustainable Development, Society as a whole and Environment: The final section of the chapter draws attention to the limitations of our efforts, which might have an effect on those who work in this field in the future.

©Daffodil International University

CHAPTER 2

Background

2.1 Introduction

Machine learning is used for sentiment analysis. My work's primary goal is to identify positive and negative reviews and ratings from text based dataset so that models with high prediction accuracy may be used to improve in product marketing strategies. Here, I implement a novel formulation of the optimal model work with outcomes. Next, we examine further and previous studies to bolster our suggested methodology. I have taken datasets include with reviews and ratings of amazon products.

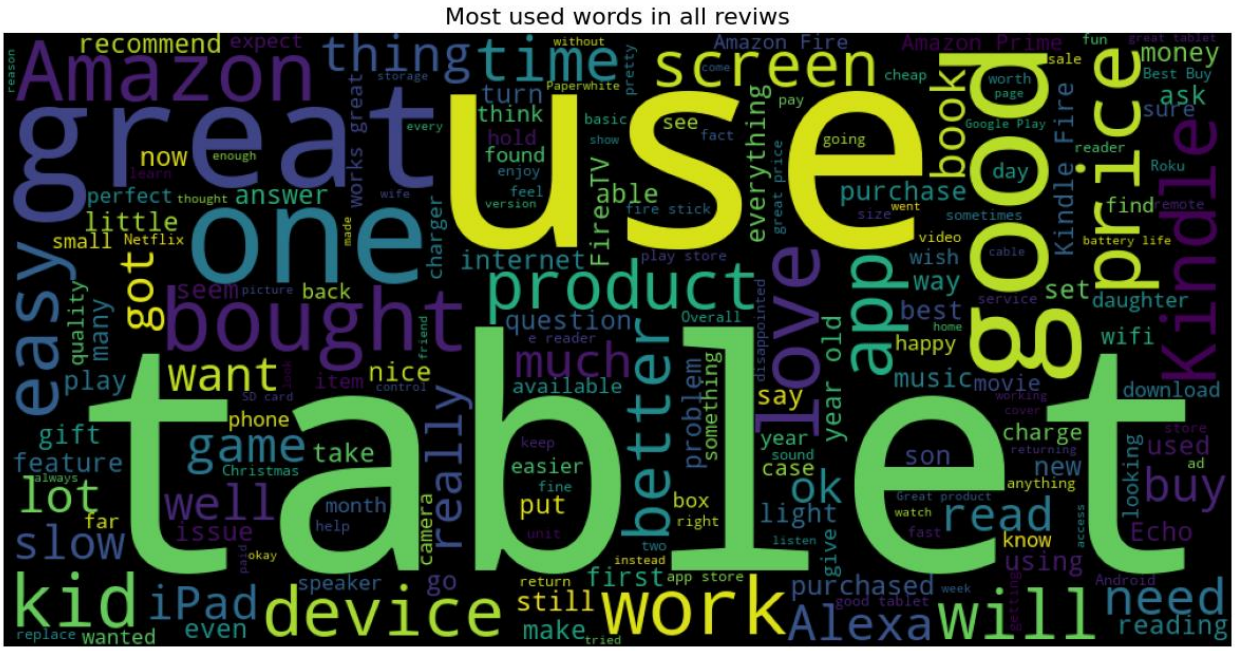


Fig 2.1: Most used words in all reviews

.

Figure 2.1 displays the word cloud from the deliverable that represents the Most used words in all reviews.

©Daffodil International University

The internet has become the primary information source for a large number of individuals nowadays, and this trend has to continue. I am going to sentiment analysis to improve product marketing startegies that will increase the confidence of seller and consumers. My models allow users to view the most current data on the main keywords or sources that are providing the most erroneous information, along with accuracy that has been charted.

2.2 Related Works

Authors of [2] has shown the various process approaches of sentiment analysis. There are Lexicon based approach, Corpus based approach, Dictionary based approach, Machine Learning approach. As I am going to use Machine Learning approach, their research will be beneficial for me. Hence they also mentioned some models such as Naïve Bayes, Support Vector Machine, Logistic Regression, Decision Tree, KNN etc.

In [1] they did a survey of sentiment analysis algorithm and illustrate a graph of number of articles which give the information about six categories of SA task among the years and said that Lexicon base algorithms usually use for sentiment analysis. They also illustrate pos classification in the overall count.

Inside [3] their study is about reviewing data of product using sentiment analysis as well. They have used three models Random forest, Naïve Bayesian and Support Vector machine. Author of [4] done a review on sentiment analysis and emotion detection from text and they have suggested Transfer Learning Approach in traditional approaches. Sentiment Analysis has been done from Twitter text in [5]. They collected the data from twitter comment, replies and reactions keywords.

Author of [6] use various amazon product dataset to do sentiment analysis. They get 72.95% accuracy on cellphone accessories, 84.44% on books, 87.33% on kindle products.

We look into [7] where the author collected dataset from amazon which contains Camera, Laptops, Mobile Phones, TVs. They used Naïve Bayes and SVM models.

©Daffodil International University

2.3 Research summary

My project endeavor's main focus is the range of approaches that the larger community has to offer. All in all, we have expanded our dataset with additional methods and distinct strategies. My publicly accessible amazon dataset has served as the primary information source in this case. As mentioned before, my collection includes the electronics product review of amazon which I take from kaggle. I'll be able to evaluate the performance of the strategies I employed and also assess the influence of the extra information I offered from the very same source. Since there are several comparable classes and labeling types, it implies. Most of the machine learning techniques and related text classification algorithms used in the data preparation and feature extraction procedures were written in Python. My go-to engine for feature extraction techniques was Python.

2.4 Scope of the problem

In essence, our work involves evaluating the given data and using machine learning techniques to build a model. With the help of our approach, reviews and ratings could be diagnosed. The individuals living in this neighborhood will be greatly impacted by this project. A problem that many people in today's society deal with is confusion about purchasing a product and the seller also can not maintain the standard. Despite this, they are unaware of the extent of their misery. They start to sell products of low quality willingly or unwillingly. Consumer wants to purchase premium or quality products as they are not able to see or can not verify in person. So they are very much dependent to the reviews and ratings. The project will collect data from Kaggle and use natural language processing (NLP) technologies to evaluate it in order to find patterns and traits of reviews and ratings. After that, this data will be used to train the algorithm to recognize the ratings with accuracy.

2.5 Challenges

Combining and assessing all of these sets of data seems to be the primary issue with this research, given how difficult it was to examine a single big data file. I cleaned and normalized the dataset using tools and techniques. The vastness of the data sets and the variety of levels they included

©Daffodil International University

from different times made it take some time to get the required results. Additional data sets are contained in this field. I never completed the study, therefore I have to put in a lot of effort on any linked assignments, which hinders my ability to quickly come up with the finest answers. The development of natural language processing (NLP) techniques to automatically classify instances of text through which it has garnered a lot of attention since it may be useful in identifying and curtailing such behavior. It might be challenging to create a unified strategy because different organizations and scholars have relatively varying definitions of what constitutes a category. Furthermore, it might be challenging for an NLP system to recognize negative and positive review because of their complex and ambiguous language when they denigrate and harm their targets. The use of the irony, sarcasm, and other figurative speech can make classification more difficult.

- Collecting data
- Data Importing,
- Data preprocessing and labeling,

©Daffodil International University

CHAPTER 3

Research Methodology

3.1 Introduction

This section covers the approach used in the research, which covers how to gather datasets, carry out each experiment, and make use of each model to improve accuracy. We also presented model technique and all-data work in this chapter. Thus, in an attempt to enhance and streamline the data offered in this chapter. A description of the full procedure and the methods used to analyze sentiment also provided in this part. This section will discuss the study methodology in its entirety. There are several ways to solve each analysis. As previously mentioned, the data set is taken from Kaggle. The amount of data id 34,661 which contains product details, id, manufacturer, ratings and reviews. The selection of an ML strategy is the next stage. As I have previously stated, generating the collection of data is a requirement to developing the framework and applying the algorithm since I am using machine learning methodologies. Firstly, I imported various required libraries for data manipulation and warnings suppression. Then load the dataset from CSV file into a pandas DataFrame, displays the first few rows and shape of the dataset. For analysis 10,000 random datasets has been sampled. Drops and checks missing values, Counts the frequency of ratings and sorts them. Used bar plot to visualize the ratings. To balance the dataset samples a specific number of reviews and also concatenated the sampled data to visualize the new distribution. Rated 1-3 to Negative ratings and 4-5 to Positive and added new colums for sentiment scores and lables. Also I used NLP for text processing(Tokenization, stopword removal, stemming, lemmatization), Feature Extraction(TF-IDF Vectorization), Visualization, Sentiment Mapping

This data is then used to train the model. This is the basis for feature selection. Next, the data was split into testing and training sets. It is often referred to as the training dataset and the testing dataset. I acquire just a significant portion of the data needed to assess our model after the input has been acquired from a training dataset and matched into many ML method models. The model's accuracy is then assessed. I will go over some of the operations with formulas and diagrams to help you better comprehend them. I have provided a description using our basic process flow chart. Below is a summary of the research project's methodology and the entire study attempt. Among

©Daffodil International University

the most important components are the data gathering, analysis, and proposed model. The suitable graph, explanation, design, and equation all help to further clarify the idea.

3.2 Subject of the Research and Equipment

A research field is an area of study that is looked at and studied to make concepts clearer for handling, creating models, gathering data, completing tasks, and training models, in addition to implementation. I talk about the equipment and methods that I as instrumentation professional's use. I made use of the Microsoft Windows operating system, the programming language Python, and other tools, including NumPy, Pandas, Matplotlib, LinearSegmentedColormap, Seaborn, re, nltk. All training and testing have been conducted using the Google Colab platform. On Google Colab, Python programmers are able to write code for methods related to data science and machine learning. These methods use machine learning-related statistical approaches to identify groups such as the seven abusive text classifications.

Libraries used:

- **Matplotlib:** Among Matplotlib's visualization tools is the Pyplot graphing, determining, plotting, and charting set of techniques. When building forms, this might be utilized, for instance, to draw attention to a narrative's boundaries or to highlight specific viewpoints in a story.
- **Seaborn:** This is a statistical data visualization library which has been used for high level interface to draw attractive statistical graphics. The next edition of this well-known Python information visualization program allows you to utilize matplotlib. This is a creative data visualization tool that is simple to use.
- **NumPy**: The Python NumPy utility has made vector processing easier. Matrix calculations, the Fourier transformed transform, and the transformation's indices are all explicitly covered in this topic. There are many tools and techniques available in the NumPy packages for Python to work with various types of matrices. NumPy can help make device design more logical and realistic. To put it simply, NumPy is a Python module

©Daffodil International University

designed largely for quantitative assessment. Another way to describe this would be "computation. Py."

- **Pandas:** Pandas provides a freeware toolbox for evaluating and working with data that is particular to a language. Systematic data administration may be aided by basic data types and statistical analysis procedures, especially when it comes to managing summary data.
- **re:** It provides matching operations of regular expression.
- **Nltk:** This library used for text processing. (Natural Language Toolkit).
- **Spacy:** A powerful and efficient open-source library for advanced natural language processing (NLP) in Python, designed specifically for production use and known for its performance and ease of use.
- **WordCloud:** A Python library that generates word clouds from text data, allowing for the visualization of the most frequent words in a text corpus, which can provide quick insights into the main themes and topics.
- **scikit-learn:** A comprehensive machine learning library in Python that offers simple and efficient tools for data mining and data analysis, including classification, regression, clustering, and dimensionality reduction.
- **transformers:** Developed by Hugging Face, this library provides state-of-the-art pre-trained models for natural language processing tasks such as text classification, translation, and summarization, enabling easy integration of advanced deep learning models.
- **xgboost:** An optimized gradient boosting library designed for speed and performance, widely used for structured/tabular data, offering efficient implementations of gradient boosting decision trees with parallel processing capabilities.
- **tf-keras:** The high-level neural networks API of TensorFlow, which allows for easy and fast prototyping, supports both convolutional and recurrent networks, and runs seamlessly on both CPU and GPU.

and many mor libraries has been used.

3.3 Workflow

It is possible that a number of methods or techniques are used in this study in order to get the desired outcomes. This experiment involved choosing a workflow, processing, gathering, and

©Daffodil International University

creating word clouds and stop words lists, cleaning up the text, and calculating the efficiency using its output.

Step 1: Data Collection: I processed the data after gathering it from Kaggle source. Since it is difficult to gather data for the specific text analyzer and kind of categorization, there isn't a large, comprehensive dataset accessible in this industry.

Step 2: Data Processing: Load the Amazon product reviews dataset from a CSV file. Displayed initial data and check its structure, including the number of rows and columns, to understand the dataset's format and content.

Step 3: Data Sampling and Cleaning: I have sampled a manageable subset of the data for analysis. Removed any missing values, generate basic descriptive statistics, and visualize the distribution of product ratings to get a sense of the data's overall quality and distribution.

Step 4: Data Balancing: Addressed class imbalance by resampling the data to ensure an even distribution of sentiment classes. Visualize the new distribution to confirm that the classes are balanced, which is crucial for effective model training.

Step 5: Label Mapping: Converted numerical ratings to binary sentiment scores (positive or negative) and map these scores to corresponding sentiment labels. Visualized the distribution of these sentiment labels to ensure correct mapping.

Step 6: Text Preprocessing: I have cleaned the review text by converting it to lowercase, removing extra whitespaces, HTML tags, digits, and punctuations. Remove stopwords and lemmatize the text to reduce words to their base forms, improving the quality of the textual data for analysis.

Step 7: Feature Extraction: Transformed the cleaned text data into numerical features using TF-IDF vectorization. Splited the data into training and test sets, and ensure the training data is balanced to avoid biased model training.

©Daffodil International University

Step 8: Model Training and Evaluation: Trained various classification models (Multinomial Naive Bayes, XGBoost, Logistic Regression, Random Forest, and SVM). Evaluate their performance using classification reports and ROC-AUC scores, and visualize confusion matrices to assess model accuracy and error rates.

Step 9: Advanced Model: Implemented BERT (Bidirectional Encoder Representations from Transformers) for advanced sentiment analysis. Train the model and evaluate its predictions, leveraging BERT's powerful natural language understanding capabilities for improved performance.

Step 10: Ensemble Method: Lastly, combined the predictions from multiple models using ensemble techniques to enhance overall predictive performance. Evaluate the ensemble model to determine its effectiveness compared to individual models, aiming for improved accuracy and robustness.

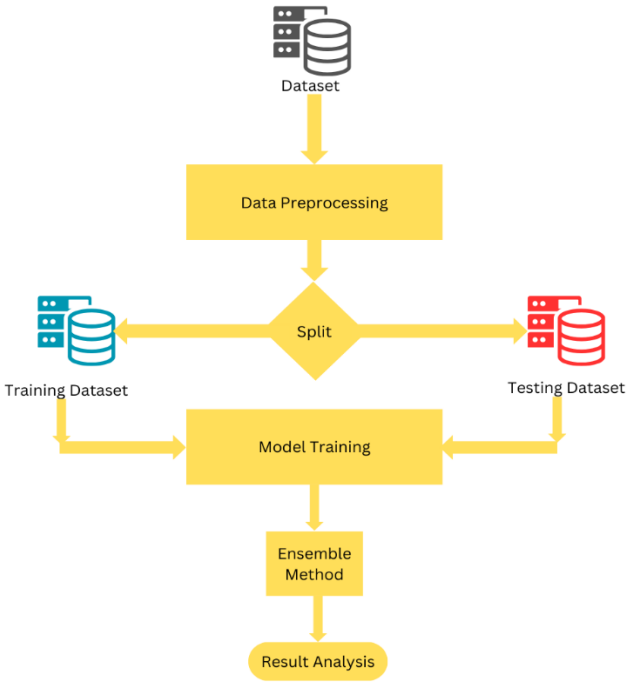


Fig 3.1: Workflow for the Entire Research Proposed.

©Daffodil International University

Fig. 3.1 illustrates the stages of our study methodology that may help product marketing strategies. Our online information is factual and has been compiled from a number of sources. I have reviewed data collection, removed any unnecessary words, and cleaned up the language to ensure that every single data set will only contain accurate and relevant information. To make the frequently used phrases simpler to read, word clouds are employed as visual aids. I study machine learning approaches by creating and refining models using pre-existing data. Apart from giving a summary, my approach searches for positive and negative reviews and ratings. By integrating language comprehension techniques with machine learning technologies, I can deliver a comprehensive solution that will be efficient for the sellers and consumers.

3.4 Data Collection

The dataset, which has been collected from Kaggle is actually the data of Amazon product reviews. The products are Electronics,iPad & Tablets,All Tablets,Fire Tablets,Tablets,Computers & Tablets. Accessories, power adopter, tech toys, iPad, stereos etc. The dataset also contain with Id, Name, Manufacture company, Review texts, Ratings, Review title, Reviewers name. The dataset contains with 34,661 overall data. Training and testing portions make up the two halves of the dataset.

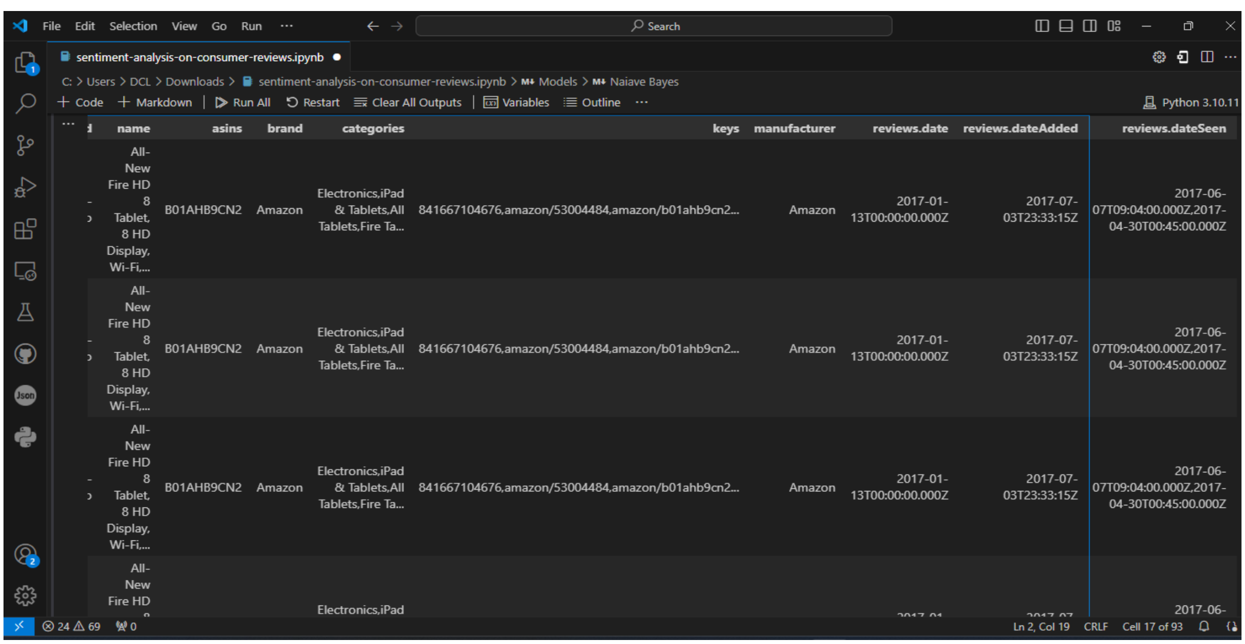


Fig 3.2: Text Sample data.

©Daffodil International University

	reviews.text	reviews.rating
21536	Bought as a Mother's Day Gift. This is great f...	4.0
20669	I can hold this next to my Kindle Paperwhite a...	5.0
30656	Love this device and went on to buy 2 as gifts...	5.0
25297	With some technical savvy, you can quickly hav...	5.0
9016	bought for grandkids they love them. wise choi...	5.0

Fig 3.3: Exploratory data analysis

3.5 Statistical Analysis

3.5.1 Analyzing Data

My goal for the data preliminary processing stage is to transform the raw textual data into a framework suitable for assessment and model training. Getting the necessary text and features from datasets that contain ratings and reviews is the initial step.

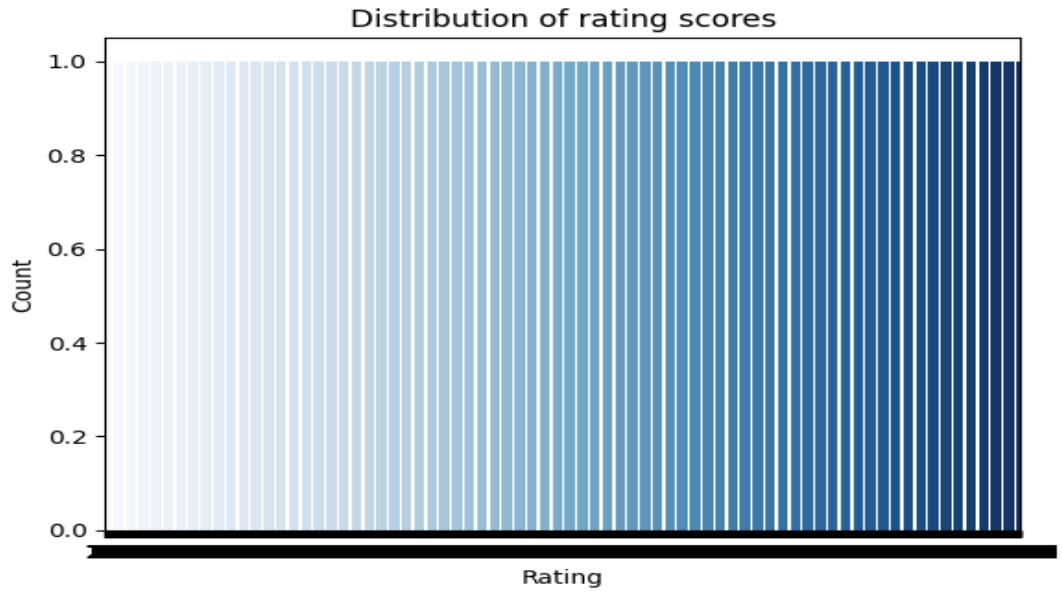


Fig 3.4: Distribution of rating scores

©Daffodil International University

After finding huge imbalance in data to the high rate classes I try to add more data with low rate classes.

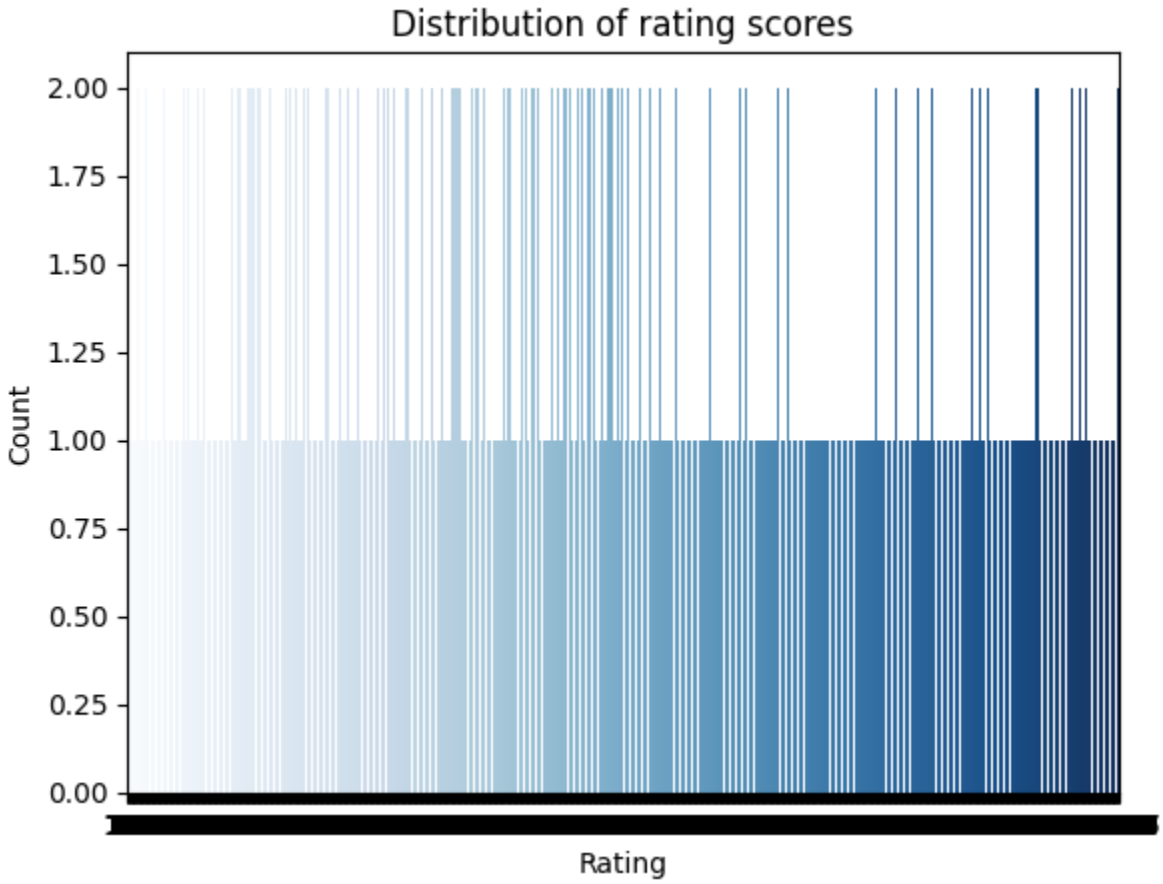


Fig 3.5: Updated distribution of rating scores

©Daffodil International University

3.5.2 Visualize Sentiment Distribution

By plotting the frequency of sentiments from data using positive and negative levels I got an outcome which is visually represented by a pie chart.

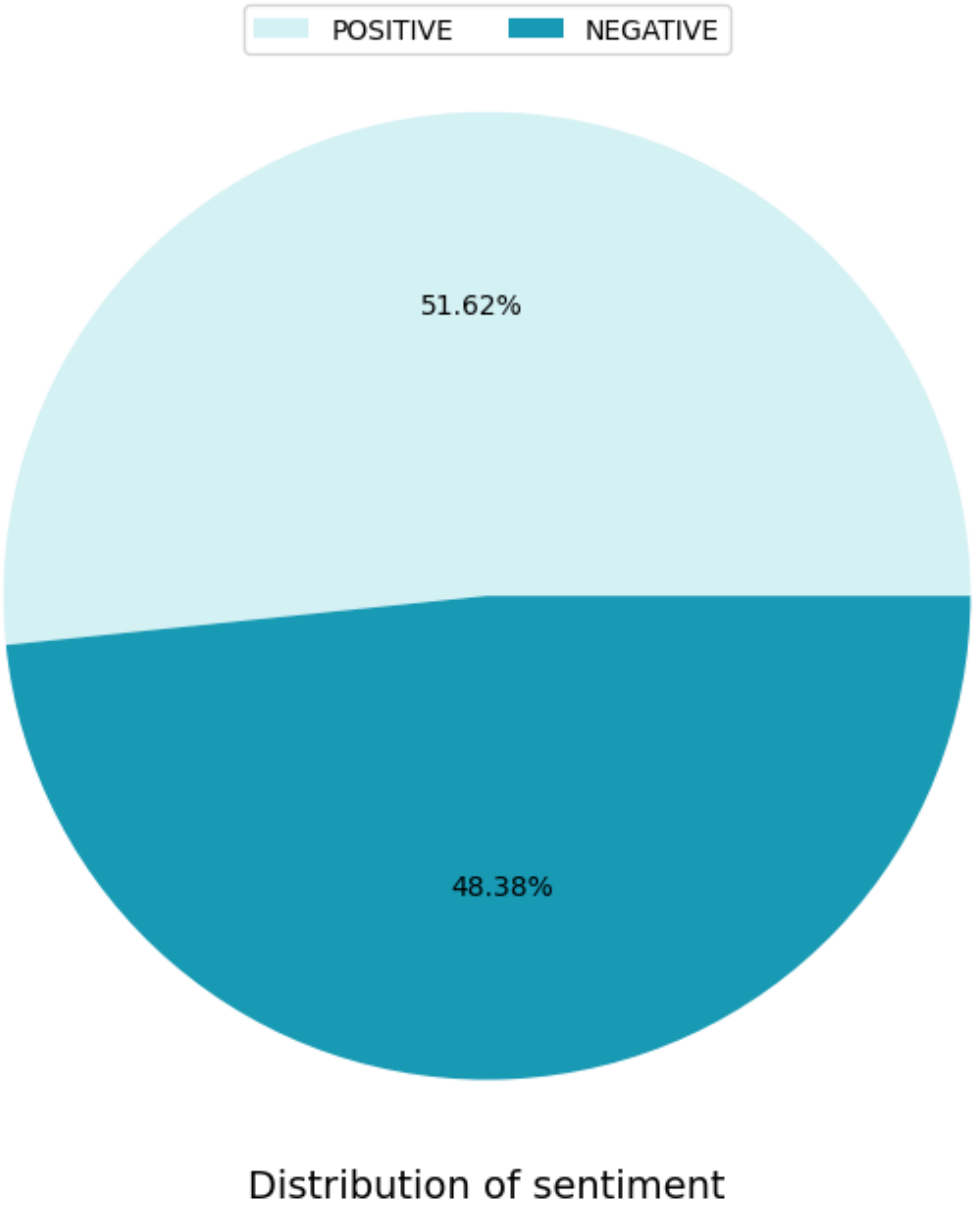


Fig 3.6: Distribution of sentiment

Fig 3.6 shows the Positive and Negative ratings percentage gradually 51.62% and 48.38%.

©Daffodil International University

3.5.3 Word Cloud Visualization

Word clouds are generated to visualize the most used words in all reviews as well as the positive and negative reviews. As seen in Figures 3.7 to 3.9 most used words in all reviews, positive reviews and negative reviews are gradually given.

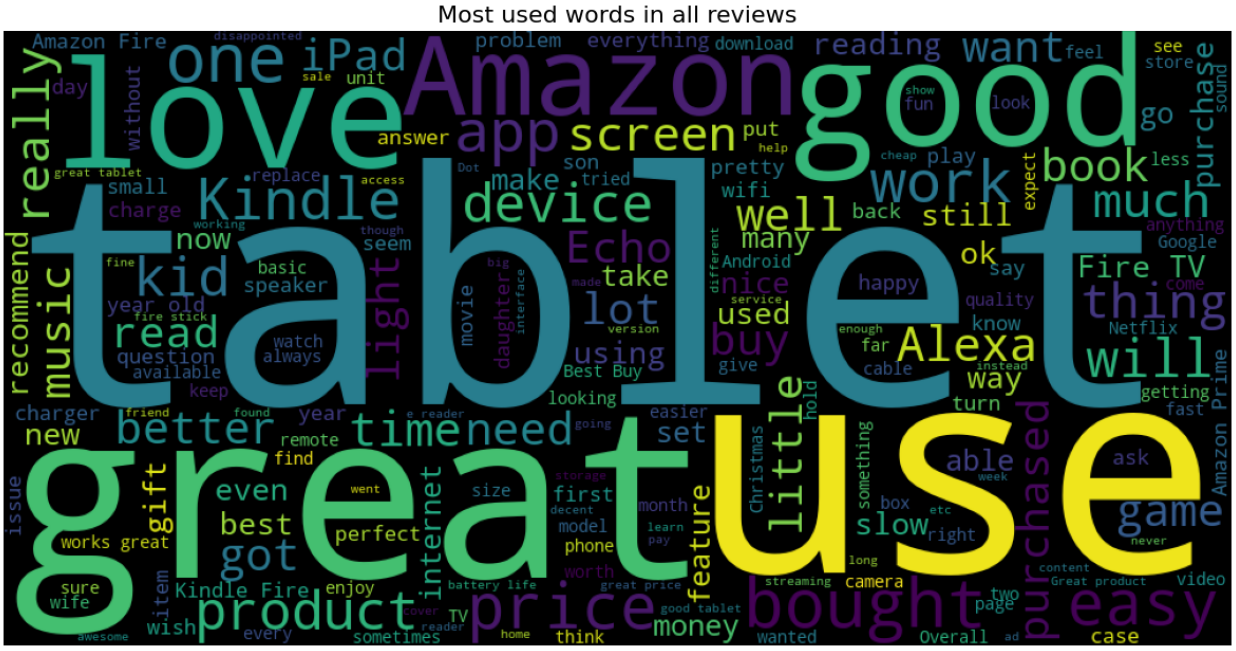


Fig 3.7: Most used words in all reviews

©Daffodil International University

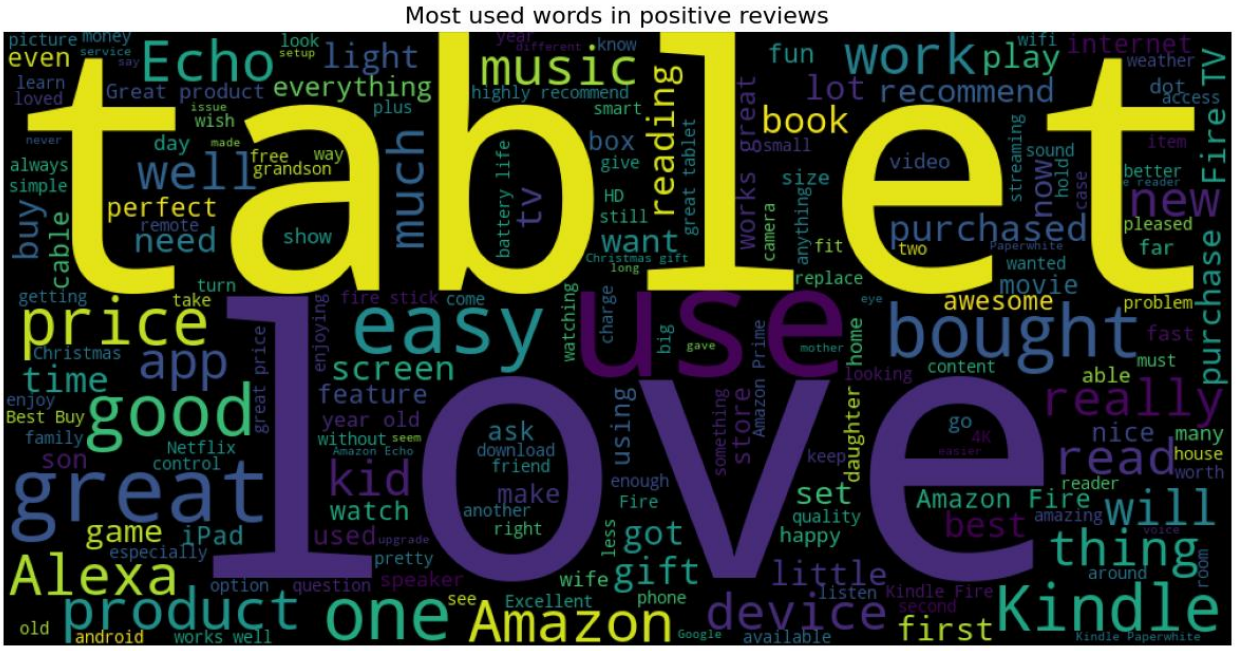


Fig 3.8: Most used words in positive reviews.

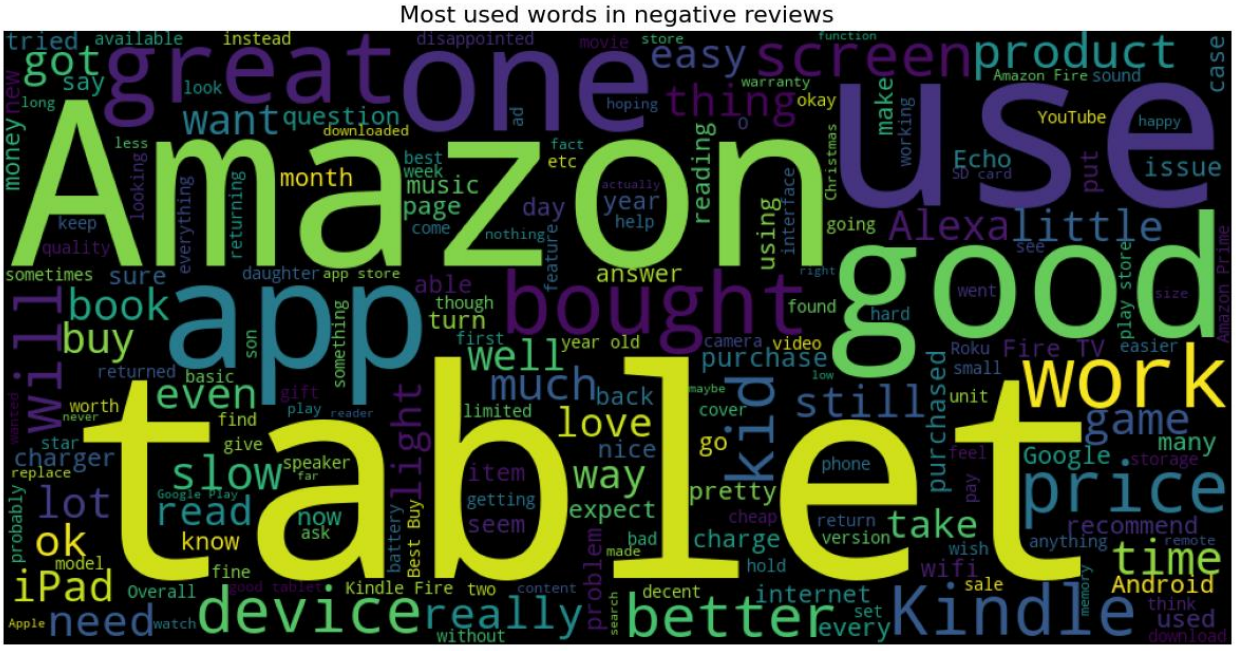


Fig 3.9: Most used words in negative reviews.

©Daffodil International University

3.5.4 Preparing text data for modeling

To prepare or preprocess the text data some steps have been done. Those are cleaning text data, removing unnecessary words, reducing words to their base form by using stemming and lemmatization. For sentiment analysis based on NLP tasks, this pre-processing is necessary because it helps to reduce the dimensionality of the data and improve the performance of the models.

Stop word Removal

Some words in the review sentences frequently occurs such as “a” , ”an”, “the”, “this”, “that”, “is”, “it”, “to”, “and”. So these words need to be removed.

```
# Stop word lists can be adjusted for your problem
stop_words = ["a", "an", "the", "this", "that", "is", "it", "to", "and"]
```

Fig 3.10: Stop Words.

Stemming and Lemmatization

Tokenizing the text into words, applying the snowball stemmer to each word we need to do stemming which reduce the words to their root.

Purpose of Lemmatization is same as stemming where the text need to tokenize inti words, getting the tag for each words, applying the WordNet lemmatize to each word and joining the lemmatized words into single string again.

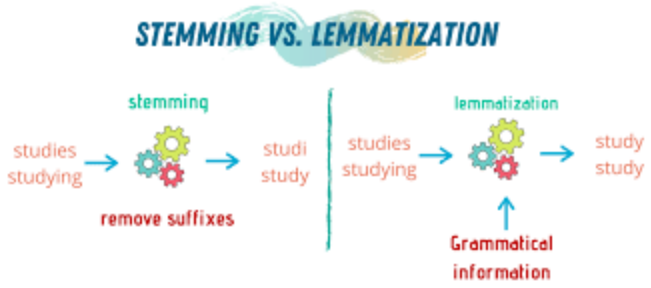


Fig 3.11: Stemming vs Lemmatization

©Daffodil International University

Vectorize text data and Split dataset

To review texts the text processing functions are being applied and by using TF-IDF vectorization the text data is being vectorized to convert the text data into numerical features. After that I split the dataset into train and test set. Then by doing over sampling the positive class to match the number of negative class the training data got balanced. Then I have defined the modeling function.

3.6 Hardware/ Software Requirements

1. Equipment and Hardware/ Software specifications

- Operating System (Windows 7 or above)
- Hard Disk (minimum 1 TB)
- Ram (Minimum 4 GB)

2. Tools

- Python Environment
- PyCharm.
- Google Colab.
- VS code.

©Daffodil International University

CHAPTER 4

Implementation

4.1 Overview

To ensure accuracy, I have to collect the data after the previous steps are finished. some components were required for our project in order to complete the fundamental configuration.

- Data Collection.
- Data preprocessing.
- Word clouds visualization.
- Stop Words remove.
- Text Cleaning.
- Data Preparation
- Tokenization.
- Deploy models for all five algorithms.
- BERT Model
- Ensemble Method (Voting)
- Discuss on the result and accuracy.

Data collection for the research starts with the Amazon product review dataset. Data preparation is then carried out to tidy and get the raw data ready for analysis. The most frequent words in the collection are represented intuitively through the use of word clouds for visualization. The next stage is to remove stop words from the text in order to get rid of common but useless words. After that, punctuation, special characters, and other noise are eliminated during text cleaning to further improve the data. After the data has been cleaned, it is ready for model training, which involves tokenizing the language to create units of data that machine learning algorithms may use. The sentiment of the reviews is then classified using five distinct algorithms: Support Vector Machine

©Daffodil International University

(SVM), XGBoost, Random Forest, Logistic Regression, and Multinomial Naive Bayes. The performance of these models is then assessed by debating their outcomes and accuracy. Furthermore, a BERT (Bidirectional Encoder Representations from Transformers) model is used, utilizing its sophisticated textual context comprehension capabilities. In order to increase the sentiment analysis's overall accuracy and robustness, the predictions from several models are finally combined using an ensemble method that uses voting.

4.2 Train Model

Applied Models

As I mentioned previously I have used 5 machine learning methods such as Naive Bayes, XGBoost, Logistic Regression, Random Forest Classifier, Support Vector Machine. I used many models to see the accuracy level as if they are almost same or not.

1. **Logistic Regression:** For tasks involving classification like deciding between yes or no true or false or 0 and 1 options when dealing with an outcome, with two possibilities statistical regression comes into play. Logistic regression stands out as a form of regression that predicts the chance of an event belonging to a category rather than having a continuous result.

   In regression the model uses the function also known as the sigmoid function to represent how independent variables relate to the probability of a binary outcome. The logistic function ensures that the predicted probability falls within the range of 0 to 1.

©Daffodil International University

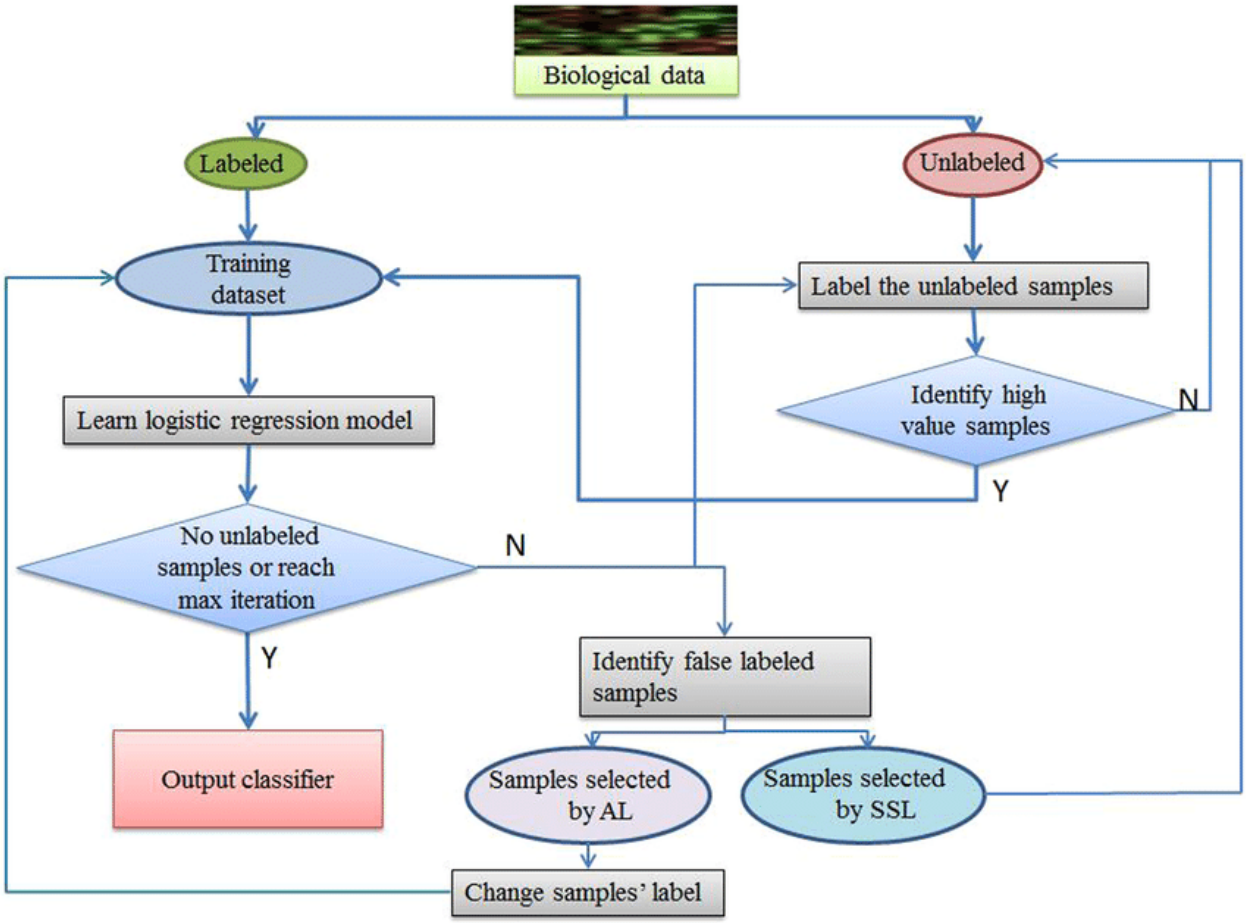


Fig 4.1: Logistic Regression work process.

2. **Random Forest:** Regression and classification may both be performed using the tree-based method of the RF Classifier. An algorithm for machine learning creates a hierarchical tree. This method is used by artificial intelligence to create hierarchical "decision trees". The random forest approach for classification creates several decision trees using a combination technique, which are then averaged. This makes the issues related to overfitting clearer. Machine learning is one of the hottest subjects in business right now since it can be used in any scenario, anyplace there is a lot of data. Because RF Classifier based on ML offers a number of benefits over alternative techniques, it is highly favored. When the approach was initially described in 1997, it was intended to work with huge datasets.

©Daffodil International University

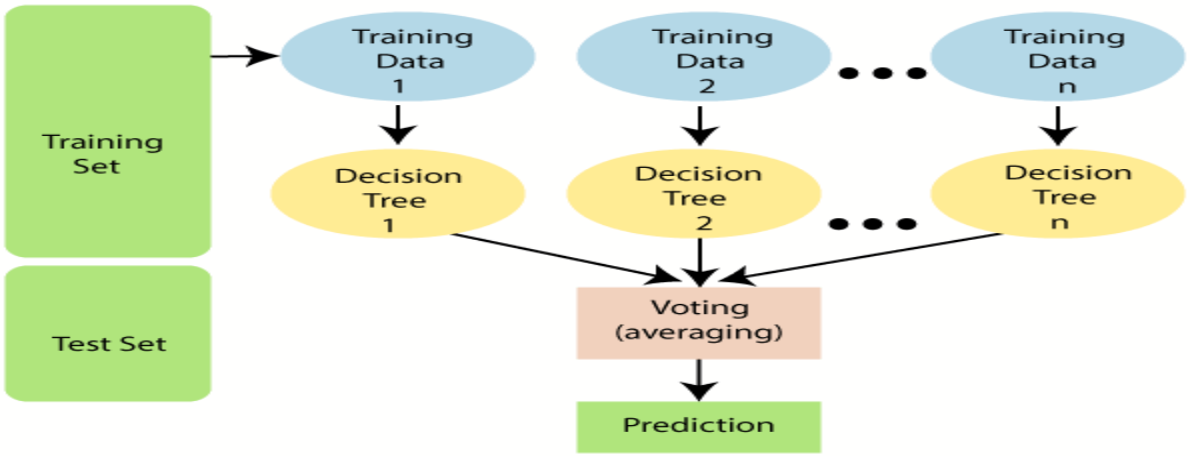


Fig 4.2: Random Forest Classifier Work process.

3. **SVM:** Support Vector Classifier, or SVM, is the acronym for it. SVC is by necessity a probabilistic clustering approach since it makes no presumptions about the fundamental foundation of the data, such as the amount of groups and the size of each one. As a result, you will often need to perform some prior manufacturing, such principal component analysis, if your data is extremely dimensional. Previous research has shown that it performs well with low-dimensional data. Numerous changes have been made to the original method that offer particular methods for computing the clusters by examining a portion of the regions contained in the connection matrices. Multiple updates are accessible.

©Daffodil International University

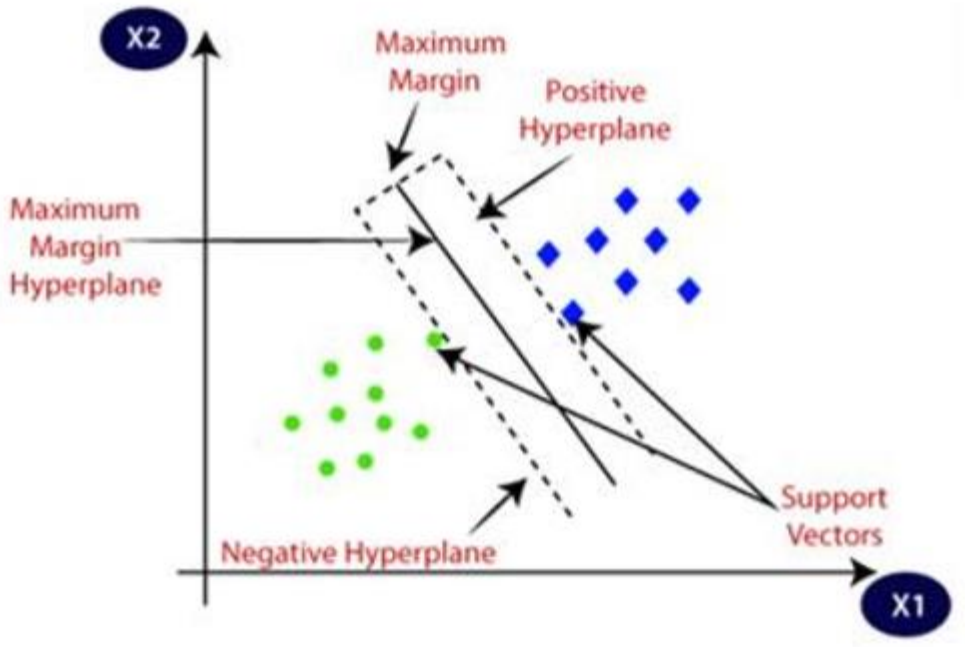


Fig 4.3: Figure of SVM.

4. **Naïve Bayes:** Because naïve Bayes model for machine learning can handle large amounts of data, it is suggested to use it even when working with millions of records. When it comes to tasks involving natural language processing (NLP) like abusive text analysis, it excels. Sorting is an easy and quick procedure. Before we can understand the naive Bayes classifier, we need to understand the Bayes theorem. First, let's look at the Bayes Hypothesis. For this theorem, conditional probability is the foundation. Conditional probability is the chance of an event occurring given the possibility of another event occurring. Conditional probability and our prior knowledge can be used to calculate the likelihood of an event. Naive Bayes algorithms are widely used in many fields, including spam filtering, sentiment analysis, and recommendation systems. Although they are quick and simple to use, their main disadvantage is the requirement for independent predictors. When the forecasting parameters are reliant, as they are in the majority of real-world applications, the classifier perform badly.

©Daffodil International University

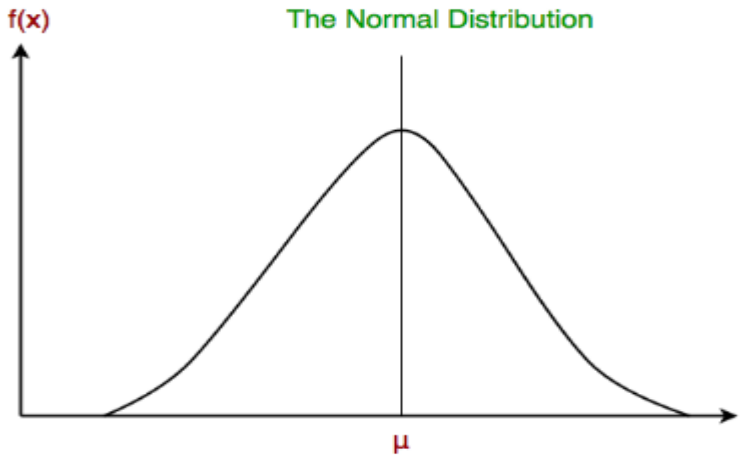


Fig 4.4: How Naive Bayes Functions works.

5. **XGBoost:** 1. Extreme gradient booster, or XGBoost, is a well-liked and successful machine learning technique that may be used to address regression and classification problems. Using this technique, gradient enhanced decision trees may be easily and successfully created. XGBoost is a method of ensemble learning that combines the predictions of several weak learners—often decision trees—to strengthen a model. As they proceed, one weak student fixes the shortcomings of the one before them, educating the underprivileged students one by one. To prevent overfitting, XGBoost has many normalization penalties. Penalty-based regularizations result in effective training, which enables the model to achieve an appropriate degree of generalization. The non-linear: Non-linear data structures may be recognized and learned from using XGBoost. Cross-checking: assembled and ready to use right out of the box.

©Daffodil International University

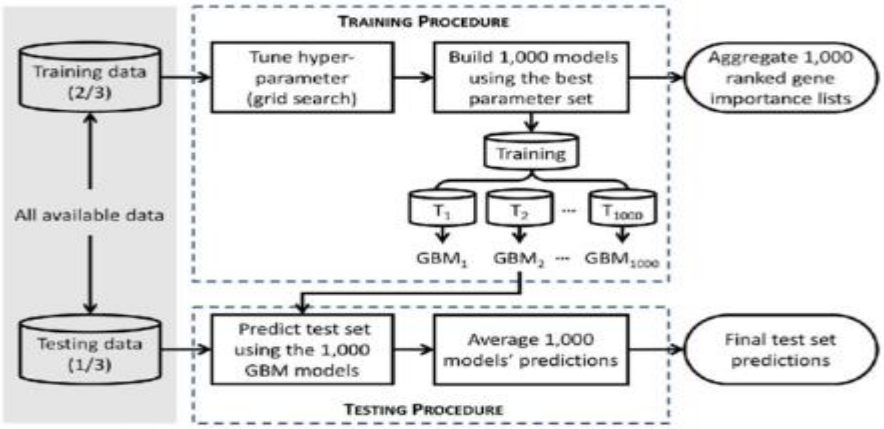


Fig 4.5: XGBoost model architecture.

6. **BERT:** Bidirectional Encoder Representations from Transformers is a model. It is a state-of-the-art machine learning model used for natural language processing (NLP) tasks. BERT is unique in its ability to consider context from both directions (left-to-right and right-to-left) simultaneously, making it highly effective at understanding the nuances of language. It is pre-trained on a massive corpus of text and can be fine-tuned for specific tasks such as sentiment analysis, question answering, and text classification, achieving state-of-the-art results in many NLP benchmarks.

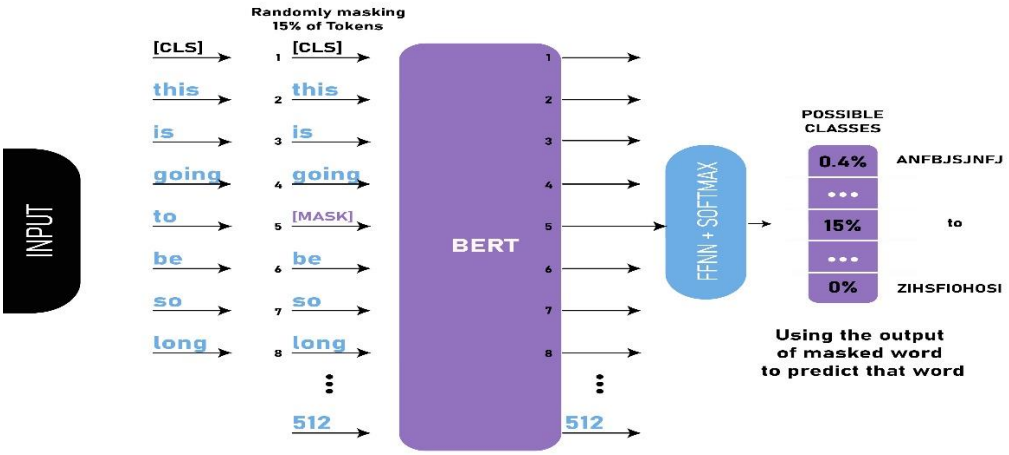


Fig 4.6: BERT model architecture.

©Daffodil International University

7. **Ensemble Method:** The Ensemble Method, specifically through voting, aggregates predictions from multiple individual models to improve overall prediction accuracy and robustness in sentiment analysis. By combining diverse models such as Naïve Bayes, XGBoost, Logistic Regression, Random Forest, SVM, and BERT, each contributing with its unique strengths and perspectives, the ensemble leverages their collective wisdom to make a final prediction. This approach typically results in higher accuracy than any single model alone, as it mitigates the weaknesses of individual models while amplifying their strengths.

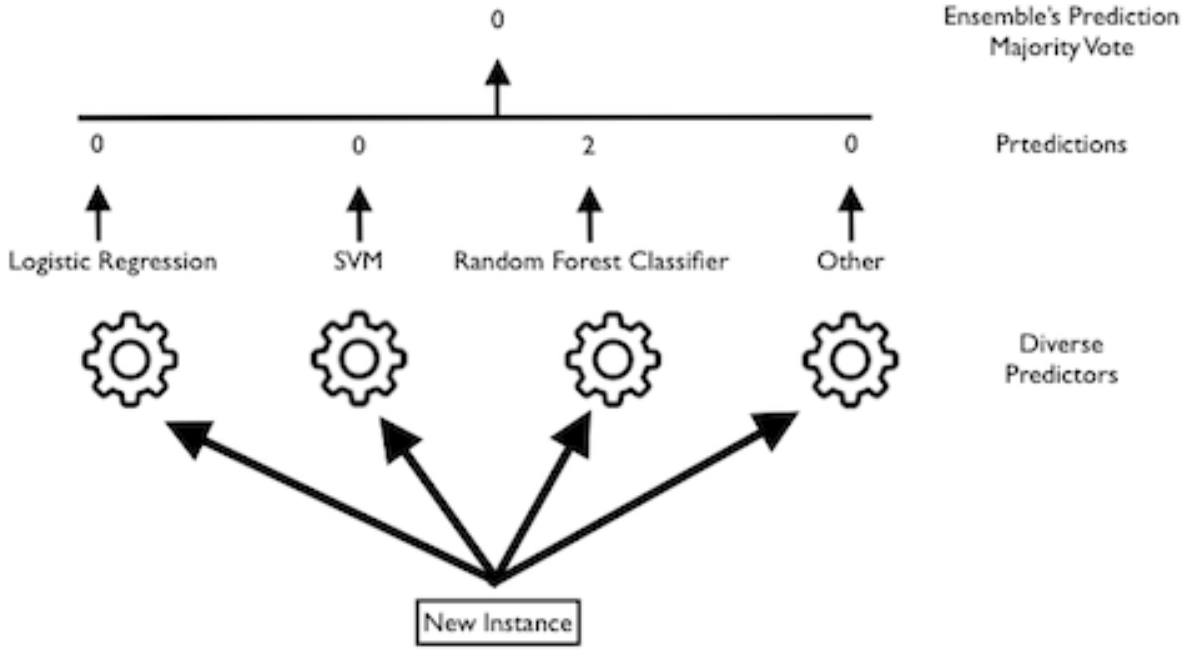


Fig 4.7: Ensemble Method (Voting) architecture.

©Daffodil International University

CHAPTER 5

Experiment Results and Discussion

5.1 Experimental Result

We understand that no form of technology can provide flawless outcomes. Similarly, we may adjust the parameters of our model during training to improve accuracy. But using a range of methods, we find that the level of precision is rather high. This is a graphic summary of the work we have been doing. The data shown in these visualizations includes a heat maps, recall, precision, f1 score, and supports.

5.2 Classification and Descriptive Analysis using ML Models

The technique we employed determined the various outcomes we obtained. We were able to accurately identify the ratings using five machine learning techniques. These methods were used to assess each component of the overall structure's relative efficacy, and several verification procedures were then employed in arriving at the forecast's outcome. Metrics like the total F1 score, accuracy, precision, and recall provide a thorough picture of the algorithms' efficiency.

Table 5.1: Accuracy Table.

Classifier	Accuracy Score (AUC)
Naïve Bayes	79.50%
XGBoost	82.02%
Logistic Regression	80.54%
Random Forest	82.86%
SVM (classifier)	83.86%
BERT	73.78%
Ensemble Method	83.55%

©Daffodil International University

This section displays the performance of several classifiers. PyCharm and CoLab were two free tools that were used during the entire process. In all, five classifiers were used: Niave Bayes, XGBoost, Logistic Regression, Random Forest, and Support Vector Classifier and BERT too

	precision	recall	f1-score	support
0	0.82	0.79	0.80	218
1	0.77	0.80	0.79	196
accuracy			0.79	414
macro avg	0.79	0.80	0.79	414
weighted avg	0.80	0.79	0.79	414

AUC 0.7950056169256692

Fig 5.1: Naive Bayes classification reports.

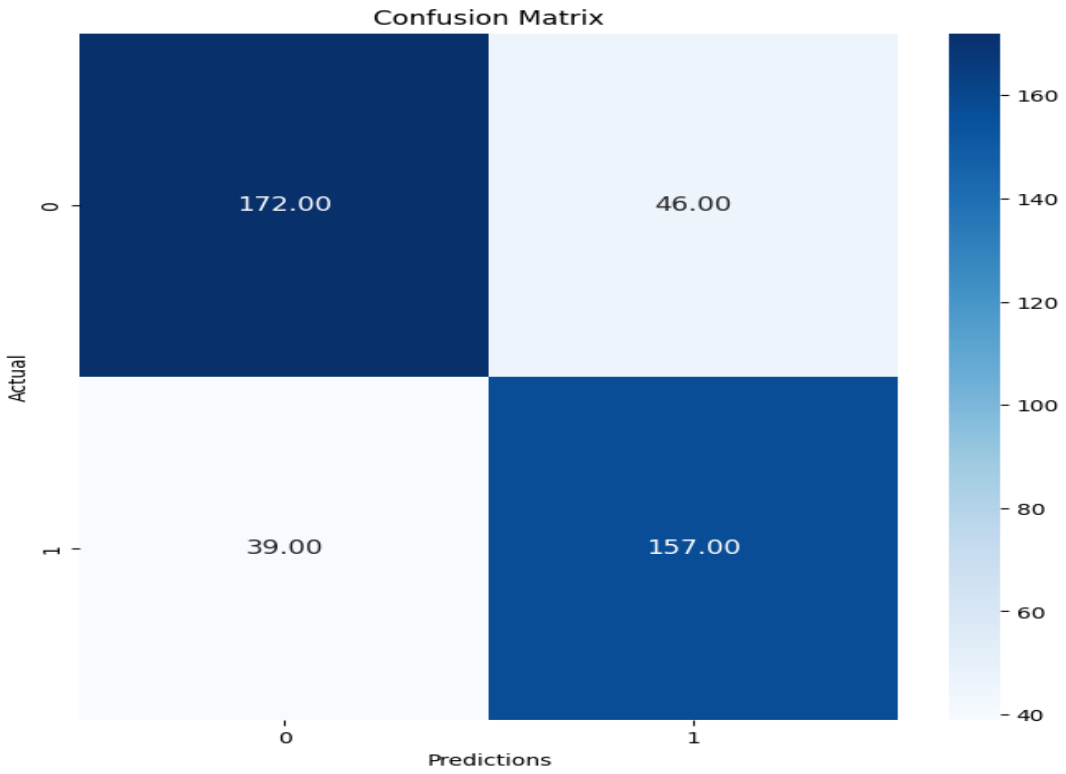


Fig 5.2: Naive Bayes Confusion matrix.

©Daffodil International University

	precision	recall	f1-score	support
0	0.85	0.79	0.82	218
1	0.79	0.85	0.82	196
accuracy			0.82	414
macro avg	0.82	0.82	0.82	414
weighted avg	0.82	0.82	0.82	414

AUC 0.8202583785807902

Fig 5.3: XGBoost classification report

.

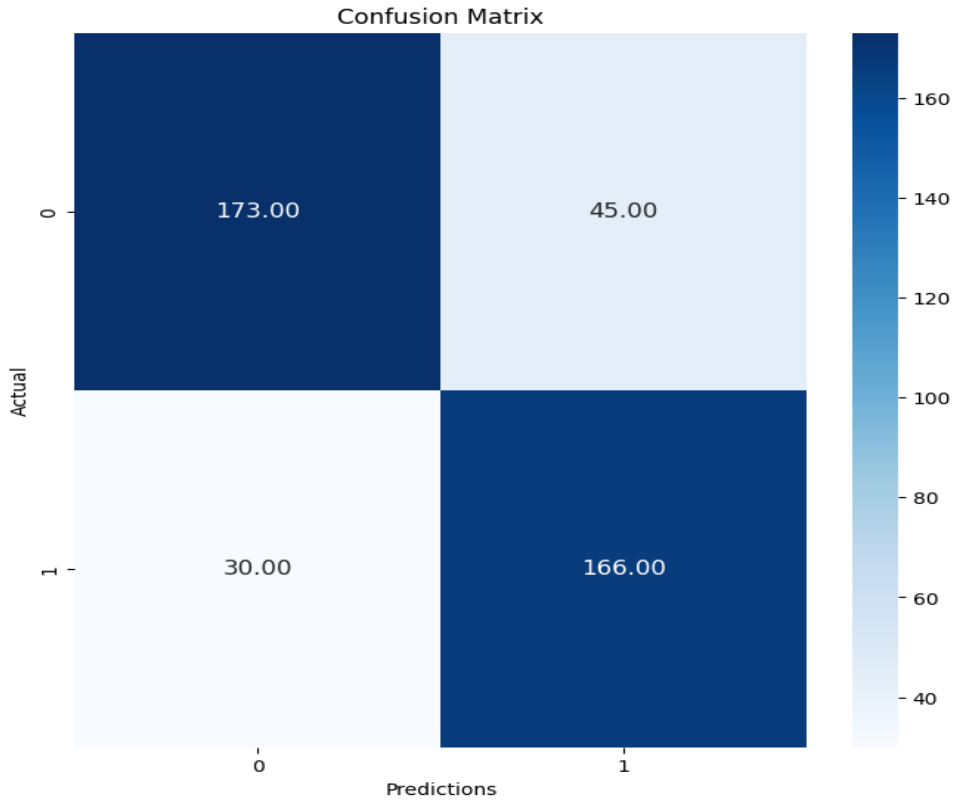


Fig 5.4: XGBoost confusion matrix

©Daffodil International University

	precision	recall	f1-score	support
0	0.81	0.83	0.82	218
1	0.81	0.78	0.79	196
accuracy			0.81	414
macro avg	0.81	0.81	0.81	414
weighted avg	0.81	0.81	0.81	414

AUC 0.8054437371278788

Fig 5.5: Logistic Regression classification report

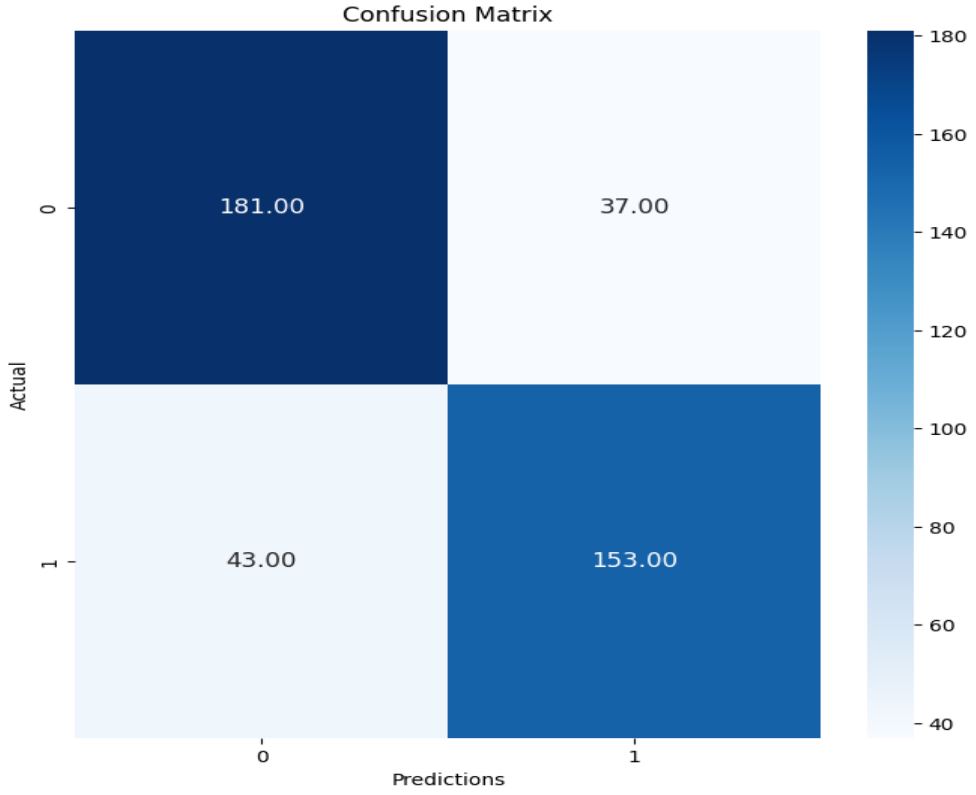


Fig 5.6: Logistic Regression confusion matrix

©Daffodil International University

	precision	recall	f1-score	support
0	0.88	0.78	0.83	218
1	0.78	0.88	0.83	196
accuracy			0.83	414
macro avg	0.83	0.83	0.83	414
weighted avg	0.83	0.83	0.83	414

AUC 0.8286837670848156

Fig 5.7: Random Forest classification report

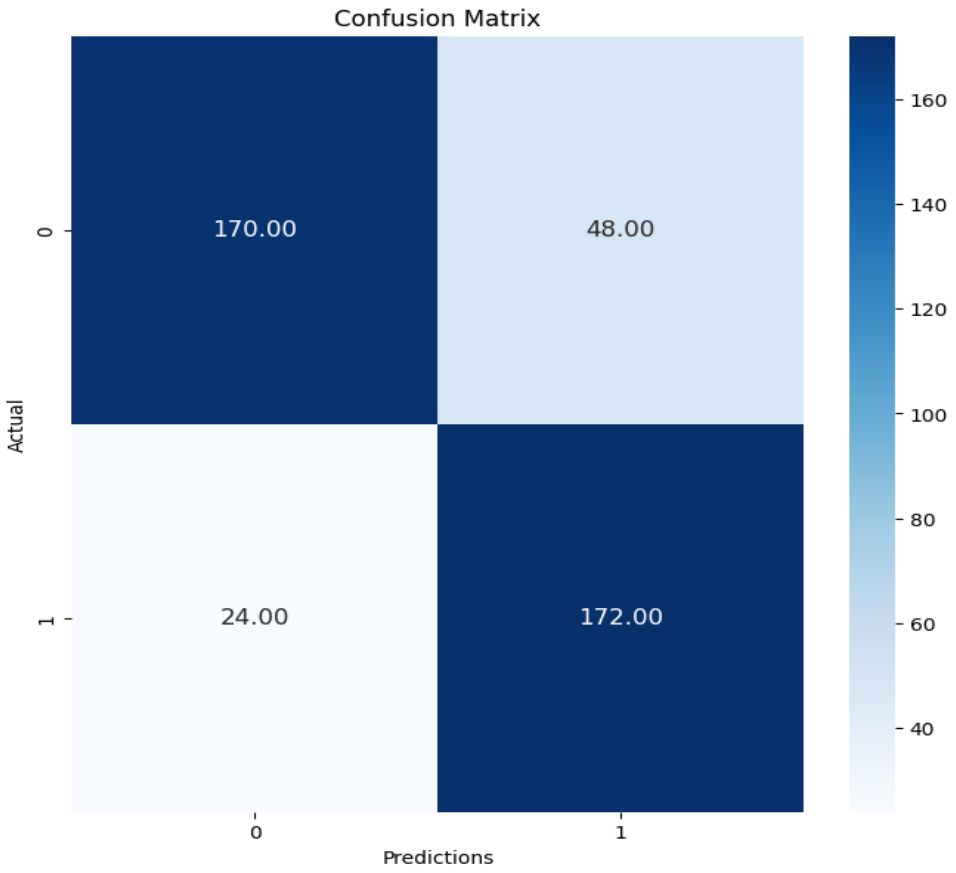


Fig 5.8: Random Forest confusion matrix

©Daffodil International University

```
Model: SVC
              precision    recall  f1-score   support

           0       0.89      0.78      0.83       218
           1       0.79      0.89      0.84       196

    accuracy                           0.84       414
   macro avg       0.84      0.84      0.84       414
weighted avg       0.84      0.84      0.84       414

AUC: 0.8386304062909566
```

Fig 5.9: SVM classification report

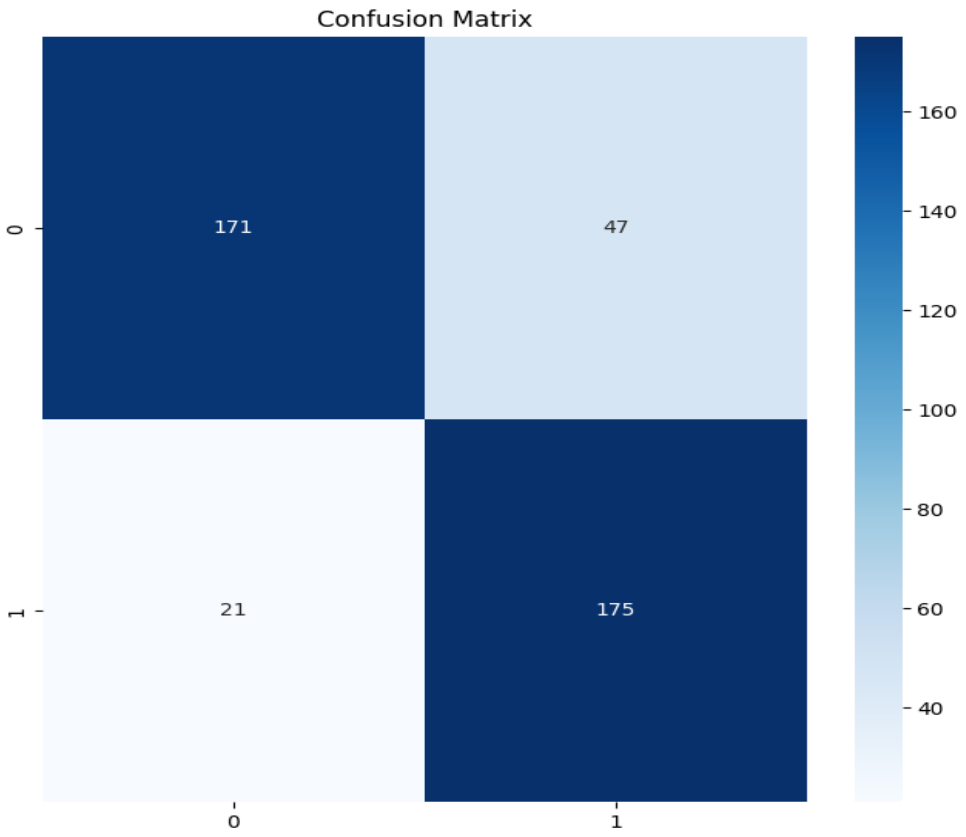


Fig 5.10: SVM confusion matrix

©Daffodil International University

	precision	recall	f1-score	support
0	0.75	0.74	0.74	1067
1	0.72	0.74	0.73	1000
accuracy			0.74	2067
macro avg	0.74	0.74	0.74	2067
weighted avg	0.74	0.74	0.74	2067

AUC: 0.7378223992502342

Fig 5.11: BERT report

	precision	recall	f1-score	support
0	0.82	0.89	0.85	218
1	0.86	0.79	0.82	196
accuracy			0.84	414
macro avg	0.84	0.84	0.84	414
weighted avg	0.84	0.84	0.84	414

AUC 0.8355176933158585

Fig 5.12: Ensemble Method report

5.3 Discussion

In this work, we will apply predictions ML approaches to predict the positive and negative ratings. In any field of study, each word should be given significant weight throughout the classification process. By applying many machine learning techniques to the averages and dependability ratings, we were able to accomplish our aim. We employed a total of five distinct algorithms for this project.

©Daffodil International University

Before we began our work, there were quite a few items we needed to find. We did start working on the algorithm after selecting it. We then assessed the accuracy of each method. The SVM classifier, which acts as the predictor for the next five models, I able to achieve the greatest accuracy of 83.86% with the use of approaches. Then BERT has given 73.78% accuracy. After doing ensemble prediction by combining all the models I get 83.55% accuracy.

Precision: Accuracy is the ability of a way to discern real successes from all predicted ones. It is possible to derive the following formula:

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall is the ability of the model to discriminate between each true positive reaction and the predicted affirmative response. The following is an explanation of the equation's structure:

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: When recall and reliability are combined, the F1 score provides a useful assessment of the model's performance. It may be calculated using this formula:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Accuracy: Accuracy is defined as the proportion of all reported cases and correctly anticipated occurrences in a collection of data. calculated as follows:

$$Accuracy = \frac{TP + TN}{Total\ Instances}$$

©Daffodil International University

CHAPTER 6

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

6.1 Impact on Life

Sentiment analysis plays a pivotal role in enhancing consumer experiences and thereby impacting lives positively. With SVM and Ensemble Method models achieving higher accuracies in your thesis, businesses can better understand and respond to customer sentiments. This understanding enables personalized marketing strategies that cater to individual preferences and needs more effectively. Customers benefit from receiving tailored recommendations and offers that resonate with their emotions and aspirations, leading to increased satisfaction and engagement with brands. By leveraging sentiment analysis insights, companies can foster deeper connections with their audience, potentially improving overall quality of life by reducing decision-making stress and enhancing shopping experiences.

6.2 Impact on Society & Environment

The societal and environmental impacts of accurate sentiment analysis are profound. High-accuracy models like SVM and Ensemble Method enable businesses to optimize their marketing efforts, minimizing the societal impact of excessive and irrelevant advertisements. By targeting consumers more precisely based on their sentiments and preferences, businesses reduce wasteful resource consumption associated with broad, indiscriminate marketing campaigns. This targeted approach not only conserves resources but also promotes a more sustainable marketing ecosystem where consumer needs drive business decisions. Moreover, by reducing unnecessary consumption and enhancing consumer satisfaction, accurate sentiment analysis contributes to a societal shift towards mindful consumerism and sustainable practices, aligning economic goals with environmental responsibility.

©Daffodil International University

6.3 Ethical Aspects

Ethical considerations in sentiment analysis are crucial given the sensitive nature of consumer data and the potential impact on individuals' privacy and autonomy. As SVM and Ensemble Method models deliver higher accuracies, ensuring ethical practices becomes even more imperative. Transparency in data collection, usage, and storage is essential to maintain consumer trust. Businesses must prioritize fairness by mitigating biases in algorithms and ensuring that decisions based on sentiment analysis do not discriminate or disadvantage any group. Respecting consumer consent and privacy rights is paramount, with stringent adherence to data protection regulations and ethical guidelines. By upholding these ethical standards, businesses can leverage sentiment analysis responsibly, fostering trust and accountability while safeguarding individual rights in an increasingly digital and data-driven marketplace.

6.4 Sustainability Plan

A sustainable approach to sentiment analysis encompasses proactive measures to uphold ethical standards and minimize environmental impact. This involves integrating ethical considerations into AI and machine learning practices, ensuring algorithms are continuously monitored and refined to mitigate biases and uphold transparency. Businesses should implement robust data security measures to protect consumer information and comply with regulatory frameworks. Adopting best practices in data handling and respecting consumer privacy rights are foundational to a sustainable sentiment analysis strategy. Moreover, promoting responsible consumption through targeted marketing efforts reduces societal and environmental footprints associated with excessive consumerism. By aligning business objectives with ethical principles and sustainability goals, companies can harness the power of sentiment analysis to not only drive profitability but also contribute positively to societal well-being.

©Daffodil International University

CHAPTER 7

Conclusion and Future Research

7.1 Summary of the Study

The study sought to examine the effectiveness of five classification techniques for text data sentiment analysis. Among these techniques were Naive Bayes, XGBoost, Random Forest, Support Vector Machine (SVM), Logistic Regression, BERT and Ensemble method. The data was cleaned up, word counts were reduced, and it was standardized in preparation for analysis.. Each classifiers performance was assessed using accuracy as the measure along with precision, recall, F1 score and ROC AUC score. The outcomes displayed the precision as the validation and training accuracies, for each classifier enabling a comparison of their effectiveness. The research concluded by suggesting a sentiment analysis classifier tailored to the datasets needs after considering the pros and cons of each model. These results could potentially lead to exploration and real world implementations, in fields.

7.2 Conclusion

I investigated sentiment analysis with text data in our study. evaluated the performance of five distinct classifiers, BERT and Ensemble method. After comparing and analyzing them, I found that SVM and Ensemble method was the most effective classifier for determining sentiment in text. Its exceptional performance on metrics like as training, validation, and accuracy demonstrates its capacity to pick up on minute emotional nuance within the dataset. Additionally, SVM demonstrated its strength in handling problems including sentiment classification with balanced precision, recall, F1 score, and ROC AUC score. My study focused on the importance of preprocessing techniques in enhancing the quality of text data for sentiment analysis, in addition to comparing classifiers. Lemmatization, stopword removal, and cleaning were among of the techniques that helped increase the signal to noise ratio.

In essence, this study serves as a stepping stone towards enhancing sentiment analysis methodologies, paving the way for more accurate, reliable, and scalable solutions in deciphering sentiment from textual data.

©Daffodil International University

7.3 Future Work

While conducting this research, I've learned a lot and will continue to learn. Marketing strategy can be improve more by using vast amount of dataset. In future, I would aim to do this study with survey based dataset and it will increase getting the chances of higher accuracy. The survey will have done for the purpose of getting audio, video and image based data. It will be beneficial for the consumers to purchase products as well as the sellers to improve their product marketing strategies in near future.

©Daffodil International University

References

[1] Walaa Medhat, Ahmed Hassan, Hoda Korashy, Sentiment analysis algorithms and applications: a Asurvey, Ain Shams Engineering Journal, Volume 5, Issue 4, 2014

[2] Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif Intell Rev* **55**, 5731–5780 (2022).

[3] Fang, X., Zhan, J. Sentiment analysis using product review data. *Journal of Big Data* **2**, 5 (2015)

[4] Nandwani, P., Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **11**, 81 (2021)

[5] Kashfia Sailunaz, Reda Alhajj, Emotion and sentiment analysis from Twitter text,Journal of Computational Science, Volume 36, 2019

[6] T. U. Haque, N. N. Saber and F. M. Shah, "Sentiment analysis on large scale Amazon product reviews," *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, Bangkok, Thailand, 2018

[7] Jagdale, R.S., Shirsat, V.S., Deshmukh, S.N. (2019). Sentiment Analysis on Product Reviews Using Machine Learning Techniques. In: Mallick, P., Balas, V., Bhoi, A., Zobaa, A. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 768. Springer, Singapore.

[8] Z. Singla, S. Randhawa and S. Jain, "Sentiment analysis of customer product reviews using machine learning," *2017 International Conference on Intelligent Computing and Control (I2C2)*, Coimbatore, India, 2017

[9] C. Chauhan and S. Sehgal, "Sentiment analysis on product reviews," *2017 International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida, India, 2017

[10] Wassan, Sobia, et al. "Amazon product sentiment analysis using machine learning techniques." *Revista Argentina de Clínica Psicológica* 30.1 (2021): 695.

©Daffodil International University

[11] Yi, Shanshan, and Xiaofang Liu. "Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review." *Complex & Intelligent Systems* 6, no. 3 (2020): 621-634.

[12] Nandal, Neha, Rohit Tanwar, and Jyoti Pruthi. "Machine learning based aspect level sentiment analysis for Amazon products." *Spatial Information Research* 28, no. 5 (2020): 601-607.

[13] G. Chauhan, A. Sharma and N. Dwivedi, "Amazon Product Reviews Sentimental Analysis using Machine Learning," *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Greater Noida, India, 2024.

[14] Thanu, Velam, and David Yoon. "Sentiment Analysis of Product Reviews on Social Media." In *Advances in Software Engineering, Education, and e-Learning: Proceedings from FECS'20, FCS'20, SERP'20, and EEE'20*, pp. 899-907. Springer International Publishing, 2021.

©Daffodil International University

Report.docx

ORIGINALITY REPORT

23%	21%	8%	18%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to University of Southern Mississippi Student Paper	4%
3	Submitted to Daffodil International University Student Paper	3%
4	Submitted to University of Strathclyde Student Paper	1%
5	ijsrcseit.com Internet Source	1%
6	ijarsct.co.in Internet Source	1%
7	link.springer.com Internet Source	<1%
8	www.ijert.org Internet Source	<1%
9	Xing Fang, Justin Zhan. "Sentiment analysis using product review data", Journal of Big	<1%

©Daffodil International University