Sentiment Analysis of Amazon Reviews

PROJECT REPORT

Submitted by

ARYA ANILKUMAR(AEC18CS009)
NEERAJA.S.RAJ(AEC18CS014)
R.P.ABHIJITH(AEC18CS016)
VISHNUMAYA(AEC18CS020)

in partial fulfillment for the award of the degree

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
COLLEGE OF ENGINEERING ARANMULA,
ARANMULA, PATHANAMTHITTA, 689533.

JUNE 2022

Acknowledgements

To bring something into existence is truly the work of ALMIGHTY. Thank GOD ALMIGHTY for making this venture a success. Express our wholehearted thanks Co-oprative Academy of Professional Education for providing an opportunity to do our studies in this esteemed institution. Thanks to our Principal Dr. Indhu P Nair for providing the facilities for our studies and constant encouragement in all achievements. At the outset wish to place on record our sincere thanks to quite a few people without whose help, this venture could not have been a success. Would like to express profound gratitude to our head of the department Prof.Siju Koshy for his encouragement and for providing all the facilities for presenting our project and project coordinator Asst. Prof.Muhammad Azharudeen Sahib and project guide Asst. Prof.Ms.Binitha S who provided the necessary guidance and serious advice to present our project. Also extend our sincere thanks to all staff members of Computer Science Engineering Department who guided us throughout the course.

ARYA ANILKUMAR
NEERAJA.S.RAJ
R.P.ABHIJITH
VISHNUMAYA

DECLARATION

*We hereby declare that the the project report "**SENTIMENT ANALYSIS OF AMAZON REVIEWS**", submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University,Kerala is a bonafide work done by us under the supervision of Ms. Binitha S. This submission represents our ideas in our own words and where ideas or words of others have been included. We have adequately accurately cited and referenced the original sources. We also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for evoke penal action from the sources which have thus not been obtained properly cited from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University .*

Signature :

Place : Aranmula

Date :

ARYA ANILKUMAR

NEERAJA.S.RAJ

R.P.ABHIJITH

VISHNUMAYA

CERTIFICATE

This is to certify that the project report entitled: ”SENTIMENT ANALYSIS OF AMAZON PRODUCT REVIEWS” submitted by ***Arya Anilkumar(AEC18CS009),Neeraja.S.Raj(AEC18CS014),R.P.Abhijith(AEC18CS016),Vishnumaya(AEC18CS020)*** *to the APJ Abdul Kalam Technological University, for the award of the Degree of Bachelor of Technology in Computer Science and Engineering is a bonafide record of the project work carried out under the guidance of Ms.Binitha S.The content of the project report, in full or parts have not been submitted to any other Institute or University for the award of any other degree or diploma.*

Research Supervisor:
Ms.BINITHA S
Assistant Professor
Dept. of Computer Science
College of Engineering
Aranmula.

Head of the Department
Mr.SIJU KOSHY
Assistant Professor
Dept. of Computer Science
College of Engineering
Aranmula.

Place: Aranmula

Date:

Office Seal

Contents

Abstract

Ecommerce has recently grown at a breakneck pace. As a result, online shopping has increased, which has resulted in an increase in product reviews from customers. Customers' purchasing decisions are heavily influenced by the inferred opinions in customer reviews. Because a customer's view of a product is influenced by the opinions of other customers, a sentimental analysis approach is used to categorise these opinions as negative, positive, or neutral.Text mining and computational linguistics studies have recently become increasingly interested in sentiment analysis for product reviews.We're looking to see if there's a link between Amazon product reviews and consumer ratings.Traditional machine learning algorithms, as well as Naive Bayes analysis, are used here. Support Vector Machines, the K-nearest neighbour approach, and deep neural networks like the Recurrent Neural Network are all examples of neural networks like the Recurrent Neural Network are all examples of deep neural networks. We could also operate as a supplement to existing fraud scoring detection approaches by comparing these results.

Abbreviations

P(A/B)	Posterior Probability
P(B/A)	Likelihood Probability
P(A)	Class Prior Probability
P(B)	Predictor Prior Probability
NLP	Natural Language Processing
TF	Term Frequency
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
BoW	Bag of Words
CBOW	Continuous Bag of Words
SG	Skip-gram

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 GENERAL BACKGROUND

Sentiment analysis is a type of natural language processing that examines a text and determines the attitude that underpins it.The classification of emotions(positive, negative and neutral) in text data using text analysis techniques is known as sentiment analysis.It is an important part of dealing with clients on companies online portals and websites.They do it all the time to categorise a comment as a question,a complaint,a suggestion,an opinion, or simply a love for a product.This allows them to quickly filter through the comments or queries, prioritise what has to be addressed first, and even arrange them in a more appealing manner.

Through the availability of products within reach of clients, e-commerce is gaining traction in our digitalized world. People are depending more and more on the opinions of others customers.Several real-world applications necessitate sentiment analysis in order to obtain extensive investigate information. Discover which components or attributes of a product appeal to buyers in terms of product quality, for example, through product analysis. Amazon is one of the world's largest e-commerce companies, with a diverse selection of products including books, medications, medical equipment, ornaments, and gadgets. Customers can rate products on a scale of 1 to 5 stars and leave text reviews on Amazon.Customers can also rate other customers' review as helpful or unhelpful based on their own experiences.Often,these reviews are fraudulent because the merchant score.This calls into doubt the product's authenticity,as customers cannot rely on the product reviews.Our algorithm will do this by analysing Amazon product reviews, classifying the language into comparable categories, and extracting the buyer's sentiments [7].

1.2 EXISTING SYSTEM

The sentimental analysis can be done using various approaches.The most common approaches are the Lexicon based Approach,Hybrid Approach and the Machine Learning Approach.

Here, we are considering the machine learning approaches to perform the analysis of reviews.Machine Learning approach of sentimental analysis can either be accomplished

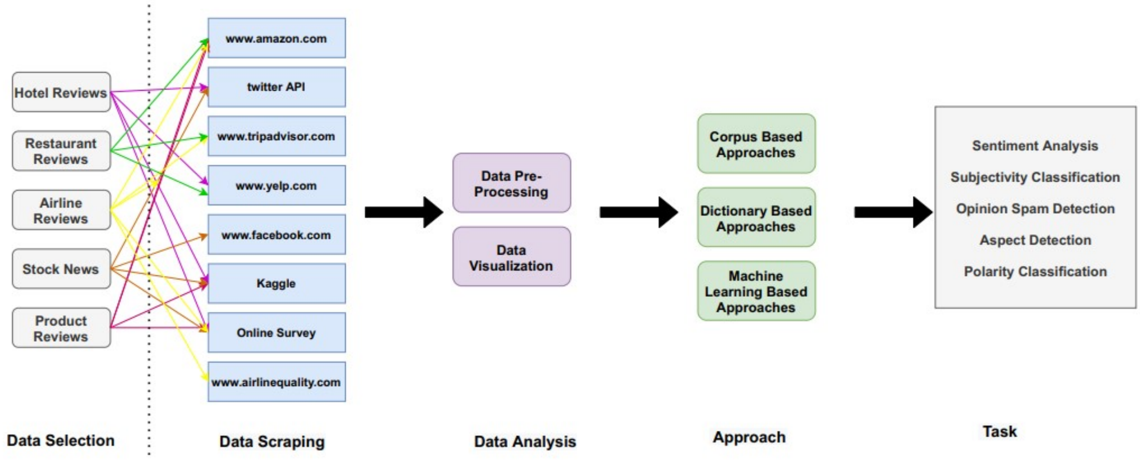


Figure 1.1: General procedure of sentimental analysis

by Supervised Learning method or Unsupervised Learning method.Machine learning allows computers to learn new skills without having to be explicitly programmed. Sentiment analysis algorithms may be trained to understand context, sarcasm, and misapplied terms in addition to straightforward definitions. The most commonly used algorithms to conduct sentimental analysis are Naive-Bayes algorithm, Support Vector Machine, K-Nearest Neighbours,Decision Trees,Logistic Regression, etc. [1]

1.2.1 NAIVE BAYES CLASSIFIER

The Naive Bayes method is a set of learning algorithm derived from the Bayesian theorem.The Bayesian methods provide a helpful perspective for the understanding of many learning algorithms that do not explicitly manipulate probabilities. Bayes rule is used to determine the probability of the features occurring in each class and to return most likely class. Bayes' theorem, often known as Bayes' Rule or Bayes' law, is a mathematical formula for calculating the probability of a hypothesis given previous information. It is conditional probability that determines this. The Bayes theorem's formula is as follows:

$$P(A/B) = \frac{P(B/A) * P(A)}{P(B)}$$

Where,

P(A/B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B/A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

1.2.2 K NEAREST NEIGHBOURS

The K-Nearest Neighbour method is based on the Supervised Learning technique and is one of the most basic Machine Learning algorithms.The K-NN method assumes that the new case/data and existing cases are comparable and places the new case in the category that is most similar to the existing categories.This method saves all available data and classifies a new data point based on its similarity to the existing data. This implies that fresh data may be quickly sorted into a well-defined category using the K-NN method. This algorithm may be used for both regression and classification, however it is more commonly utilised for classification tasks. This is a non-parametric algorithm, which means it makes no assumptions about the underlying data.The K-Nearest Neighbour method is based on the Supervised Learning technique and is one of the most basic Machine Learning algorithms.It assumes that the new case/data and existing cases are comparable and places the new case in the category that is most similar to the existing categories.The K-NN method saves all available data and classifies it. It is also known as a lazy learner algorithm because it does not learn from the training set right away;instead,it stores the dataset and performs an action on it when it comes time to classify it.During the training phase, the KNN algorithm simply stores the dataset, and when it receives new data, it classifies it into a category that is quite similar to the new data.

1.2.3 SUPPORT VECTOR MACHINE

A support vector machine (SVM) is a technique for locating a hyperplane(linear) in an N- dimensional space that clearly identifies data points with the greatest margin of error.Maximizing the distances between nearby data points of both classes improves classification robustness, however having a small margin increases the risk of missclassification. In most classification jobs, however, more complicated structures are required to achieve optimal separations. In most circumstances, a nonlinear separation is required to better identify the groups than a single linear line. SVM solves this problem by reorganising the data set using kernel functions,which are mathematical functions. To execute the linear separation, the technique involves mapping the data into a new space and transforming it into a higher dimensional feature space.

1.2.4 DECISION TREES

Decision Tree is a supervised learning technique that can be applied to classification and regression problems, however it is most commonly employed to solve classification problems. Internal nodes represent dataset attributes, branches represent decision rules, and each leaf node represents the outcome in a tree-structured classifier.The Decision Node and the Leaf Node are the two nodes of a Decision Tree. Decision nodes are used to make any decision and have several branches, whereas Leaf nodes are the results of such decisions and have no additional branches.The decisions or tests are based on the characteristics of the given dataset.It is a graphical depiction for obtaining all feasible solutions to a problem/decision based on certain criteria.It is termed a decision tree because, like a tree, it starts with the root node and grows into a tree-like structure.The CART algorithm (Classification and Regression Tree algorithm) is used to construct a tree.A decision tree simply asks a question and then divides the tree into subtrees based on the answer (Yes/No).Decision Trees are designed to mirror human thinking abilities when making decisions, making them simple to comprehend.Because the decision tree has a tree-like form, the rationale behind it is simple to comprehend.

1.2.5 LOGISTIC REGRESSION

Under the Supervised Learning approach, logistic regression is one of the most used Machine Learning algorithms. It's used to predict a categorical dependent variable from a group of independent variables.A categorical dependent variable's output is predicted using logistic regression. As a result, the result must be a discrete or categorical value. It can be Yes or No, 0 or 1, true or false, and so on, but instead of giving exact values like 0 and 1, it delivers probabilistic values that are somewhere between 0 and 1.The use of Logistic Regression is similar to that of Linear Regression. For regression problems, Linear Regression is employed, and for classification difficulties, Logistic Regression is used. Rather than fitting a regression line, we fit a ”S” shaped logistic function in logistic regression, which predicts two maximum values (0 or 1). The logistic function’s curve reflects the probability of something, such as whether the cells are cancerous or not, whether a mouse is obese or not depending on its weight, and so on. Because it can generate probabilities and classify new data using continuous and discrete datasets, logistic regression is a key machine learning algorithm.

1.3 PROPOSED SYSTEM

In our proposed system, we train the system with training datasets.We implement this using machine learning algorithms such as Naive Bayes classifier,Support Vector Machine, K-Nearest Neighbour classifier,Decision tree and Logistic Regression.We train the dataset using these algorithms and we discovered that Logistic regression provides maximum accuracy. Therefore, logistic regression is saved as the best model.For live data scrapping, the user sends a Get request using a valid Amazon URL of the product.Our system scraps the product reviews from the comment section using a python package called

Beautiful Soup. After fetching comments using Beautiful Soup, the comments are preprocessed and stop words are removed.The processed dataset is then set as the test dataset to perform the analysis.The program then classifies the reviews into positive ,negative and neutral based on the reviews by the customers.Here the training dataset used is Amazon Musical instruments dataset and for testing scrapped live dataset is used. After classification a report of the analysis is send to the users email id.

1.3.1 DATAFLOW DIAGRAM

1. Check whether the model is trained or not.
2. If the model is not trained, get the Amazon dataset and the dataset is encoded using GloVe model.
3. Train the dataset using the Logistic Regression Algorithm.
4. After training the model , testing is done with another dataset.
5. If the model is trained, get user reviews from Amazon URL by using Beautiful Soup.
6. Each review is evaluated using the trained model.
7. After evaluating the model it shows the statistical parameters of the product as Positive, Negative or Neutral.

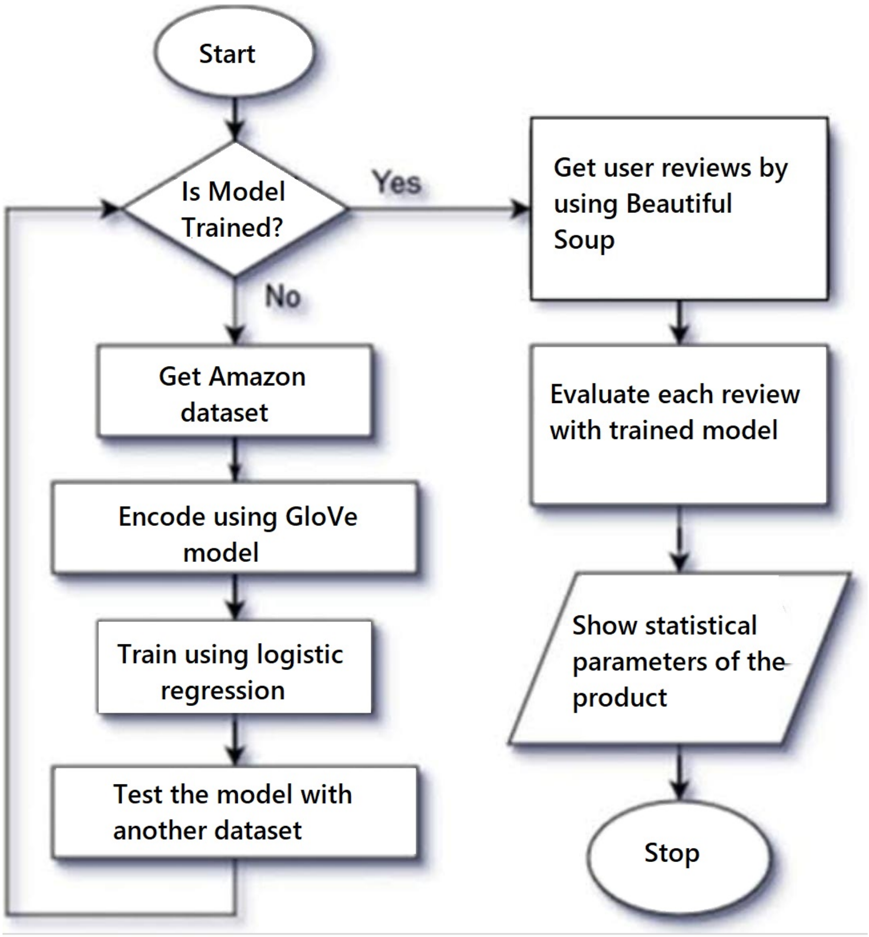

Start
Is Model Trained?
Yes
No
Get user reviews by using Beautiful Soup
Get Amazon dataset
Evaluate each review with trained model
Encode using GloVe model
Train using logistic regression
Show statistical parameters of the product
Test the model with another dataset
Stop

Figure 1.2: Data flow diagram of sentimental analysis

1.4 PROBLEM STATEMENT

- Customers can now share their experiences by rating items and services from specific shops.
- Customer reviews indicate what customers think about prices, value, quality, customer service, ease of shopping, and other aspects of online buying.
- The customer reviews are unstructured, and sentiment analysis will aid in the efficient and cost-effective extraction of meaning from these unstructured texts.

- Using sentiment analysis, businesses will obtain a better knowledge of top-rated products and services, what customers appreciate, and what they dislike.

- In today's competitive and information-driven business climate, it is critical for a company to understand how its consumers feel about the products and services it offers.

- Customers' favourable ratings must be maintained, while neutral and bad customer reviews must be improved.

1.5 AIMS AND OBJECTIVES

The aim is to categorize customer feedback as POSITIVE, NEGATIVE or NEUTRAL.

The objectives are as follows;

1. To determine the intensity of the feelings elicited by customer reviews.

2. To analyze the association between customer reviews concerning different amazon products.

3. To scrap live datasets and classify it as Positive,Negative and Neutral.

Chapter 2

LITERATURE SURVEY

2.1 SENTIMENT ANALYSIS TECHNIQUES INVOLVING SOCIAL MEDIA AND ONLINE PLATFORMS

Opinion analysis also known as sentiment analysis, is critical in achieving the best possible approximation. This is a critical consideration since, well planned and conducted sentiment analysis can produce better and more accurate projections in both politics and business. At its most basic level, sentiment analysis is based on the shared or experienced thoughts of individuals and users. Individuals post and exchange millions of bytes of data on social networking platforms,blogs,product review sites and other sites in the internet space, which pervades practically on every known sphere and sector of human activity on our globe. The ability to collect and analyse such data can provide crucial insights into how products, services, political figures, organisations, governments, and other entities are seen and evaluated. Sentiment Analysis can be used to address a variety of concerns, including accuracy issues, binary classification problems, data sparsity problems, and polarity shift. While different methods for sentiment analysis have been proposed and developed, there is still a need for an effective way for extracting and generating reliable sentiment analysis on a consistent basis. Although machine learning algorithms have come a long way, with Nave Bayes, Support Vector Machine, and Maximum Entropy being the most prominent to appear in research and mainstream use, sentiment classification by category, involving positive and negative sentiments, is a topic of research interest in and of itself. This study includes a survey of popular Sentiment Analysis approaches and methodologies, with the goal of submitting a clear evaluation report that can serve as a foundation for future research. [3] [2]

2.2 SENTIMENT ANALYSIS IN SOCIAL MEDIA

In today's world, sentiment analysis is the most popular research area in the Natural Language Processing field (NLP). The primary goal of this research topic is to use language to identify the emotions and views of customers or users. Despite the fact that several research studies have been conducted in this subject using various models, sentiment analysis remains a difficult problem to address due to the numerous conflicts that must be resolved. Slang terms, new accents, grammar and spelling errors, and other issues are among the current concerns. This research intends to conduct a literature review utilising a variety of machine learning techniques and data.The current literature study entails a scan of roughly 20 papers that span various sorts of sentimental analysis applications. The examination begins by presenting each work's contributions and observing the types of machine learning methods used. Furthermore, the investigation focuses on determining the type of data that was used. In addition, the used environment and performance measures covered in each work are reviewed, and correct research gaps and challenges are concluded, which aids in identifying the non-saturated application for which sentimental analysis is most needed in future study. [8]

2.3 SENTIMENT ANALYSIS TECHNIQUES FOR MICROBLOGGING SITE

Sentiment is any opinion or review provided by an individual via which feelings, text messages, attitudes, and thoughts can be expressed. Sentiment analysis, often known as opinion mining, is the type of data analysis obtained from news reports, user reviews, social media updates, or microblogging sites. It is a method for analysing the sentiment of input data. Sentiment analysis can reveal what people think about certain events, brands, products, or companies. Researchers collect and improvise replies from the general public in order to conduct evaluations. The popularity of sentiment analysis is expanding today, as the number of people sharing their opinions on microblogging sites grows.Positive, negative, and neutral sentiments are the three categories that can be used to categorise all of the emotions. The most popular microblogging service, Twitter, is employed to collect the data for research. The source data was extracted from Twitter using Tweepy. This study employs the Python programming language to implement the classification method on the given data. The N-gram modelling technique is used to extract the features. A supervised machine learning method known as K-Nearest Neighbor is used to identify the attitudes as positive, negative, or neutral. [6]

2.4 SENTIMENT ANALYSIS USING MACHINE LEARNING TECHNIQUES

Many businesses use social media networks to provide various services, communicate with clients, and gather information about people's thoughts and opinions. Sentiment analysis is a machine learning technique that detects polarity like positive or negative ideas in text, whole texts, paragraphs, lines, or subsections. Machine Learning (ML) is a multidisciplinary discipline that combines statistics and computer science methods to develop predictive and classification algorithms. The common strategies for assessing sentiment from a machine learning perspective are presented in this work. In light of this, a systematic examination and assessment of business and community white papers, scientific research articles, journals, and reports was conducted to examine and analyse the concept of sentiment analysis.The goal and key objectives of this article are to categorise and examine the most common research approaches and Machine Learning implementations in Sentiment Analysis on various applications. The drawback of this approach is that it focuses solely on the application side, ignoring the hardware and theoretical exposure relevant to the problem. The study's shortcoming is that it focuses mostly on the application side of things, ignoring the hardware and theoretical parts of the issue. Finally, this study provides a research proposal for sentiment analysis in an e-commerce setting using machine learning techniques. [5]

Chapter 3

SYSTEM OVERVIEW

This section gives the overview of the approach.The sentimental analysis for Amazon reviews is done using the Machine Learning Approach of the sentimental analysis.Figure 3.1 shows the general framework involved in the sentimental analysis of text based sentimental analysis.The primary steps includes Data collection,Feature selection,Feature extraction,Sentiment detection,Sentiment classification and Sentiment scoring,visualisation, output and scoring.

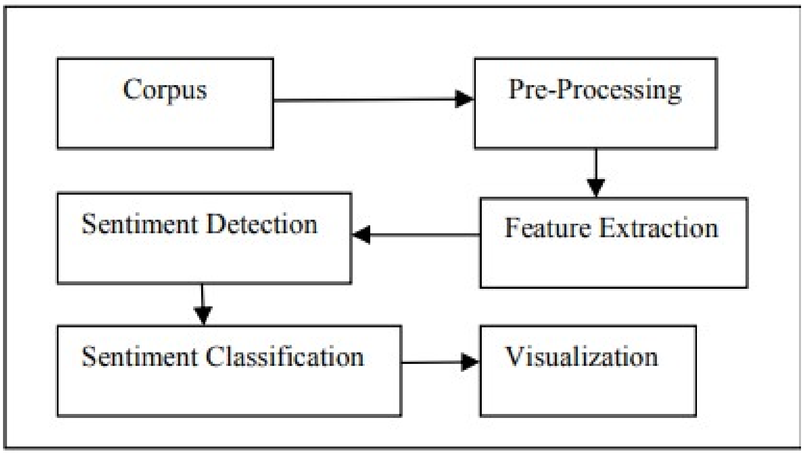


Figure 3.1: Schematic representation of work flow

3.1 DATA COLLECTION

Data collection is the process of collecting of data needed for the process from the internet via web scraping, Weblog, social media,forums or from websites such as www.amazon.com, Kaggle, twitter API etc.In our system, we are taking into consideration the Amazon reviews data set for the sentimental analysis purpose.in our system,we are considering live datasets to implement the sentiment analysis.The live dataset is scrapped from the Amazon URL of the product provided by the user using a python package called Beautiful Soup.The Beautiful Soup is used for parsing HTML and XML files.It extracts the reviews from the URL provided by the user.The scrapped data is assigned as the test for conducting the analysis of reviews.

3.2 FEATURE SELECTION

The identification of the features of the model is an important step in the classification of a model.The data set needs to be classified based on the requirement of the model.Data set is decoded into words during model training and it is added to the feature vector.. For single word is considered, the technique is called a “Uni-gram”; when two words are considered, the technique is called a “Bi-gram”; and if three words are considered, the technique is referred to as a “Tri-gram.” combination of unigram and bigram helpful for analysis; the context feature which helpful for getting results most accurate.Pragmatic features are those that emphasize the application of words rather than a methodological foundation. Pragmatics is the study of how context relates to perception in linguistics and related sciences. Pragmatics is the study of phenomena such as implicature, speech acts, relevance, and conversations.

3.3 FEATURE EXTRACTION

If using the machine learning approaches, feature extraction is must for the sentiment analysis. Feature extracting is also one of the most significant stages building effective classifiers . The achievement or disappointment of the sentiment classification model is strongly reliant on the quality of the features.Feature extraction is a key task in sentiment classification as it involves the extraction of valuable information from the text data, and it will directly impact the performance of the model. The approach tries to extract valuable information that encapsulates the text’s most essential features. In most cases, punctuation’s are removed from the text after lowering it in the pre-processing stage, but they used them to extract features and hashtags and emoticons commonly used techniques for feature extractions are listed below.

Terms frequency It is one of the simplest ways to express features that are more frequently used in various NLP applications, including Sentiment Analysis, for information retrieval. It considers a single word, i.e., uni-gram or group of two-three words, which can be in bi-gram and tri-gram, with their terms count representing features.Term's presence gives the word a value of either 0 or 1. Term frequency is the integer value, which is its count in the given document. TF-IDF can be used as a weighted scheme for better results that will measure the importance of any token in the given document.

Negations These are the words that can change or reverse the polarity of the opinion and shift the meaning of a sentence. Commonly used negation words include not, cannot, neither, never, nowhere, none, etc. Every word appearing in the sentence will not reverse the polarity; therefore, removing all negation words from stop-words may increase the computational cost and decrease the model's accuracy. Negation words must be handled with at most care. Negation words such as not, neither, nor, and so on are critical for sentiment analysis since they can revert the polarity of a given phrase. For instance, the line "This movie is good." is a positive sentence, but "The movie is not good." is a negative sentence. Regrettably, some systems eliminate negation words because they are included in stop word lists or are implicitly omitted since they have a neutral sentiment value in a lexicon and do not affect the absolute polarity. However, reversing the polarity is not straight forward because negation words might occur in a sentence without affecting the text's emotion. [4]

*Bag of Words (BoW)*BOW is one of the simplest approach for extracting text features. BoW will describe the occurrence of words in a document. Bag represents the vocabulary of words using which a vector is formed for each sentence. The main problem with this model is that it does not consider the syntactic meaning of the text. For instance, consider two sentences s1= "the food was good", s2= "the service was bad". The vocabulary is created for two sentences where v= 'the', 'food', 'was', 'service', 'bad', 'good' and the length of the vector is 6 and is represented as v1= [1 1 1 0 0 1] and v2= [1 0 1 1 1 0]. BoW approach performance evaluated using (TF-IDF) which performs better in most cases.

3.3.1 WORD EMBEDDING

Word embeddings represent words in a vector space by clustering words with similar meanings together. Each word is assigned to a vector, which is then learned in a manner similar to neural networks. It learns and chooses a vector from a predetermined vocabulary. The dimension of the words may be chosen by passing it as a hyper parameter. SG model and the continuous CBOW model are two of the most well-known algorithms for word embeddings. Both of these are shallow window approaches methods in which a short window of some size, such as four or six, is specified, and the current word is anticipated using context words in CBOW, while context words are forecasted using the current word in the SG model. Word embeddings are concerned with learning about words in the context of their local usage, which is specified by a window of nearby terms.Global Vectors (GloVe) Global Vectors for word representation have developed by an unsupervised learning approach to generate word embeddings from a corpus word-to-word co-occurrence matrix. GloVe is a popularly used method as it is straightforward and quick to train GloVe model because of its parallel implementation capacity.

3.4 SENTIMENT DETECTION

In sentiment analysis during the sentiment detection stage, each sentence removed from the opinion and review is examined for subjectivity. The statements of sentences that consists of subjective terminologies are kept and the objective sentences are disallowed. Sentiment analysis as well as opinion mining is completed at dissimilar levels of language for instance at the morphological, lexical, pragmatic levels and semantic discourse Sentences consists of subjective expressions (opinions, views and beliefs) are retained and sentences that consists of objective communication (factual information, facts) are rejected.

3.5 SENTIMENT CLASSIFICATION

Sentiment categorization is a well-known researched task in sentiment analysis.The classification of sentiments is executed following procedures that ensure the raw data is cleaned, converted into a corpus (a collection of words) and a term document matrix (a matrix showing the frequency of words in the corpus). These objects make it easy for the classification models to learn the relationship between the words and the sentiments and the term "Opinion analysis" is frequently used while referring to Sentiment Analysis. It is a little duty aimed on determining the sentiment of each piece of text.The sentiment is either positive negative or neutral.With a trained classifier, the cross-domain analysis predicts the sentiment of a target domain. Extracting the domain invariant features and where they are distributed is a commonly used approach.

3.6 SENTIMENT SCORING AND VISUALISATION

The performance of the classification methods can be found by using Accuracy, F-Score, Cross entropy, Recall, and Precision. These parameters are helpful to evaluate the performance of supervised machine learning algorithms, based on the element from a matrix known as the confusion matrix or contingency table.These annotations from the earlier analysis are the output of the system and the visualization tools will present the results to the user. The basic assumption of sentiment analysis is to alter unstructured data into important or data that is meaningful. At the end after the conclusion of analysis, the acquired results are presented on graph such as bar chart and line graphs.

3.6.1 CONFUSION MATRIX

A confusion matrix is typically used for allowing visualization of the performance of an algorithm. From the classification viewpoint,terms such as'True Positive (TP)', 'False Positive (FP)','True Negative (TN)', 'False Negative (FP)' are used to compare labels of classes in this matrix. True Positive represents positive reviews that were classified as positive by the classifier, whereas False Positive is predicted as negative but is actually classified as

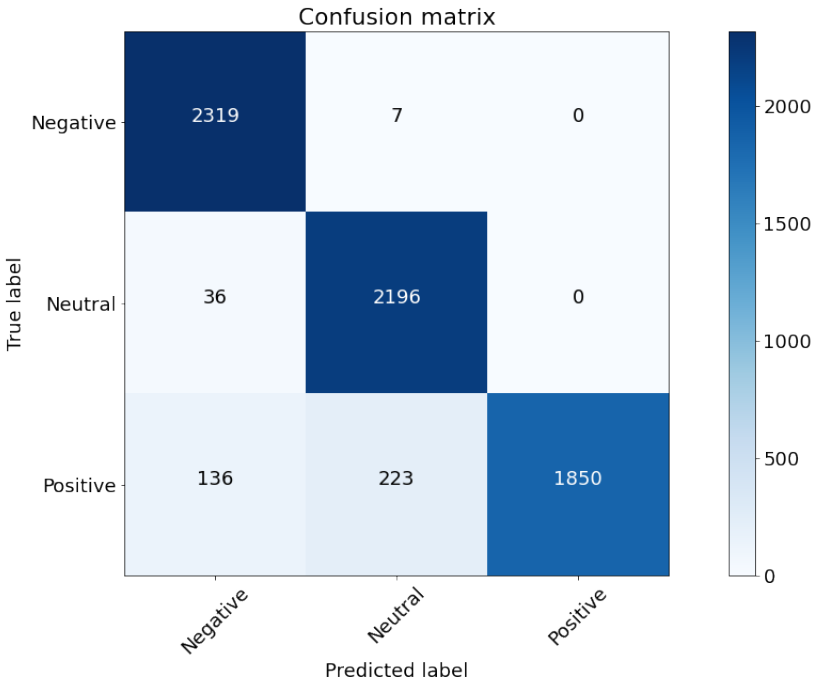


Figure 3.2: Confusion matrix of sentimental analysis based on positive, negative and neutral reviews

negative. Conversely, True Negative represents negative reviews that were classified as negative by the classifier, whereas False Negative is predicted as positive actually classified as negative. According to the data of the confusion matrix, precision,recall, f-measure, and accuracy are used for evaluating the performance of classifiers.

Chapter 4

SYSTEM DESIGN

4.1 INPUTS

- A set of training data and testing data consisting of unique ASIN ID.
- A sentence that is reasonable for sentimental analysis.
- Set of words called STOPPING WORDS which is used to remove unnecessary words during sentimental analysis.

4.2 OUTPUTS

Produces the following output:

- Classify the sentence into positive,negative and neutral.
- Predict the overall sentiment of a product along dif-

ferent stages.

- Complete the missing values with predicted one.

4.3 SYSTEM ARCHITECTURE

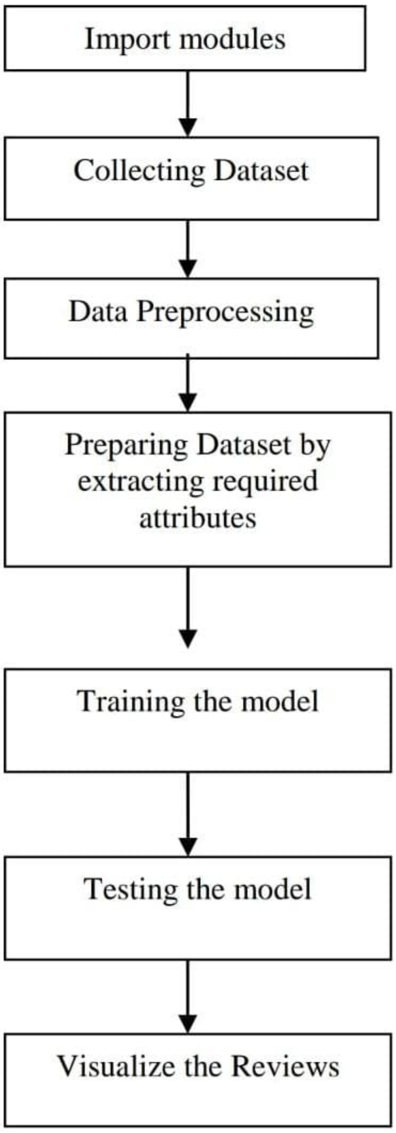


Figure 4.1: Processes involved in sentiment analysis

The huge dataset of reviews obtained from amazon.com comes in a .json file format. A small python code has been implemented in order to read the dataset from those files and dump them in to a pickle file for easier and fast access and object serialization.. Each review includes the following information: 1) reviewer ID; 2) product ID; 3) rating; 4) time of the review; 5) helpfulness; 6) review text. Every rating is based on a 5-star scale, resulting all the ratings to be ranged from 1-star to 5-star with no existence of a half-star or a quarter-star.Hence initial fetching of data is done in this section using Python File Handlers.

The pickle file is hence loaded in this step and the data besides the one used for sentiment analysis is removed.The review dataset contains numerous columns out of which the columns only rating and text review is what we require. So, the column, "reviewSummary" is dropped from the data file.) After that, the review ratings which are 3 out of 5 are removed as they signify neutral review, and all we are concerned of is positive and negative reviews.The entire task of preprocessing the review data is handled by this utility class- "NltkPreprocessor".

The data preprocessing is the vital step in the sentiment analysis process.Here, Words present in the file are accessed both as a solo word and also as pair of words. Because, for example the word "bad" means negative but when someone writes "not bad" it refers to as positive. In such cases considering single word for training data will work otherwise. So words in pairs are checked to find the occurrence to modifiers before any adjective which if present which might provide a different meaning to the outlook.Tokenization of reviews after removal of STOP words which mean nothing related to sentiment is the basic requirement for POS tagging. After proper removal of STOP words like “am, is, are, the, but” and so on the remaining sentences are converted in tokens. These tokens take part in POS tagging.In natural language processing, part-of-speech (POS) taggers have been developed to classify words based on their parts of speech. For sentiment analysis, a POS tagger is very useful because of the following two reasons: 1) Words like nouns and pronouns usually do not contain any sentiment. It is able to filter out such words with the help of a POS tagger; 2) A POS tagger can also be used to distinguish words that can be used in different parts of speech.

The training of dataset model is the process of training the dataset using various algorithms to perform and classify the dataset items into negative,neutral and positive based on numerous parameters.Machine learning algorithms like Support Vector Machine(SVM),Logistic Regression algorithm,Naive Bayes algorithm,K-Nearest Neighbour(KNN) model,Decision tree algorithm are used to train the dataset.The Accuracy, Precision, Recall, and Evaluation time is calculated and displayed.The model that exhibit the highest accuracy is considered the better model and is applied to perform sentimental analysis for the testing datasets.Prediction of test data is done and Confusion Matrix of prediction is displayed.Total positive, neutral and negative reviews are counted.A review like sentence is taken as input on the console and if positive the console gives 1 as output and 0 for negative input.

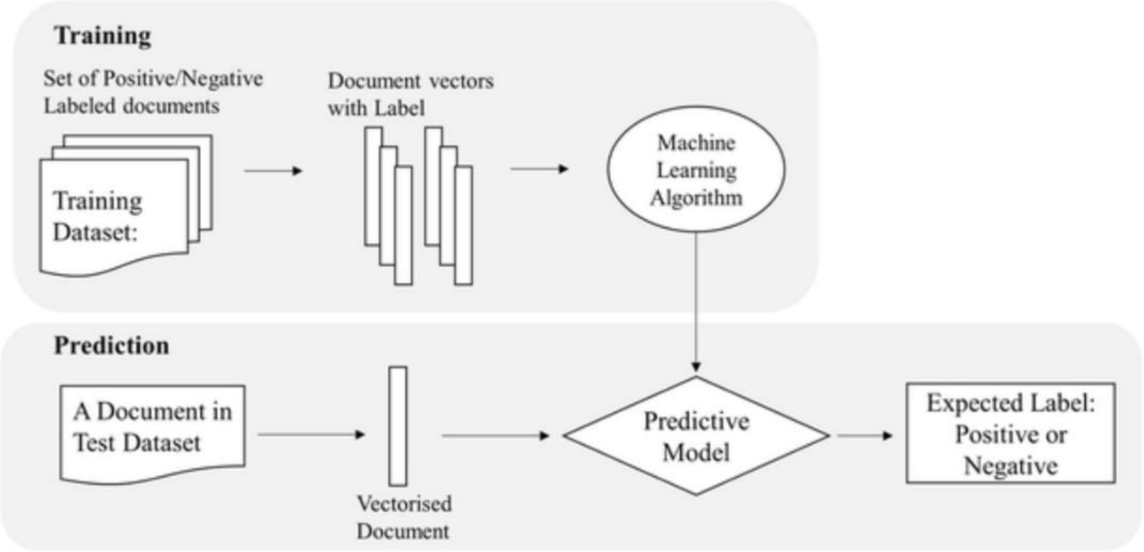

Training
Set of Positive/Negative Labeled documents
Training Dataset:
Document vectors with Label
Machine Learning Algorithm
Prediction
A Document in Test Dataset
Vectorised Document
Predictive Model
Expected Label: Positive or Negative

Figure 4.2: Training process involved in sentiment analysis

The testing model is the conducting of the sentimental analysis using the trained model.The test data is scrapped from the Amazon URL provided by the user.The predictive model uses the machine learning model that induced higher accuracy i.e, the logistic regression model is used for conducting sentimental analysis on the test dataset.The preprocessing of the test dataset is performed similar to that of the training dataset.The predictive model obtained from training of dataset is used to perform sentimental analysis on test dataset.The process is illustrated in the figure 4.3 given below. We can see the training dataset undergoes the sentimental analysis processes to develop a predictive model which is then used to conduct the analysis on test data after data preprocessing and feature extraction processes.The output obtained is the Positive, Neutral and Negative classification of the input reviews from the dataset.

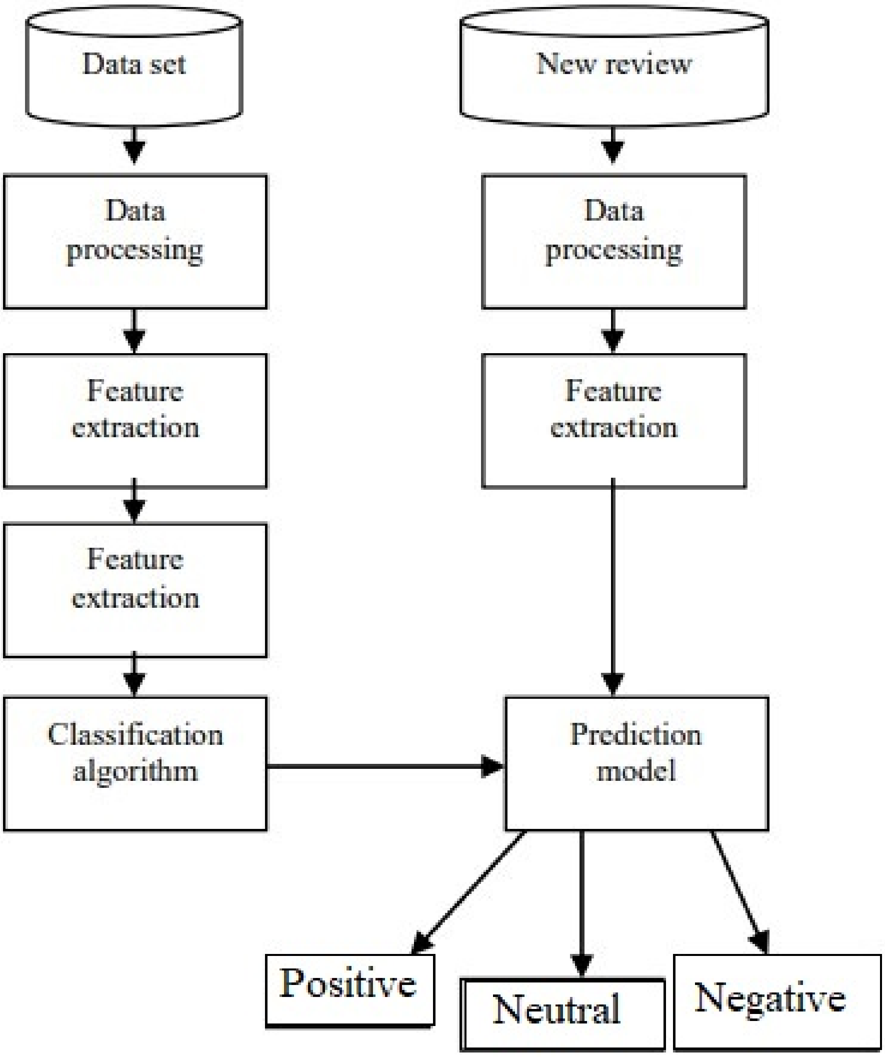


Figure 4.3: The testing process of a dataset using trained dataset model

The visualisation of sentimental analysis output include graphical representations,histogram representations,ROC curve,confusion matrix.These are portrayed based on the classification of reviews using parameters positive,neutral and negative,The confusion matrix is obtained from the count of the positive ,neutral and negative reviews.The ROC curve is constructed indicating the Receiver operator characteristics of the multi-class.The bigram and trigram plotting of review words can be done based on the positive ,neutral and negative reviews.This is helpful in understanding the words that appear together often in the reviews and their word count can evaluated using the representation with the help of graphs.The frequency of words in positive,neutral and negative reviews,i.e, word count plots is constructed to evaluate the word count in review classification.

Chapter 5

IMPLEMENTATION

The goal of this project is to make an overall sentiment analysis of a product by counting the value of each review. The model includes methods to count positive negative and neutral values by analyzing each key words.The program must be start by running anaconda navigator, from that select the Jupyter notebook for perform the project.

Notes on running the code:

- The anaconda navigator must be updated recently for better GUI experience

- It should be available with latest python module (python version 3.9) is preferable

- Navigator should be satisfied with pip or pip3 carries latest version

- Note that the necessary libraries should be in recent mode so that bug or implementation error can be resolved.

Steps in execution:

- Run the code after importing necessary packages such as ipynb,pandas,numpy,sklearn,nltk.

- Ensure the training dataset must be available in the code directory.

- Import the dataset using pandas and open it in read mode

- Execute the code and ensure the dataset is loaded by printing either head() or tail()

- Avoid unnecessary coloumns from the dataset and add or skip missing values in it

- Train the model and select the best model by checking its performance analysis.

- After that apply normalization to obtain the best model.

- Save the model using pickle and perform the live data scrapping.

- The scrapped data is directed to the predefined model so that testing is occur and result can be evaluated

Chapter 6

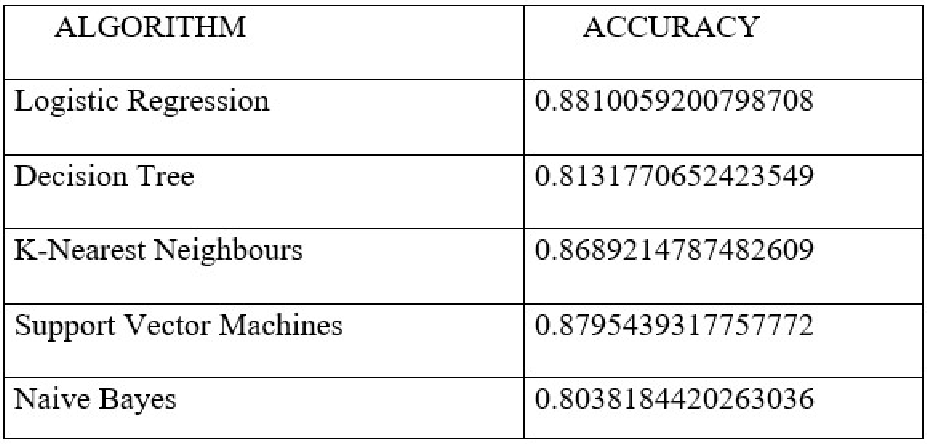
RESULTS

- The proposed system successfully scraps the live dataset from Amazon product URL.

- The system successfully classifies the dataset into positive, negative and neutral.

- The model predicts the sentiment of reviews with an accuracy of 88%.Higher accuracy was demonstrated by the training dataset when employing logistic regression.Therefore, logistic regression is considered for testing dataset in the sentimental analysis process.

- The total number of input data and number of data after classification will be same. For example, if our number of inputs is 50, then the output number will also be a total of 50 classified into positive, negative and neutral.



ALGORITHM	ACCURACY
Logistic Regression	0.8810059200798708
Decision Tree	0.8131770652423549
K-Nearest Neighbours	0.8689214787482609
Support Vector Machines	0.8795439317757772
Naive Bayes	0.8038184420263036

Table 6.1: Accuracy of machine learning models

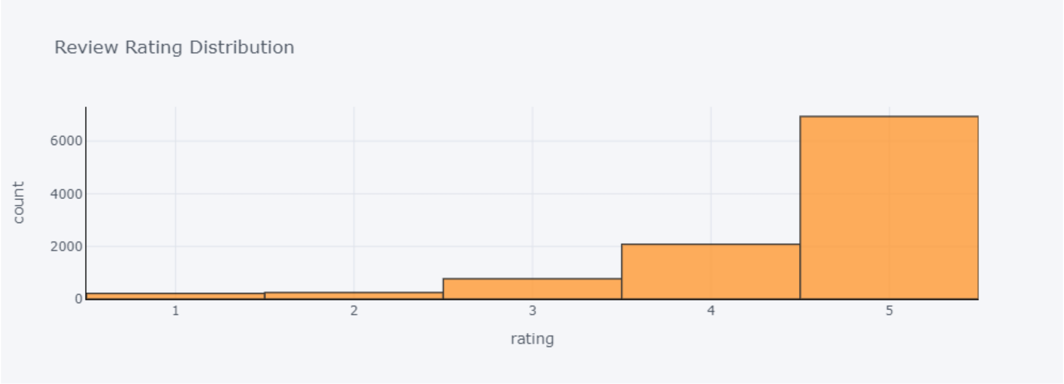


Figure 6.1: Review Rating Distribution

- Overall ratings are distributed and have a large number of 5 ratings followed by 4,3,2,1. It's linear in nature.

- Considering the ROC curve for classes, class 2 and 0 have been classified pretty well a their area under the curve is high. We can chose any threshold between 0.6-0.8 to get the optimal number of TPR and FPR.

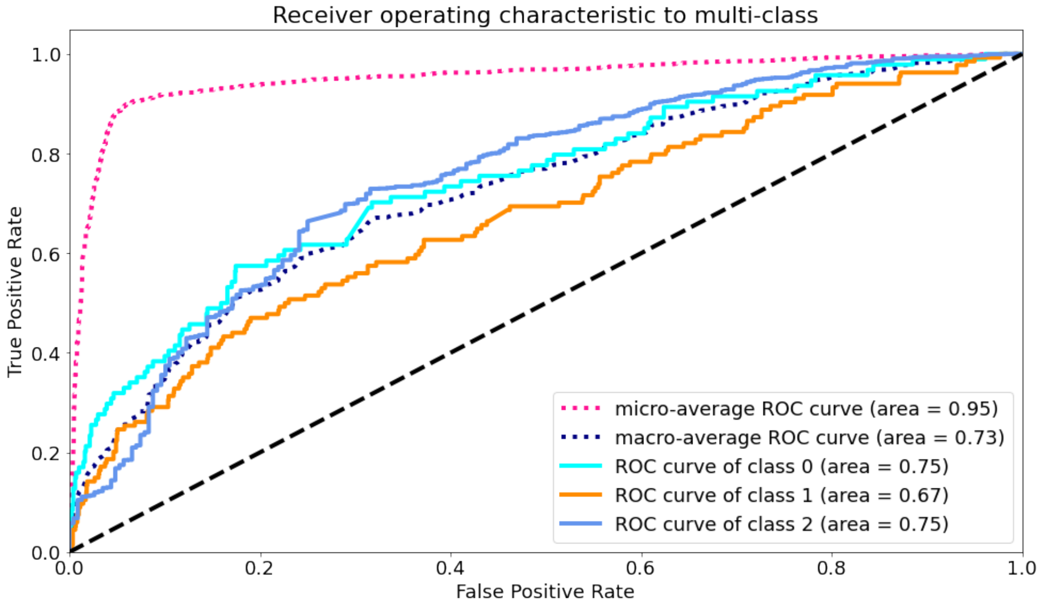


Figure 6.2: ROC Curve indicating the Receiver operating characteristics to multi-class

- Coming to micro and macro average, micro average preforms really well and macro average shows a not very good score.

- A macro-average will compute the metric independently for each class and then take the average (hence treating all classes equally), whereas a micro-average will aggregate the contributions of all classes to compute the average metric. In a multi-class classification setup, micro-average is preferable if you suspect there might be class imbalance.

- The classification report contain the positive,neutral and negative classification of reviews.It also includes precision,recall,f1-score,support of reviews.The total accuracy,macro average and weighted average is also computed for the total review analysis.

Chapter 7

CONCLUSION

Sentiment analysis is a necessary and widely used method for collecting information from text data on eCommerce websites. In the form of ideas, feedback, tweets, and comments, e-commerce portals generate a vast volume of text data every day. Aside from that, the public's opinion is important.Reviews, ratings, and emoticons all imply humans. Taking information about a product from a website.The review will assist a customer in learning more about the product and making a decision.This study has applied different machine learning algorithms like Naive Bayes, Support Vector Machine, K Nearest Neighbours, Decision Tree and Logistic Regression on Amazon product reviews. The result from the study showed that in terms of accuracy the Logistic Regression approach achieves better results than all other algorithms when the whole data set was used as training and testing data set. Here the training dataset used is Amazon Musical instruments dataset and for testing scrapped live dataset is used. The dataset is tested using the trained data model to perform the Sentimental analysis.The model classifies the total reviews into positive,neutral and negative and provide the count of each classification.After classification a report of the conducted analysis is send to the users email id.

7.1 FUTURE WORK

For future work we are planning to detect whether the user is fake or not using ASIN ID.Amazon Standard Identification number(ASIN ID), is a 10 character alphanumeric unique identifier number assigned by Amazon.com for product identification within the Amazon organization.By using this Asin id we can find out the users activity on Amazon product reviews and detect whether he/she is a paid user or not.Future result should continue to explore another efficient sentiment classifiers like Decision tree, Support Vector Machines etc.

Bibliography

[1] MJ Budhwar and Sukhdip Singh. Sentiment analysis based method for amazon product reviews. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICACT*, 9(08), 2021.

[2] Raktim Kumar Dey, Debabrata Sarddar, Indranil Sarkar, Rajesh Bose, and Sandip Roy. A literature survey on sentiment analysis techniques involving social media and online platforms. *International Journal Of Scientific & Technology Research*, 1(1), 2020.

[3] Sureshkumar Govindaraj and Kumaravelan Gopalakrishnan. Intensified sentiment analysis of customer product reviews using acoustic and textual features. *ETRI Journal*, 38(3):494–501, 2016.

[4] Sepideh Paknejad. Sentiment classification on amazon reviews using machine learning approaches, 2018.

[5] Megha Rathi, Aditya Malik, Daksh Varshney, Rachita Sharma, and Sarthak Mendiratta. Sentiment analysis using machine learning approach. In *2018 Eleventh international conference on contemporary computing (IC3)*, pages 1–3. IEEE, 2018.

[6] Priyanka Tyagi, Sudeshna Chakraborty, RC Tripathi, and Tanupriya Choudhury. Literature review of sentiment analysis techniques for microblogging site. In *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India*, 2019.

[7] Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, pages 1–50, 2022.

[8] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2):617–663, 2019.