MINI PROJECT REPORT [EC 65]

Sentiment Analysis on Product Reviews Using Machine Learning

BACHELOR OF ENGINEERING

in

Electronics and Communication

Names of the Student	USN
Narva Rakesh Reddy	1MS20EC054
Ravi N	1MS20EC079
S Taranya	1MS20EC085
Sagar Maruthi	1MS20EC086

Under Guidance of

H. Mallika

Assistant Professor, Department of E & C

Department of Electronics and Communication

RAMAIAH INSTITUTE OF TECHNOLOGY

(Autonomous Institute, Affiliated to VTU)

Accredited by National Board of Accreditation & NAAC with 'A+' Grade

MSR Nagar, MSRIT Post, Bangalore-560054

www.msrit.edu

2023

CERTIFICATE

This is to certify that the Mini Project work entitled **"**Sentiment Analysis on Product Reviews Using Machine Learning **"** is carried out by Narva Rakesh Reddy (1MS20EC054), Ravi N (1MS20EC079), S Taranya (1MS20EC085), Sagar Maruthi (1MS20EC086), bonafide students of Ramaiah Institute of Technology, Bangalore, in **Electronics and Communication** of the Visvesvaraya Technological University, Belgaum, during the year 2022 -2023. It is certified that all corrections / suggestions indicated for Internal Assessment have been incorporated in the report.

Guide	**HOD**
H. Mallika	**Dr. Maya V Karki**
Assistant Professor	Professor and HOD
Department of E & C	Department of E & C
RIT, Bangalore	RIT, Bangalore

Name & Signature of Examiners with Date:

1)

2)

DECLARATION

We hereby declare that the Mini Project entitled “Sentiment Analysis on Product Reviews using Machine Learning” has been carried out independently at Ramaiah Institute of Technology under the guidance of **H. Mallika, Assistant Professor, Department of Electronics and Communication, RIT, Bangalore.**

Signature of Students:

1. Narva Rakesh Reddy **[**USN: 1MS20EC054]
2. Ravi N **[**USN: 1MS20EC079]
3. S Taranya **[**USN: 1MS20EC085]
4. Sagar Maruthi **[**USN: 1MS20EC086]

Place:

Date:

ACKNOWLEDGEMENT

The immense satisfaction that accompanies the successful completion of the project would be incomplete without the mention of the people who made it possible. We consider it our honour to express our deepest gratitude and respect to the following people who always guided and inspired us during the course of the Project.

We are deeply indebted to **Dr. N. V. R. Naidu**, Principal, RIT, Bangalore for providing us with a rejuvenating master course under a very creative learning environment.

We are much obliged to **Dr. Maya V Karki**, Professor & HOD, Department of Electronics and Communication Engineering, RIT, Bangalore for her constant support and motivation.

We sincerely thank our guide **H. Mallika, Assistant Professor,** Department of Electronics and Communication Engineering, RIT, Bangalore and express our humble gratitude for **her** valuable guidance, inspiration, encouragement and immense help which made this work a success.

We sincerely thank the **Chairperson of the group Dr. K. Indira** and **Dr. Rajendra Prasad P** for reviewing our work and providing valuable suggestions. We also thank all the faculty members of Department of E&C, RIT for their kind support to carry out this project successfully.

ABSTRACT

In today's digital age, the abundance of online product reviews has made it increasingly challenging for consumers to make informed purchasing decisions. Sentiment analysis, a branch of natural language processing, offers a promising solution by automating the analysis of textual data to determine the sentiment expressed within product reviews. This abstract presents a comprehensive overview of sentiment analysis on product reviews using machine learning techniques.

The primary objective of this study is to develop an effective sentiment analysis model that can accurately classify product reviews into positive, negative, very positive, very negative or neutral sentiments. To achieve this, a corpus of product reviews from diverse domains will be collected, preprocessed, and annotated for sentiment labelling. Various machine learning algorithms, such as Support Vector Machines (SVM), Naïve Bayes, and Random Forests, will be implemented and evaluated for their performance in sentiment classification.

The preprocessing phase involves techniques such as tokenisation, stop word removal and stemming. Feature extraction methods, such as bag-of-words, n-grams, and word embeddings, will be explored to capture the essential information from the text and transform it into numerical representations suitable for machine learning models.

The sentiment analysis models will be trained on labeled data using a supervised learning approach, where the input features are the transformed textual data, and the target variable is the sentiment label. The performance of the models will be evaluated using standard evaluation metrics such as accuracy, precision, recall, and F1-score.

Furthermore, the study will investigate the impact of different factors on sentiment analysis performance, such as the size of the training data and the effect of using different machine learning algorithms. Additionally, techniques for addressing challenges like imbalanced data, sarcasm, and sentiment intensity will be explored to enhance the overall sentiment analysis accuracy.

The potential applications of sentiment analysis on product reviews are extensive. Companies can utilise this technology to gain insights into customer opinions and preferences, improve product development, and enhance customer satisfaction. Consumers can benefit from sentiment analysis by obtaining a summarised sentiment analysis score for a product, helping them make informed decisions before making a purchase.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Introduction

The rise of e-commerce has revolutionised the way people make purchasing decisions. With the availability of numerous products and brands at the click of a button, customers heavily rely on the opinions and experiences of other customers to make informed decisions. Online customer reviews, therefore, have become a critical source of information for both customers and businesses. Sentiment analysis, a subfield of Natural Language Processing (NLP), aims to automate the process of determining the sentiment expressed in a text document, such as a customer review.

The objective of this project is to perform sentiment analysis on product reviews to understand the overall sentiment of customers towards a product. The insights from this analysis can provide valuable information to companies and sellers in improving their products and customer satisfaction. Various machine learning and deep learning models can be used for the same.

The results of the sentiment analysis are presented in the form of visualisations and statistical analysis, allowing for easy interpretation of the findings. The report provides a comprehensive overview of the sentiment analysis performed on the product reviews data.

In conclusion, sentiment analysis on customer reviews has the potential to provide valuable insights into customer opinions and preferences.

Other examples of application of sentiment analysis :

- Market Research :

Sentiment analysis can be a valuable tool for market research. By analysing customer feedback from various sources (such as product reviews, social media, and customer surveys), businesses can gain insight into the overall sentiment towards their brand or products, as well as identify specific areas where they may need to improve.

Sentiment analysis can also help businesses understand customer preferences and trends, which can inform product development and marketing strategies.

- Reputation Monitoring :

Sentiment analysis can be used to monitor a business's reputation by analysing customer feedback from various online sources. By tracking sentiment towards their brand, businesses can quickly identify any negative feedback or complaints and take steps to address them. Reputation monitoring can also help businesses identify trends and patterns in customer feedback, which can inform product development and marketing strategies.

- Customer Support :

Sentiment analysis can be used in customer support to analyse customer feedback and identify any issues or areas where customers may be struggling. By analysing customer feedback in real-time, businesses can quickly identify any negative sentiment and take steps to address the issue. Sentiment analysis can also be used to identify common customer concerns or complaints, which can inform the development of self-service support resources or training for customer service representatives.

1.2 Problem Definition

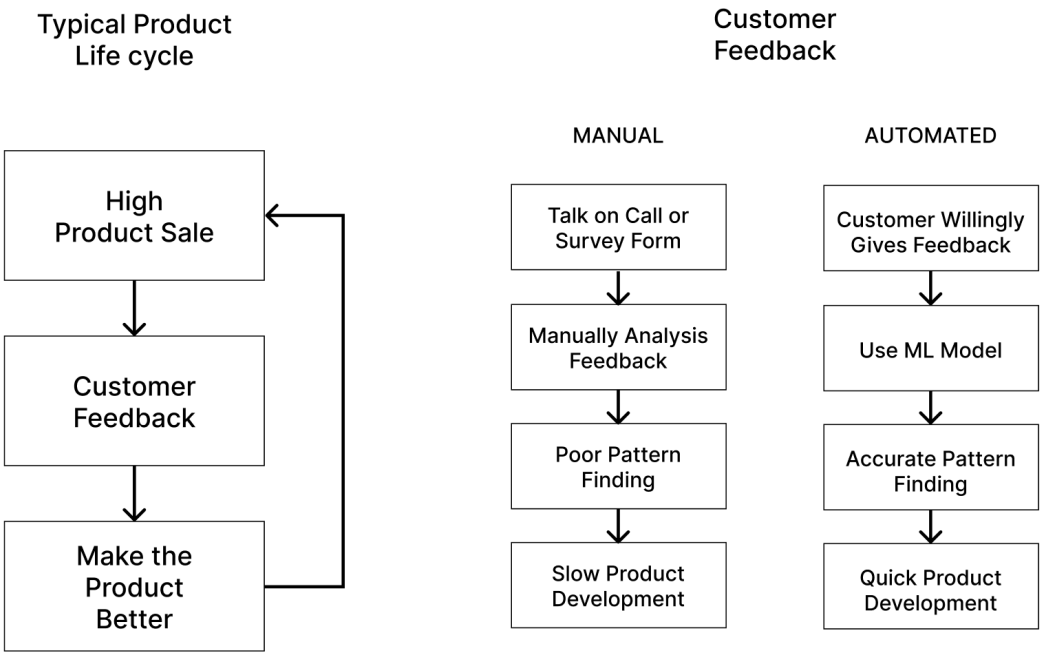
The problem addressed in this project is to perform sentiment analysis on product reviews to understand the overall sentiment of customers towards a product. The objective is to provide valuable insights to companies and sellers in improving their products and customer satisfaction. The manual process of analysing customer reviews can be time-consuming and may not provide an accurate representation of customer sentiment. Sentiment analysis using machine learning algorithms can automate this process and provide a more accurate representation of customer sentiment.

The project aims to address the following :

1. How do models like SVM, Random Forest, Naive Bayes, etc. perform?
2. What is the overall sentiment of customers towards a product?
3. Can sentiment analysis provide valuable insights to companies and sellers in improving their products and customer satisfaction?
4. What are the limitations and challenges of sentiment analysis on customer reviews and how can they be addressed in future studies?

The results of this project will contribute to the growing body of research in sentiment analysis and its applications in the e-commerce industry.

With the rise of Internet, data has become abundant and analysing such huge data cannot be done manually. Considering that companies and brands are looking for every possible way to increase their sales and customer satisfaction, providing them with a powerful tool to understand their customers sentiment is a big win for the companies as they can built out a feedback mechanism that helps them improve the product.



Typical product life cycle:

1. High product sale: The initial stage of the product life cycle is characterised by high sales as the product is introduced to the market. During this phase, customers purchase the product and start using it.

2. Customer feedback: As customers use the product, they provide feedback based on their experiences. This feedback can be crucial for understanding the product's strengths, weaknesses, and areas for improvement. Customer feedback plays a vital role in shaping the future development and success of the product.

3. Make the product better: Upon receiving customer feedback, the product development team analyses and evaluates the feedback to identify areas for improvement. This feedback can involve suggestions, complaints, or requests for additional features. The team then works on implementing these improvements, whether they are related to enhancing existing features, fixing issues, or introducing new functionalities.

4. Linked back to high product sale: By addressing customer feedback and making the necessary improvements, the product becomes better aligned with customer needs and preferences. This, in turn, increases customer satisfaction, loyalty, and positive word-of-mouth. As a result, the product has a higher likelihood of generating increased sales and market demand.

Customer feedback can be of two types: Manual and Automated.

Manual feedback:

- Take calls or use survey forms: Customers may provide feedback through customer service calls or by filling out survey forms. These methods involve direct interaction with customers, allowing them to express their opinions, concerns, or suggestions.

- Manually analyse feedback: The feedback collected through manual means is manually analysed by the product development team or customer support representatives. This process involves reading, categorising, and interpreting the feedback to extract valuable insights.

- Poor pattern finding: Manual analysis of feedback can be time-consuming and prone to human biases or limitations. It may be challenging to identify patterns or trends within a large volume of feedback, leading to inefficiencies in identifying actionable insights.

- Slow product development: Due to the manual nature of analysing feedback and identifying patterns, the process of implementing improvements and updates to the product can be slower. The time taken to address customer concerns and enhance the product may result in delayed development cycles.

Automated feedback:

- Customer willingly gives feedback in the form of product reviews: Automated feedback is generated when customers voluntarily share their experiences, opinions, and ratings through product reviews on various platforms such as e-commerce websites, social media, or dedicated review platforms.

- Use ML Model: Machine learning models can be employed to analyse the vast amount of textual data available in product reviews. These models use natural language processing techniques to extract sentiment, identify key themes, and detect patterns in customer feedback.

- Accurate pattern finding: By leveraging machine learning algorithms, automated analysis of customer feedback can provide more accurate pattern finding capabilities. The models can efficiently process large volumes of reviews, detect sentiment polarity, and identify recurring themes or issues.

- Quick product development: With the assistance of automated analysis, the product development team can swiftly identify actionable insights from customer feedback. This allows for a more rapid and efficient product development cycle, enabling timely improvements and updates based on customer needs and preferences.

In summary, both manual and automated customer feedback play essential roles in the product life cycle. While manual feedback collection and analysis can be time-consuming and prone to limitations, automated feedback through product reviews, coupled with machine learning analysis, offers the advantage of faster and more accurate pattern finding. By leveraging customer feedback, whether collected manually or through automation, companies can continuously improve their products, enhance customer satisfaction, and ultimately drive higher product sales.

1.3 Motivation of the Work

In today's digital era, the proliferation of online shopping platforms and the ease of accessing product information have revolutionised the way consumers make purchasing decisions. However, the abundance of product options and the sheer volume of available information pose significant challenges for customers. As a result, consumers often turn to product reviews as a valuable source of insights and guidance to inform their buying choices.

Product reviews provide a platform for customers to share their experiences, opinions, and sentiments regarding a particular product. However, manually scouring through countless reviews to make sense of the sentiments expressed by customers is a time-consuming and arduous task. This is where sentiment analysis, coupled with machine learning techniques, can offer a solution.

The primary motivation of this project is to leverage the power of machine learning to automate the sentiment analysis process and enable efficient and accurate classification of product reviews into positive, negative, or neutral sentiments. By employing machine learning algorithms to analyse textual data, patterns can be extracted, sentiments can be identified, and meaningful insights can be derived at scale.

The importance of sentiment analysis in the realm of product reviews is multifold. Firstly, sentiment analysis can provide businesses with invaluable insights into customer perceptions

and sentiments regarding their products. Understanding customer sentiment allows companies to identify areas of improvement, refine marketing strategies, and enhance overall customer satisfaction.

Moreover, sentiment analysis empowers consumers by offering them a summarised sentiment analysis score for a product. With this information at hand, customers can make informed decisions, thereby saving time and minimising the risk of purchasing products that do not meet their expectations. By providing quick access to sentiment analysis results, machine learning-based sentiment analysis enables consumers to navigate the vast array of available options efficiently.

Another motivation for this project is to contribute to the advancement of machine learning techniques in the field of sentiment analysis. While sentiment analysis has made significant strides in recent years, there are still challenges to overcome. By exploring various preprocessing techniques, feature extraction methods, and machine learning algorithms, this project aims to push the boundaries of sentiment analysis accuracy and robustness, particularly in the context of product reviews.

Ultimately, the outcomes of this project can benefit both businesses and consumers. Businesses can leverage the insights gained from sentiment analysis to improve product development, refine marketing strategies, and cultivate customer loyalty. Consumers, on the other hand, can make more informed purchasing decisions, save time, and enhance their overall satisfaction with products.

In conclusion, the motivation behind this project lies in addressing the challenges faced by consumers in navigating the vast amount of product reviews and extracting meaningful insights. By leveraging machine learning techniques for sentiment analysis, this project aims to automate the process, enhance the accuracy of sentiment classification, and contribute to the advancement of sentiment analysis in the context of product reviews. Ultimately, the project seeks to empower businesses and consumers alike by providing valuable insights and facilitating informed decision-making processes.

1.4 Objective

The overall objective of this project is to leverage the exponentially growing fields of e-commerce and machine learning, specifically focusing on sentiment analysis, to extract valuable insights from product reviews. The project aims to achieve the following objectives:

1. Perform data analysis on product reviews: The project involves collecting and analysing a significant volume of product reviews from various online platforms. Through data analysis techniques, such as data cleaning, preprocessing, and exploratory analysis, useful insights can be extracted. These insights can provide valuable information to sellers and companies regarding customer preferences, satisfaction levels, and areas of improvement.

2. Perform sentiment analysis using various machine learning algorithms: Sentiment analysis involves classifying the sentiment expressed in product reviews as positive, negative, very positive, very negative or neutral. This project aims to explore and implement different machine learning algorithms, such as Support Vector Machines (SVM), Naïve Bayes, and Random Forests, to perform sentiment analysis on the collected data. The performance and accuracy of each algorithm will be evaluated and compared.

3. Measure and analyse the accuracy and performance of sentiment analysis models: The project involves assessing the accuracy and performance metrics of the sentiment analysis models. Standard evaluation measures, such as accuracy, precision, recall, and F1-score, will be used to evaluate the models' effectiveness in classifying sentiments accurately. This analysis will provide insights into the strengths and weaknesses of each model and guide the selection of the most accurate one.

4. Fine-tune the selected model for improved accuracy: Once the most accurate sentiment analysis model is identified, further optimization will be performed to enhance its accuracy. Fine-tuning techniques, such as hyper-parameter tuning and feature selection, will be employed to optimise the model's performance. This iterative process aims to achieve the highest possible accuracy in sentiment classification.

5. Deploy the sentiment analysis model as a website: The final objective of the project is to deploy the selected and optimised sentiment analysis model as a functional website. This website will allow users to input their product reviews and obtain the sentiment classification results. By providing an accessible and user-friendly platform, the project aims to make sentiment analysis easily accessible to a wide range of users, including sellers, customers, and businesses.

By achieving these objectives, this project aims to bridge the gap between e-commerce and machine learning by utilising sentiment analysis to extract meaningful insights from product

reviews. The findings can inform business strategies, product development decisions, and customer engagement initiatives, leading to improved customer satisfaction, enhanced decision-making processes, and ultimately, increased success in the dynamic e-commerce landscape.

1.5 Scope

A project on sentiment analysis of product reviews is feasible and doable, although there are some challenges to consider. Collecting and preprocessing a diverse dataset of product reviews can be time-consuming, and labelling the data for sentiment analysis requires manual effort or alternative approaches. Selecting and training an appropriate model, addressing domain-specific challenges, and evaluating model performance are important steps in the project. Scalability and real-time processing considerations should be taken into account, along with ethical considerations regarding privacy, security, and potential biases. With careful planning and implementation, sentiment analysis on product reviews can be successfully conducted.

1. Ambiguous Sentiment: Some product reviews may express mixed opinions or exhibit ambiguous sentiment, making it challenging to assign a clear positive or negative label. Dealing with such cases requires advanced techniques like aspect-based sentiment analysis or fine-grained sentiment classification.

2. Generalisation across Products: Sentiment analysis models trained on one product category may struggle to generalise well to other categories. Each product domain has unique characteristics and vocabulary, necessitating domain-specific adaptation or transfer learning techniques to achieve accurate sentiment analysis across different product types.

3. Evolving Language and Trends: Language usage and sentiment expressions change over time, with new trends and phrases emerging. To maintain an effective sentiment analysis model, continuous monitoring and updating based on evolving language patterns and emerging trends are essential. Regular retraining or adaptation of the model is required to stay up-to-date with the changing nature of customer sentiment.

4. Handling Unstructured Data: Product reviews can contain unstructured data in various formats, such as text, images, or videos. Integrating and processing these diverse data types poses technical challenges. Effective feature extraction methods, multimodal analysis techniques, and robust data preprocessing pipelines are needed to handle different data formats effectively in sentiment analysis.

1.6 Organisation of the Report

The report is organised into several chapters and sections, each addressing a specific aspect of the research. Here is the breakdown of the organisation of the report:

CHAPTER 1: INTRODUCTION

This research focuses on a specific problem in the field and aims to provide a comprehensive understanding of its significance. The problem is clearly defined, and the research question or objective is identified. The motivation behind this work is discussed, emphasising its importance and potential impact. The specific goals and objectives of the research are stated, providing clarity on what the study aims to achieve. The scope of the research is defined, including the boundaries and limitations. The report's organisation is outlined, providing a brief description of each chapter or section.

CHAPTER 2: LITERATURE SURVEY:

Conducts a comprehensive review of relevant literature and summaries existing research, theories, and findings. Identifies gaps in knowledge or areas for further investigation.

CHAPTER 3: METHODOLOGY:

This section provides an overview of various machine learning algorithms applied to the research problem. It explains the principles of Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest Classifier algorithms and describes how each algorithm is applied to the research problem. The section includes a model comparison, discussing the performance, advantages, and disadvantages of each algorithm. It further details the implementation process, including the tools, software, and datasets used. The chosen model evaluation metric is introduced, explaining how it assesses the accuracy and effectiveness of the models.

CHAPTER 4: RESULTS & DISCUSSION:

In this section, the collected or obtained data is analysed, and relevant statistical analysis or findings are presented. The accuracy of the implemented models is evaluated and compared, providing quantitative assessments of their performance. The process of fine-tuning the models is discussed, including adjustments made to optimise their performance. Finally, the section presents the final results of the research, summarising the findings and outcomes of the study.

CHAPTER 5: CONCLUSION & FUTURE WORK:

The conclusion section summarises the main findings and conclusions of the study, restating the research objectives and addressing whether they were achieved. It also highlights potential areas for future research and improvement, suggesting possible extensions or modifications to the current study.

REFERENCES:

Provides a list of all the cited references in the report.

APPENDIX A:

Includes any supplementary information that supports the main report. Contains additional data snippets.

CHAPTER 2

LITERATURE SURVEY

Literature Survey I

Sayyed Johar, Samara Mubeen "Sentiment Analysis on Large Scale Amazon Product Reviews", International Journal of Scientific Research in Computer Science and Engineering Vol.8, Issue.1, pp.07-15, February (2020)

The paper begins by highlighting the importance of customer reviews in the e-commerce landscape, particularly on platforms like Amazon. It emphasises the need for a model that can analyse and classify these reviews to provide insights into customer sentiment towards products.

The authors propose a supervised learning model that can handle large amounts of Amazon product review data. The model is designed to provide a statistical report on the number of reviews where customers are not satisfied with a specific feature of an Amazon product. The aim is to compare different classification methods in sentiment analysis on Amazon dataset to see if it works better in some particular aspects.

The paper discusses two main approaches to feature extraction: the Bag of Words approach and the TF-IDF & Chi-square approach. The authors use both manual and active learning approaches to label the datasets. In the active learning process, different classifiers are used to provide accuracy until reaching a satisfactory level.

The paper also discusses the use of Naive Bayes and Support Vector Machine (SVM) for classification. Naive Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. SVMs are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis.

The authors conclude that their system can accurately analyse Amazon product reviews and provide valuable insights for both customers and manufacturers. They suggest that future work could involve implementing the system as a web or mobile application with a user-interactive front end.

Literature Survey II

Arwa S. M. AlQahtani "Product Sentiment Analysis for Amazon Reviews", International Journal of Computer Science & Information Technology Vol.13, Issue.3, pp.15-31, June (2021)

Evaluation Metrics: The authors used several evaluation metrics including Precision, Recall, Accuracy, F-score, and Cross-entropy or log loss. These metrics are calculated using the values from the confusion matrix (True Positive, False Positive, False Negative, True Negative).

Results: The authors applied several models including Logistic Regression, Naïve Bayes, Random Forest, Bidirectional Long-Short Memory (Bi-LSTM) with GloVe embedding and joint-learned embedding, and Bidirectional Encoder Representations from Transformers (BERT).

BERT model achieved an excellent result in multi-class classification and binary classification, with accuracy of 94% and 98%, respectively.

Bi-LSTM with joint-learned embedding also provides a very good result, with accuracy of 93% for multi-class classification and 97% for binary classification.

Random Forest with word embedding (GloVe) outperforms other baseline models, LR and NB, with accuracy of 90% for multi-class classification and 94% for binary classification.

Dataset: The dataset used in the study was Amazon mobile phone reviews. The authors found that 67.5% of reviews are "Positive," 24.6% are "Negative," and 7.97% are "Neutral." The brand with the maximum number of reviews was Samsung with 58,395 reviews.

Limitations and Future Work: The authors acknowledge the limitations of their study and suggest future work could involve using more sophisticated machine learning models and different feature extraction approaches for text classification. They also suggest fine-tuning the pre-trained BERT model for the sentiment analysis task on the Amazon mobile.

Literature Survey III

Ms. Jyoti Budhwar and Prof. Sukhdip Singh "Sentiment Analysis based Method for Amazon Product Reviews", International Journal of Engineering Research & Technology Vol.9, Issue.8, pp.21-25, June (2021)

The authors propose a hybrid approach combining Naïve Bayes, K-Nearest Neighbours (KNN), and Long Short-Term Memory (LSTM) mechanisms. The Naïve Bayes approach is used for classification, KNN aids in grouping, and the dataset is trained using an LSTM-based model to improve accuracy. The research aims to resolve the issues faced during sentiment analysis in previous studies.

This sentiment analysis is important as it assists in text mining and computational linguistics, studying correlations among Amazon product reviews and considering product ratings. The authors have taken into account traditional machine learning algorithms, SVM, K-nearest neighbour mechanisms, and deep neural networks with Recurrent Neural Networks (RNN) to provide a better solution for sentiment analysis.

The authors recognise that finding and monitoring online opinion pages and interpreting the information found in them remains a challenge due to the abundance of different sites and the large amount of opinionated text in long forum posts and blogs.

The research paper mentions that the data was uneven, which required additional sampling of data. This could be seen as a limitation as the unevenness of the data could have affected the results of the sentiment analysis.

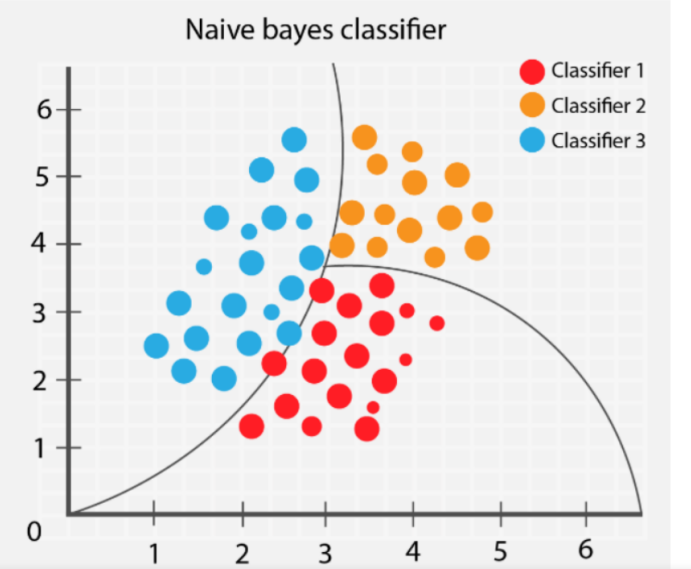
The paper also mentions that the availability of repeated samples could make the design overfit. Overfitting is a modelling error that occurs when a function is too closely fit to a limited set of data points. This could be a limitation as it might make the model less accurate when applied to new, unseen data.

Sno.	Author / Year	Objective	Methodology	Limitations
1	R. Collobert / 2011	To implement natural language processing from scratch.	Machine Learning	Research failed to provide solution for semantic reviews.
2	K. Dave / 2003	Performing opinion extraction and semantic classification of product reviews.	Semantic Classification	Research has not considered optimised solution.
3	M. S. Elli / 2018	To propose Amazon reviews, business analysis with sentiment analysis.	Analysing the Sentiment	Lack of accuracy in prediction.
4	S. Hota / 2018	Proposing Knn classifier based approach for multi-class sentiment analysis of twitter data.	Ken Classifier	This work is suffering from performance issues.
5	B. Liu / 2012	To perform opinion mining and sentiment analysis.	Sentiment Analysis	Lack of accuracy and flexibility.
6	C. Rain / 2013	Implementing sentiment analysis in amazon reviews using probabilistic machine learning.	Machine Learning	Research is providing solution on the basis of probability that leads to degradation in accuracy.
7	R. Soccer / 2013	Presenting recursive deep models for semantic compositionally over a sentiment treebank.	Recursive Deep Model	Recursive deep model waste lot of time during training.
8	Y. Xu / 2015	Implementing sentiment analysis of yelp ratings based on text reviews.	Sentiment Analysis	Research is not providing wide scope.

CHAPTER 3

METHODOLOGY

3.1 Naive Bayes



Naive Bayes is a powerful probabilistic classification algorithm rooted in Bayes' theorem. It leverages the assumption that the features in a dataset are conditionally independent of each other given the class label. This assumption simplifies the calculation of the posterior probability, making the algorithm computationally efficient and highly scalable.

One of the notable advantages of Naive Bayes is its ability to handle large datasets and high-dimensional feature spaces. This makes it well-suited for tasks involving text classification, where the number of features can be vast, such as spam filtering and sentiment analysis.

There are three main variants of Naive Bayes algorithms: Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes, each tailored to specific types of data.

1. Gaussian Naive Bayes:

Gaussian Naive Bayes assumes that the features follow a Gaussian distribution. It is suitable for continuous or real-valued features. In this variant, the algorithm estimates the mean and standard deviation of each feature for each class, and uses them to compute the likelihood of a feature value given the class. It then combines these likelihoods with prior probabilities to calculate the posterior probability.

2. Multinomial Naive Bayes:

Multinomial Naive Bayes is designed for discrete features that represent counts or frequencies. It is commonly used for text classification tasks, where features can be the

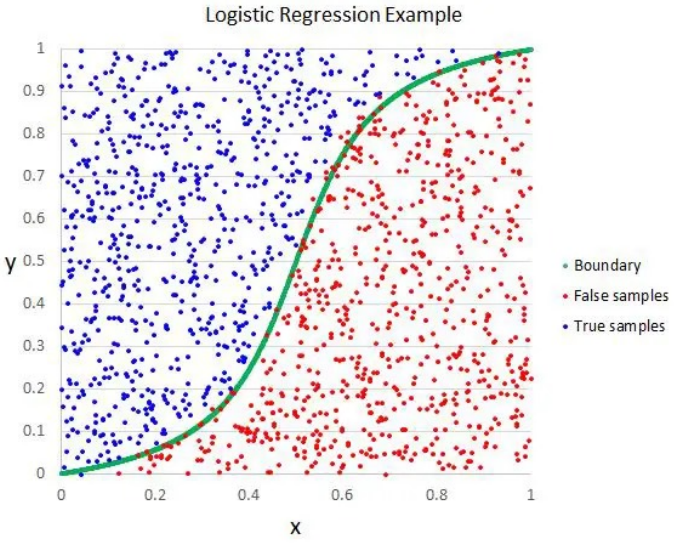
occurrence or frequency of words in a document. In this variant, the algorithm computes the likelihood of each feature value given the class as a probability distribution, typically using techniques like Laplace smoothing. It then combines these likelihoods with prior probabilities to determine the posterior probability.

3. Bernoulli Naive Bayes:

Bernoulli Naive Bayes is also suitable for discrete features but assumes that they follow a Bernoulli distribution, meaning they are binary or Boolean variables. It is often employed in tasks where the presence or absence of a feature is essential, such as document classification based on the presence of specific keywords. This variant calculates the likelihood of each feature value given the class as a probability of its presence or absence. As with the other variants, it combines these likelihoods with prior probabilities to obtain the posterior probability.

In summary, Naive Bayes is a versatile classification algorithm that makes strong assumptions about feature independence to simplify calculations. It is efficient, scalable, and particularly effective for text classification tasks. Its three main variants, Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes, cater to different types of data and provide flexibility in various applications.

3.2 Logistic Regression



Logistic Regression is a statistical classification algorithm that aims to model the relationship between the features of a dataset and the probability of belonging to a specific class label. Unlike linear regression, which predicts continuous values, logistic regression predicts the likelihood of an instance belonging to a particular class, typically by applying a logistic or sigmoid function to the linear combination of the features.

The logistic function, also known as the sigmoid function, maps any real-valued input to an output between 0 and 1. This property is crucial for logistic regression as it allows the algorithm to interpret the output as a probability. The logistic function is defined as:

sigmoid(z) = 1 / (1 + e^(-z))

Here, 'z' represents the linear combination of the features and their corresponding weights. The logistic function transforms 'z' into a probability value that quantifies the likelihood of an instance belonging to the positive class. The probability of the negative class can be calculated as 1 minus the probability of the positive class.

Training a logistic regression model involves estimating the optimal weights that minimise the difference between the predicted probabilities and the actual class labels in the training data. This process is typically accomplished through maximum likelihood estimation or minimising a cost function, often referred to as the logistic loss or binary cross-entropy loss.

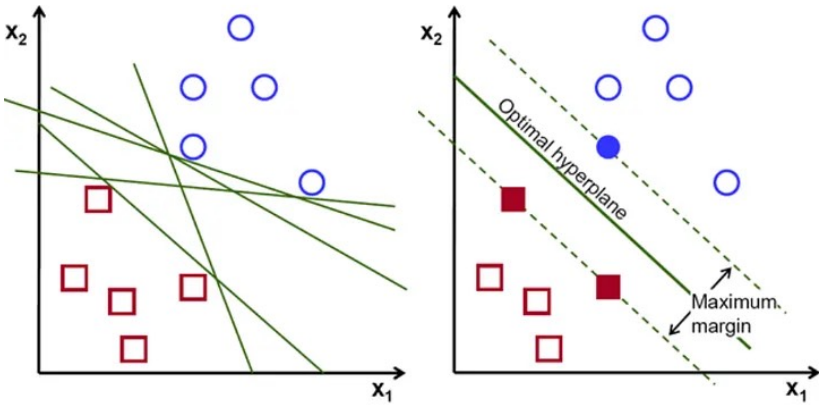
Various optimization techniques can be employed to find the optimal weights in logistic regression. One common method is gradient descent, which iteratively adjusts the weights by taking steps proportional to the negative gradient of the cost function. Other optimization algorithms, such as stochastic gradient descent, can also be utilised to enhance efficiency and handle large datasets.

Logistic regression is a versatile algorithm capable of handling both binary classification problems, where there are two classes to predict, and multi-class classification problems, where there are more than two classes. For multi-class classification, logistic regression can be extended using techniques like one-vs-rest or softmax regression.

The algorithm finds extensive application in a wide range of fields, including medical diagnosis, finance, marketing, and social sciences. In medical diagnosis, logistic regression can be employed to predict the presence or absence of a disease based on various clinical features. In finance, it can be utilised to evaluate credit risk or forecast stock market movements. In marketing, logistic regression can assist in customer segmentation or predicting customer churn.

In summary, logistic regression is a statistical classification algorithm that models the relationship between features and the probability of class labels using a logistic function. It is widely used due to its interpretability, flexibility in handling binary and multi-class problems, and its applicability across diverse domains.

3.3 Support Vector Machine [SVM]



Support Vector Machines (SVM) is a robust and flexible classification algorithm that is widely recognised for its effectiveness in separating data points in high-dimensional feature spaces using a hyperplane. SVM is particularly notable for its ability to handle linear and nonlinear decision boundaries while aiming to maximise the margin between the hyperplane and the closest data points. This margin maximisation helps reduce the risk of overfitting, enabling improved generalisation performance on unseen data.

The core idea of SVM is to find an optimal hyperplane that best separates the data points belonging to different classes. The hyperplane is defined as a subspace with one dimension less than the feature space. For example, in a two-dimensional feature space, the hyperplane is a line, while in a three-dimensional feature space, it is a plane. In higher dimensions, the hyperplane becomes a hyperplane.

The goal of SVM is to find the hyperplane that maximises the margin, which is the distance between the hyperplane and the closest data points from each class. The data points that lie on the margin are called support vectors, as they play a crucial role in defining the hyperplane. SVM optimises this margin by solving a quadratic programming problem, seeking to minimise the classification error and simultaneously maximise the margin.

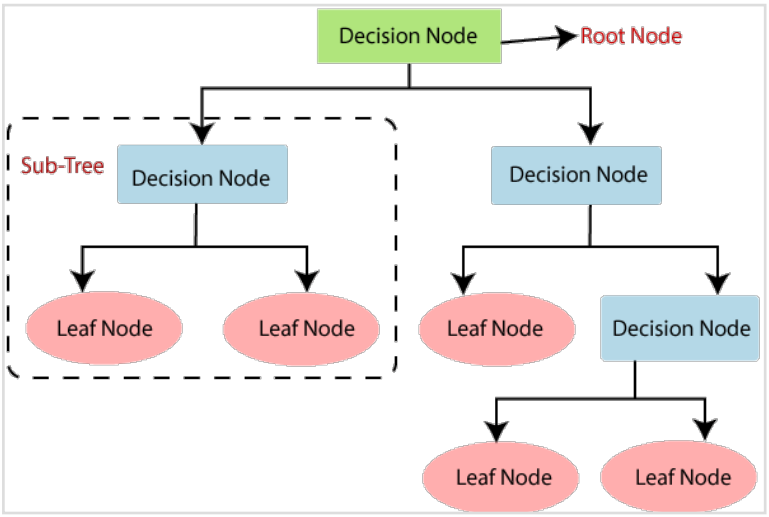
One of the significant advantages of SVM is its ability to handle nonlinear decision boundaries. SVM achieves this by employing a kernel trick, which implicitly maps the data into a higher-dimensional feature space where a linear separation becomes possible. The kernel function measures the similarity between pairs of data points in the original feature space. Commonly used kernel functions include the linear kernel, polynomial kernel, and radial basis function (RBF) kernel.

SVM is not limited to binary classification tasks but can also be extended to handle multi-class problems. One approach is the one-vs-rest strategy, where separate SVM models are trained for each class against the rest of the classes. Another approach is the one-vs-one strategy, where SVM models are trained for each pair of classes. Both strategies can effectively handle multi-class classification by combining the results of the individual SVM models.

SVM has demonstrated remarkable success in various domains and applications. In image classification, SVM has been employed for tasks such as object recognition and facial expression analysis. In text classification, SVM has been widely used for sentiment analysis, spam detection, and document categorisation. Furthermore, SVM has found applications in bioinformatics, where it has been utilised for tasks like protein classification and gene expression analysis.

In summary, SVM is a powerful and versatile classification algorithm that separates data points in high-dimensional feature spaces using a hyperplane. By maximising the margin between the hyperplane and the closest data points, SVM provides better generalisation and mitigates overfitting. It can handle both linear and nonlinear decision boundaries through the use of kernel functions. SVM is applicable to multi-class problems and has been successfully employed in diverse fields, including image classification, text classification, and bioinformatics.

3.4 Decision Tree



A Decision Tree is a popular and intuitive classification algorithm that constructs a tree-like model by recursively partitioning the feature space through a series of binary splits. Each node in the tree represents a decision based on the value of a specific feature, and each leaf node corresponds to a class label. The process of building a Decision Tree involves selecting the most informative features and determining optimal split points that maximise the homogeneity or purity of the resulting subsets.

One of the significant advantages of Decision Trees is their ability to handle both categorical and numerical features. For categorical features, the splits are based on the different categories, while for numerical features, the splits are determined by comparing the feature values with a threshold. This versatility allows Decision Trees to accommodate a wide range of data types and make decisions based on diverse feature characteristics.

Another advantage of Decision Trees is their interpretability. The resulting tree structure can be easily understood and interpreted, as it mirrors human decision-making processes.

Decision Trees can be graphically represented, enabling users to follow the path from the root to the leaves and observe the feature-based decisions made at each node. This transparency makes Decision Trees valuable in domains where explainability and comprehensibility are crucial.

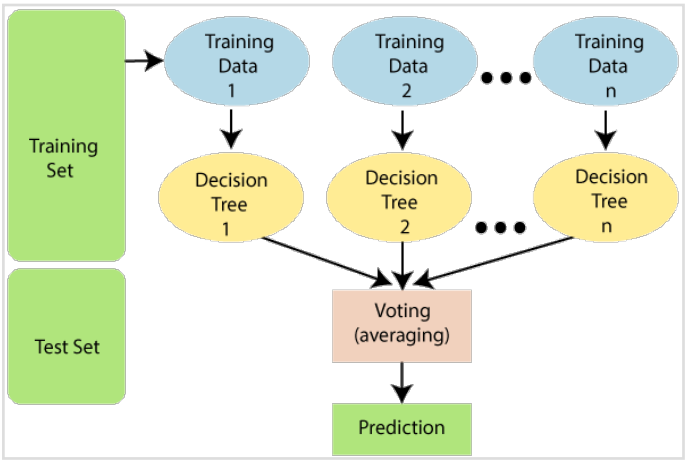
However, Decision Trees can be sensitive to small variations in the data and prone to overfitting, particularly for complex problems. Overfitting occurs when the model captures noise or idiosyncrasies in the training data, leading to poor generalisation performance on unseen data. Decision Trees have a tendency to create overly complex trees that perfectly fit the training data but may fail to generalise well to new instances.

To address the overfitting issue, various strategies can be employed. One common approach is pruning, which involves post-processing the tree by removing or collapsing certain nodes and branches. Pruning helps simplify the tree and reduce overfitting by favoring smaller, more generalisable trees. Other techniques include setting constraints on the maximum depth of the tree or the minimum number of samples required to split a node.

To mitigate the limitations of a single Decision Tree, ensemble methods such as Random Forests and Gradient Boosting are often employed. These methods combine multiple Decision Trees to make predictions, aggregating the outputs of individual trees to achieve better performance and increased robustness.

In summary, Decision Trees are simple and intuitive classification algorithms that build tree-like models through a series of binary splits on the features. They can handle both categorical and numerical features and provide interpretability. However, Decision Trees can be sensitive to small variations in the data and prone to overfitting, especially for complex problems. Strategies such as pruning and ensemble methods can be employed to address these limitations and enhance the performance of Decision Tree models.

3.5 Random Forest Classifier



Random Forest is a powerful ensemble classification algorithm that leverages the combination of multiple Decision Trees to enhance accuracy and robustness. It overcomes some of the limitations of individual Decision Trees by generating an ensemble of diverse trees and aggregating their predictions to make the final decision.

The process of building a Random Forest involves creating a collection of Decision Trees, each trained on a different subset of the original data. Random subsets, called bootstrap samples, are created by randomly selecting instances from the original dataset with replacement. Additionally, for each tree, only a subset of features is considered at each split. This feature subsampling further introduces randomness and diversity into the model.

During the training phase, each Decision Tree in the Random Forest independently learns patterns and makes predictions based on the subset of data and features it was trained on. When making predictions for a new instance, each tree in the ensemble produces its own prediction. The final decision of the Random Forest is determined by aggregating the predictions of all the trees, either through majority voting (for classification tasks) or averaging (for regression tasks).

Random Forest offers several advantages over individual Decision Trees. Firstly, it can handle high-dimensional feature spaces effectively. By considering random subsets of features at each split, Random Forest mitigates the risk of relying too heavily on a single feature, resulting in better feature importance estimation and reducing the impact of irrelevant or noisy features.

Secondly, Random Forest can handle noisy data more robustly. Since each tree in the ensemble is trained on a different bootstrap sample, the impact of noisy or outlier instances is reduced. The averaging or voting scheme used for prediction further smooths out the noise, leading to improved generalisation performance.

One of the key benefits of Random Forest is its ability to address overfitting and bias. By creating an ensemble of diverse trees, Random Forest is less prone to overfitting than individual Decision Trees. The diversity in the ensemble helps capture different aspects of the data, leading to a more robust and accurate model.

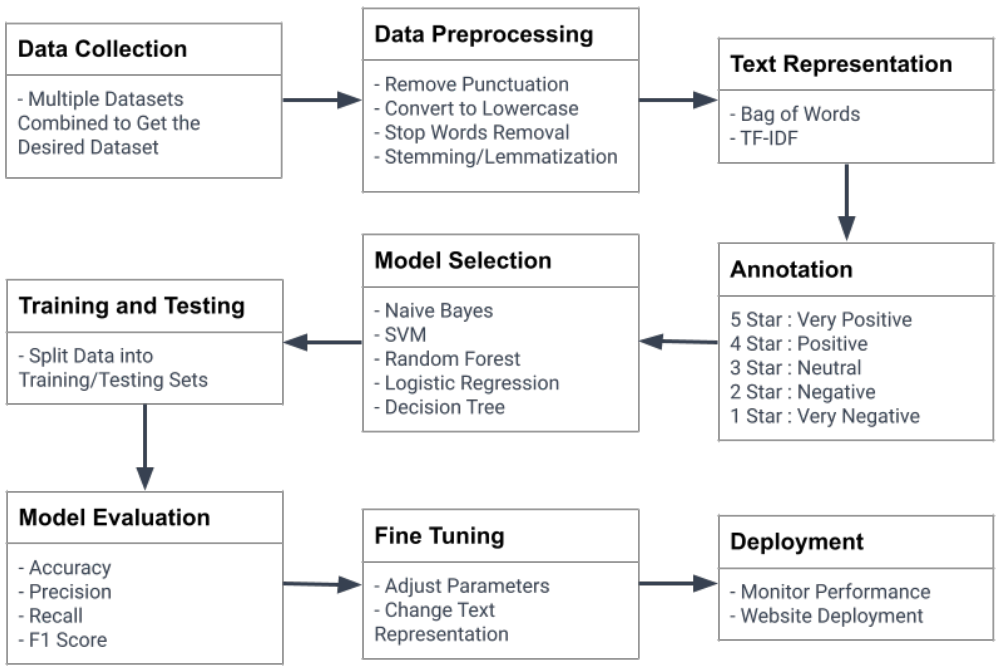
Random Forest has found widespread application in various domains. In finance, it has been used for tasks like credit scoring, fraud detection, and stock market prediction. In ecology, Random Forest has been applied to species classification, habitat modelling, and biodiversity assessment. In bioinformatics, it has been utilised for gene expression analysis, protein structure prediction, and disease diagnosis.

In summary, Random Forest is an ensemble classification algorithm that combines multiple Decision Trees to improve accuracy and robustness. It handles high-dimensional feature spaces, noisy data, and mitigates overfitting and bias. Random Forest has been successfully employed in diverse domains such as finance, ecology, and bioinformatics, contributing to improved predictive modelling and data analysis.'

3.6 Model Comparison

Algorithm	Key Feature	Pros	Cons	Applications
Naive Bayes	Assumption of feature impedance	Efficient and scalable	Oversimplified assumptions	Text classification, spam filtering, sentiment analysis
Logistic Regression	Probabilistic classification	Interpretable and easily visualized	Limited handling of nonlinearity	Medical diagnosis, finance, marketing
SVM	Hyperplane separation in high-dimensional space	Handles both linear and nonlinear boundaries	Sensitive to hyperparameter tuning	Image classification, text classification, bioinformatics
Decision Trees	Simple and intuitive	Interpretable and easily visualized	Prone to overfitting, sensitive to small variations in data	Medical diagnosis, finance, social sciences
Random Forests	Ensemble of Decision Trees	Reduces overfitting and bias, handles high-dimensional data	Less interpretable than individual Decision Trees	Finance, ecology, bioinformatics

3.7 Implementation



1. Data Collection: The Data Collection Process is iterative, continuous experimentation with different datasets is necessary. This project uses 2 datasets that are combined so that a bigger dataset can be obtained. Both the datasets are available for free use on the internet.

2. Data Preprocessing: Clean and preprocess the data by performing tasks such as removing punctuation, stop words, and special characters, converting all text to lowercase, stemming or lemmatising words, etc. The goal of this step is to reduce the dimensionality of the data and prepare it for analysis by machine learning algorithms. Scikit-learn is used in this project.

3. Text Representation: Represent the text data in a numerical form. The purpose of this step is to convert the text data into a format that can be easily processed by machine learning algorithms.

4. Annotation: Annotate the collected product reviews by labelling each review as positive, negative, or neutral. Pre-labelled sentiment analysis models can be used to do this. The goal of this step is to create a labeled dataset that can be used for training and testing the machine learning model. In this project the star rating given by the customer for a product sold online is used as the label for that particular review.

5. Model Selection: Choose a machine learning model that is suitable for sentiment analysis, such as Naive Bayes, SVM, Random Forest, Logistic Regression and Decision Tree. Consider factors such as accuracy, computational complexity, and interpretability when selecting the model.

6. Training and Testing: Train the selected machine learning model on the preprocessed and annotated data, then test the model on a separate set of reviews to evaluate its performance. It is important to split the data into training and testing sets randomly to avoid overfitting or under-fitting the model.

7. Model Evaluation: Evaluate the performance of the model by computing metrics such as accuracy, precision, recall, and F1 score. These metrics will provide insight into how well the model is performing on different classes of sentiment.

8. Fine-tuning: Fine-tune the model if needed, by adjusting the parameters, changing the representation of the text data, or using a different machine learning model. The goal of this step is to improve the performance of the model and achieve the desired accuracy.

9. Deployment: Deploy the model in a real-world application, such as a web app, to perform sentiment analysis on incoming product reviews in real-time. It is important to monitor the

performance of the deployed model and make changes as needed to ensure that it continues to perform well.

3.8 Model Evaluation Metric

1. Accuracy: Accuracy is a commonly used metric that measures the overall correctness of the model's predictions. It calculates the proportion of correctly classified instances (both positive and negative) out of the total number of instances in the dataset. The formula for accuracy is:
   Accuracy = (True Positives + True Negatives) / Total Instances
   This metric provides a general overview of the model's performance, indicating how often the model predicts the correct class.

2. Precision: Precision focuses on the correctness of positive predictions made by the model. It calculates the proportion of true positive predictions (correctly predicted positives) out of the total positive predictions. The formula for precision is:
   Precision = True Positives / (True Positives + False Positives)
   Precision is particularly relevant when the cost of false positives (incorrectly predicted positives) is high. It helps assess the model's ability to make accurate positive predictions.

3. Recall (Sensitivity or True Positive Rate): Recall measures the model's ability to capture all the positive instances in the dataset. It calculates the proportion of true positive predictions (correctly predicted positives) out of the actual positive instances. The formula for recall is:
   Recall = True Positives / (True Positives + False Negatives)
   Recall is important when the cost of false negatives (missed positives) is high. It helps evaluate how well the model identifies positive instances.

4. F1 Score: The F1 score is a balanced measure that combines precision and recall into a single metric. It is the harmonic mean of precision and recall, providing a balanced assessment of the model's performance. The formula for the F1 score is:
   F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
   The F1 score ranges from 0 to 1, with higher values indicating better model performance. It is useful when there is an uneven class distribution or when both false positives and false negatives are important to consider.

CHAPTER 4

RESULT & DISCUSSION

4.1 Data Analysis

Dataset Visualisation Before Data Processing: Dataset after removing unwanted columns in both the datasets and combining both the datasets. After combining the two datasets the number of reviews have grown to about 600 thousand reviews and the columns have been reduced to the most needed ones only.

The Columns include : Column ID, Product ID, Score (which is basically the star rating given by the user), User ID, Summary (which is basically the title of the review given by the user.)

The datasets that were used :

Amazon Fine Foods Reviews + Consumer Reviews of Amazon Product

(Both the datasets from Kaggle)

	Id	ProductId	Score	Text	UserId	Summary
0	1.0	B001E4KFG0	5	I have bought several of the Vitality canned d...	NaN	NaN
1	2.0	B00813GRG4	1	Product arrived labeled as Jumbo Salted Peanut...	NaN	NaN
2	3.0	B000LQOCH0	4	This is a confection that has been around a fe...	NaN	NaN
3	4.0	B000UA0QIQ	2	If you are looking for the secret ingredient i...	NaN	NaN
4	5.0	B006K2ZZ7K	5	Great taffy at a great price. There was a wid...	NaN	NaN
...	...	...	...	...	...	...
28327	NaN	B018T075DC	5	I got 2 of these for my 8 yr old twins. My 11 ...	AVqklhxunnc1JgDc3kg_	Xmas gift
28328	NaN	B018T075DC	4	I bought this for my niece for a Christmas gif...	AVqklhxunnc1JgDc3kg_	yes it is a great tablet.
28329	NaN	B018T075DC	5	Very nice for light internet browsing, keeping...	AVqklhxunnc1JgDc3kg_	You get a lot for the price!
28330	NaN	B018T075DC	5	This Tablet does absolutely everything I want!...	AVqklhxunnc1JgDc3kg_	You get the entire World for less than $100!
28331	NaN	B018T075DC	4	At ninety dollars, the expectionations are low...	AVqklhxunnc1JgDc3kg_	You get what your paying for

596786 rows × 6 columns

Dataset Visualisation After Data Processing: Considering the star rating given by the user (Score) the dataset has been annotated with the Sentiment. The Sentiment are annotated as follows :

5 Stars - Very Positive

4 Stars - Positive

3 Stars - Neutral

2 Star - Negative

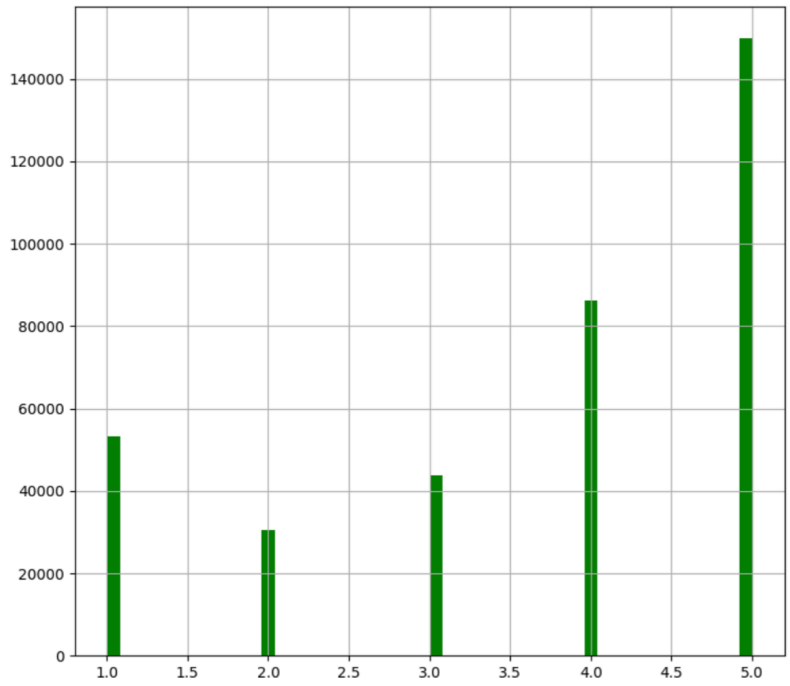
1 Star - Very Negative

The last column has the review after data processing where upper case letters have been converted to lowercase, plural words have been converted into singular words, etc. Dataset with additional Columns: Sentiment and text_stemmed column added.

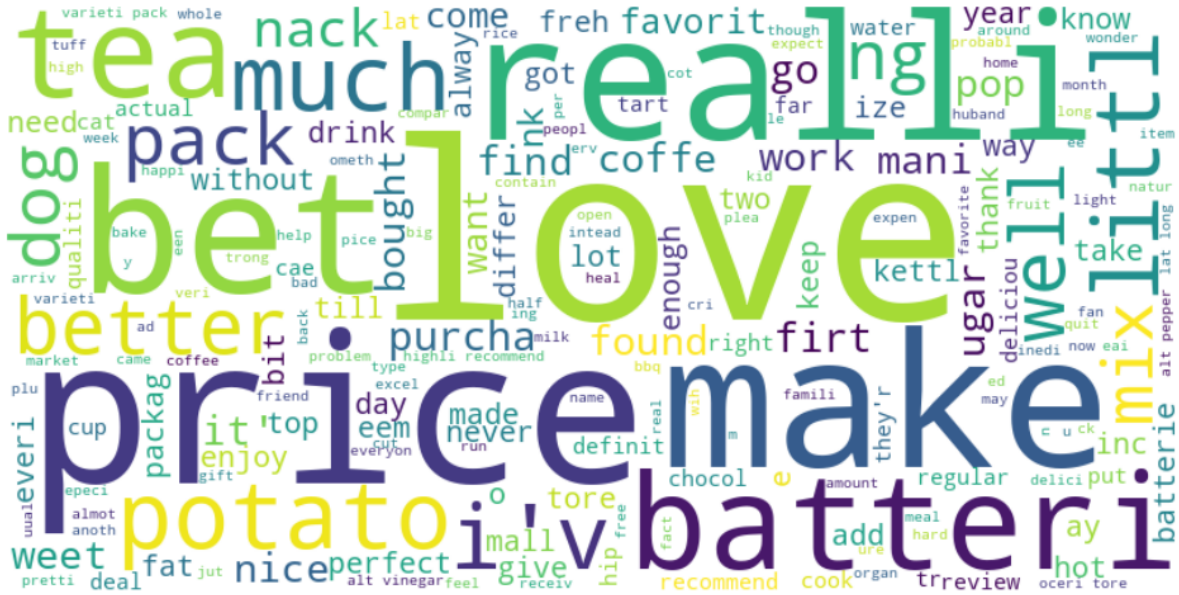
	Id	ProductId	Score	Text	UserId	Summary	Sentiment	text_stemmed
0	1.0	B001E4KFG0	5	I bought several Vitality canned dog food prod...	NaN	NaN	Very Positive	i bought sever vital can dog food product foun...
4	5.0	B006K2ZZ7K	5	Great taffy great price. There wide assortment...	NaN	NaN	Very Positive	great taffi great price. there wide assort yum...
6	7.0	B006K2ZZ7K	5	This saltwater taffy great flavors soft chewy....	NaN	NaN	Very Positive	thi saltwat taffi great flavor soft chewy. eac...
7	8.0	B006K2ZZ7K	5	This taffy good. It soft chewy. The flavors am...	NaN	NaN	Very Positive	thi taffi good. it soft chewy. the flavor amaz...
8	9.0	B000E7L2R4	5	Right I'm mostly sprouting cats eat grass. The...	NaN	NaN	Very Positive	right i'm mostli sprout cat eat grass. they lo...
...	...	...	...	...	...	...	...	...
28197	NaN	B018T075DC	1	This Kindle overloaded apps I never use. Despi...	AVqklhxunnc1JgDc3kg_	Too may preloaded useless apps	Very Negative	thi kindl overload app i never use. despit one...
28199	NaN	B018T075DC	1	I bought kindle fire 8 husband I use. Neither ...	AVqklhxunnc1JgDc3kg_	Unable to use it	Very Negative	i bought kindl fire 8 husband i use. neither o...
28254	NaN	B018T075DC	1	Freeze frequently... No way trouble shoot repa...	AVqklhxunnc1JgDc3kg_	Very poor.	Very Negative	freez frequently... no way troubl shoot repair...
28265	NaN	B018T075DC	1	cheap, run chrome stuff, returned store.	AVqklhxunnc1JgDc3kg_	was cheap, can not run chrome stuff, returned	Very Negative	cheap, run chrome stuff, return store.
28285	NaN	B018T075DC	1	Worked great awhile I unlock. After I'd enter ...	AVqklhxunnc1JgDc3kg_	Won't unlock	Very Negative	work great awhil i unlock. after i'd enter pin...

363767 rows × 8 columns

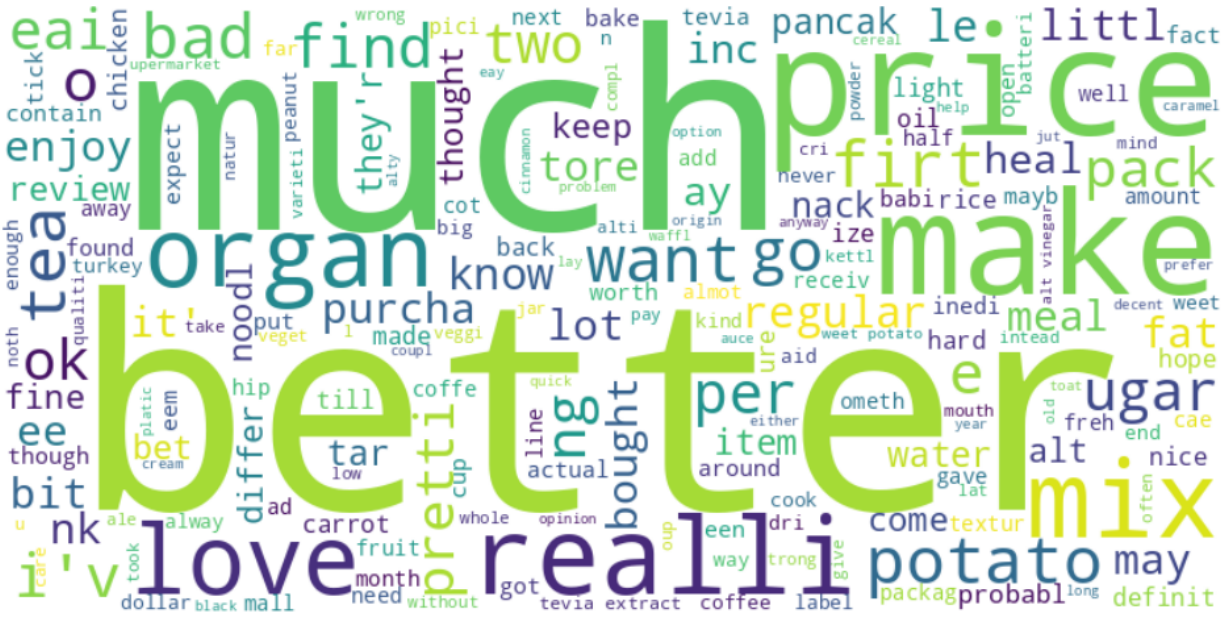
Distribution of Reviews Based on Star Rating :



WordCloud for Positive Reviews :



WordCloud for Neutral Reviews :



WordCloud for Negative Reviews :

4.2 Data Accuracy Analysis

accuracy = (true positives + true negatives) / (true positives + false positives + true negatives + false negatives)

precision = true positives / (true positives + false positives)

recall = true positives / (true positives + false negatives)

F1 score = 2 * (precision * recall) / (precision + recall)

Naive Bayes Accuracy :

	precision	recall	f1-score	support
Negative	0.00	0.00	0.00	28
Neutral	0.00	0.00	0.00	48
Positive	0.00	0.00	0.00	60
Very_Negative	0.00	0.00	0.00	36
Very_Positive	0.67	1.00	0.80	349
accuracy			0.67	521
macro avg	0.13	0.20	0.16	521
weighted avg	0.45	0.67	0.54	521

Naive Bayes Accuracy: 0.6698656429942419

Logistic Regression Accuracy :

	precision	recall	f1-score	support
Negative	0.00	0.00	0.00	28
Neutral	0.00	0.00	0.00	48
Positive	0.20	0.02	0.03	60
Very_Negative	0.00	0.00	0.00	36
Very_Positive	0.68	1.00	0.81	349
accuracy			0.67	521
macro avg	0.18	0.20	0.17	521
weighted avg	0.48	0.67	0.55	521

Logistic Regression Accuracy: 0.6717850287907869

SVM Accuracy :

	precision	recall	f1-score	support
Negative	0.67	0.07	0.13	28
Neutral	0.30	0.06	0.10	48
Positive	0.22	0.13	0.17	60
Very_Negative	0.32	0.31	0.31	36
Very_Positive	0.75	0.95	0.84	349
accuracy			0.68	521
macro avg	0.45	0.30	0.31	521
weighted avg	0.62	0.68	0.62	521

SVM Accuracy: 0.6794625719769674

Decision Tree Accuracy :

	precision	recall	f1-score	support
Negative	0.00	0.00	0.00	28
Neutral	0.22	0.17	0.19	48
Positive	0.19	0.22	0.20	60
Very_Negative	0.19	0.17	0.18	36
Very_Positive	0.71	0.73	0.72	349
accuracy			0.54	521
macro avg	0.26	0.26	0.26	521
weighted avg	0.53	0.54	0.54	521

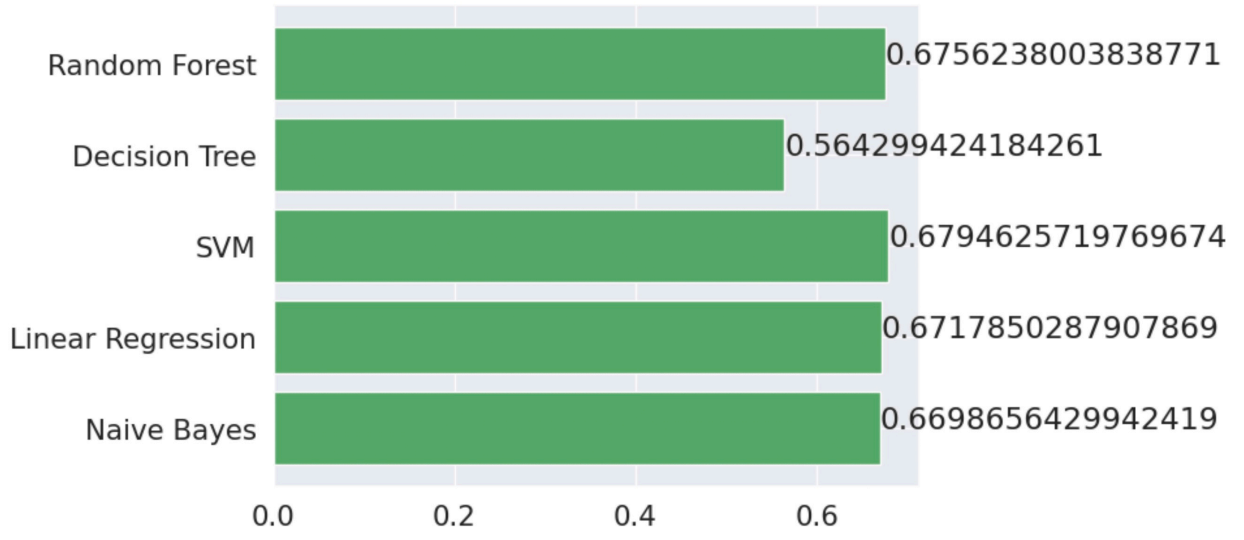
Decision Tree Accuracy: 0.5393474088291746

Random Forest Accuracy :

	precision	recall	f1-score	support
Negative	0.00	0.00	0.00	28
Neutral	1.00	0.02	0.04	48
Positive	0.00	0.00	0.00	60
Very_Negative	0.50	0.03	0.05	36
Very_Positive	0.68	1.00	0.81	349
accuracy			0.67	521
macro avg	0.44	0.21	0.18	521
weighted avg	0.58	0.67	0.55	521

Random Forest Classifier Accuracy: 0.6737044145873321

Model Accuracy Analysis :



In a comparative analysis of various machine learning algorithms, the accuracy scores were obtained for different models. Naive Bayes achieved an accuracy of 0.669, followed closely by logistic regression with a score of 0.671. SVM (Support Vector Machine) demonstrated slightly better performance with an accuracy of 0.679. However, the decision tree model yielded a lower accuracy of 0.539, indicating comparatively weaker predictive power. On the other hand, random forest exhibited a more promising accuracy score of 0.673, surpassing most other models in the evaluation. These results highlight the varying degrees of success achieved by different algorithms in capturing and understanding the underlying patterns in the data.

4.3 Fine Tuning

The process of fine-tuning a machine learning model involves selecting the best combination of hyper-parameters, which are parameters that are not learned from the data but rather set by the user before training the model.

In this example, a Support Vector Machine (SVM) model was used to classify text reviews into different sentiment categories. The pipeline was defined to include two steps - TF-IDF vectorisation and the SVM classifier.

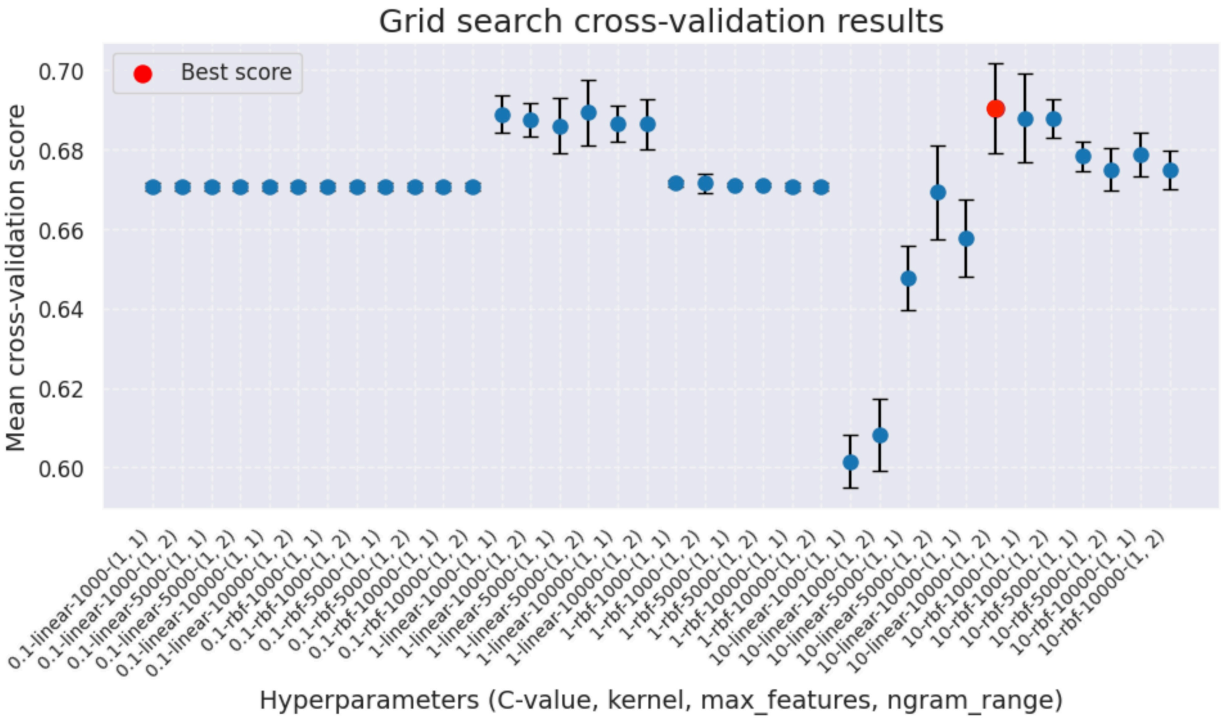
The hyper parameters that were tuned using a grid search were:

- tfidf__max_features: the maximum number of features (i.e., words) to be considered during the TF-IDF vectorisation process.
- tfidf__ngram_range: the range of n-grams (i.e., consecutive sequences of words) to be considered during the TF-IDF vectorisation process.
- svm__C: the regularisation parameter of the SVM model.
- svm__kernel: the kernel function to be used by the SVM model.

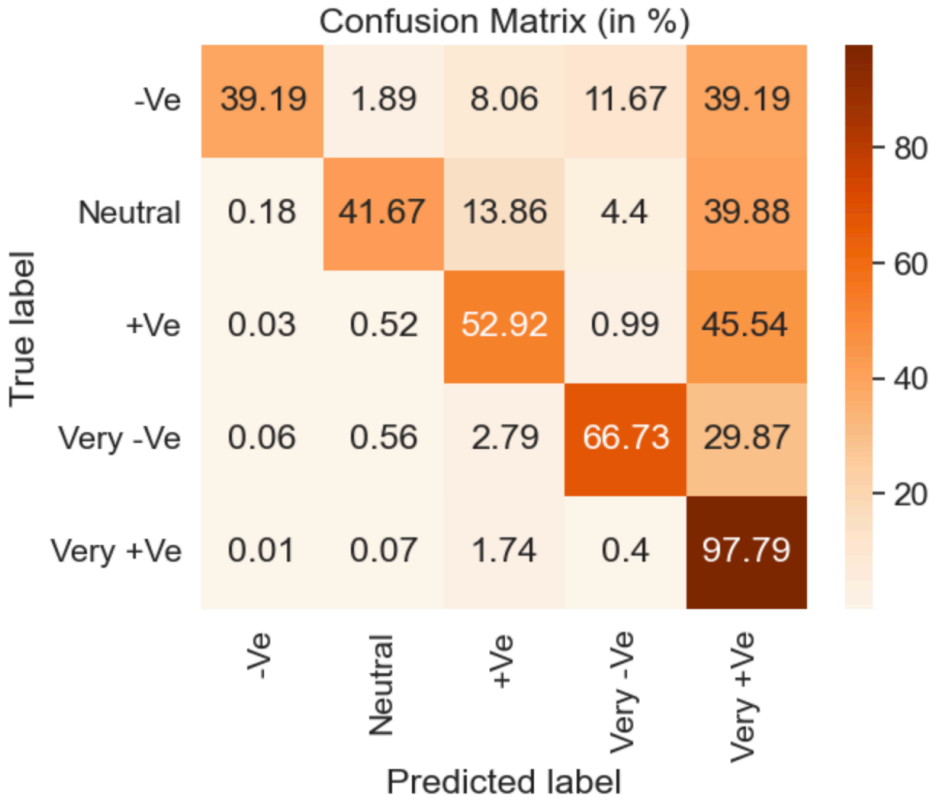
The grid search was performed using a 5-fold cross-validation approach, and the best combination of hyper-parameters was selected based on the highest average accuracy score across the different folds.

The initial accuracy score was 0.679, and after fine-tuning the hyper-parameters using the grid search, the **best accuracy score obtained was 0.690**. The best combination of hyper-parameters found by the grid search was 'svm__C': 10, 'svm__kernel': 'linear', 'tfidf__max_features': 10000, 'tfidf__ngram_range': (1, 2)'.

Overall, fine-tuning hyper-parameters is an essential step in building a robust and accurate machine learning model, as it allows the model to capture the underlying patterns in the data more effectively.



4.4 Result



Looking at the confusion matrix, we can make the following general observations:

1. The model performs relatively well in classifying instances as "Very Positive," with a high accuracy of 97.79%. This indicates that the model is successful in identifying instances with a strong positive sentiment.

2. On the other hand, the model struggles to accurately classify instances as "Very Negative." It misclassifies a significant portion of "Very Negative" instances as "Neutral" (1.89%) and "Very Positive" (39.19%). This suggests that the model may have difficulty distinguishing between negative and neutral sentiment, as well as between negative and strongly positive sentiment.

3. The model achieves moderate accuracy in classifying instances as "Neutral" (41.67%). However, it misclassifies a notable portion of "Neutral" instances as "Very Positive" (39.88%) and "Positive" (13.86%).

4. Instances labeled as "Positive" are generally well-predicted by the model, with an accuracy of 52.92%. However, there are misclassifications into other classes, such as "Neutral" (0.52%), "Very Negative" (0.03%), and "Very Positive" (45.54%).

5. The model also struggles with correctly classifying instances labeled as "Very Negative." It misclassifies a considerable portion of "Very Negative" instances as "Very Positive" (39.19%) and also makes misclassifications into other classes.

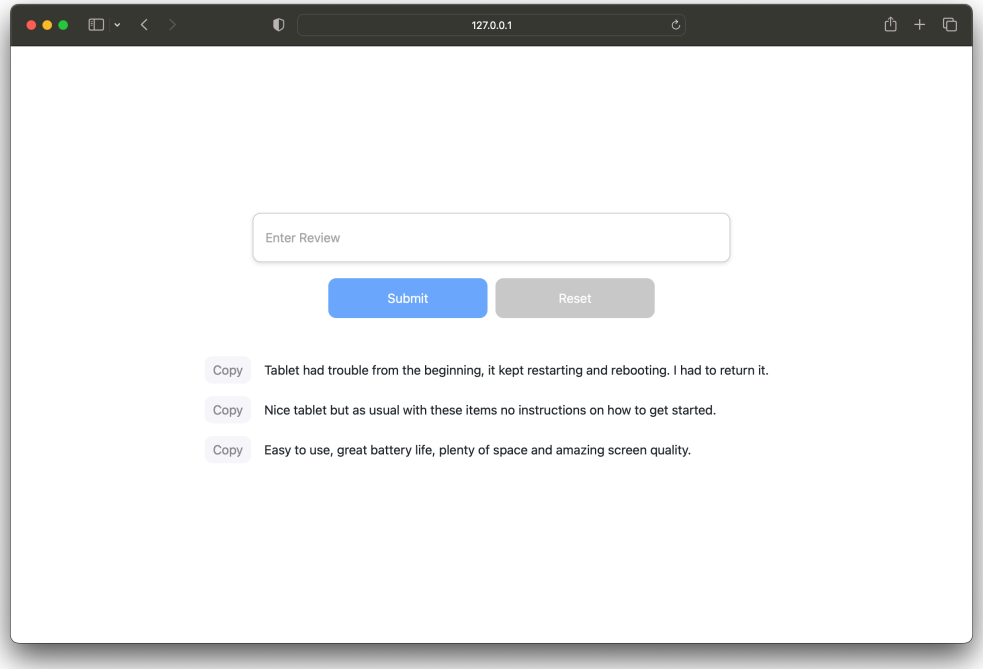
Overall, the confusion matrix reveals the strengths and weaknesses of the model across different sentiment categories. It highlights the challenges faced in distinguishing between negative and neutral sentiments, as well as between negative and strongly positive sentiments.

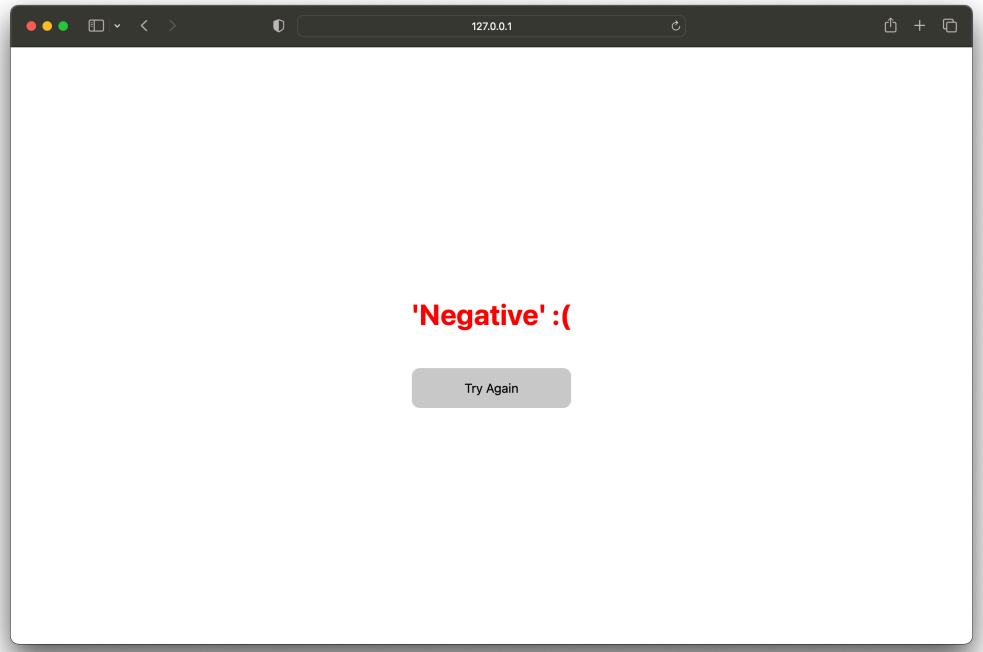
Code

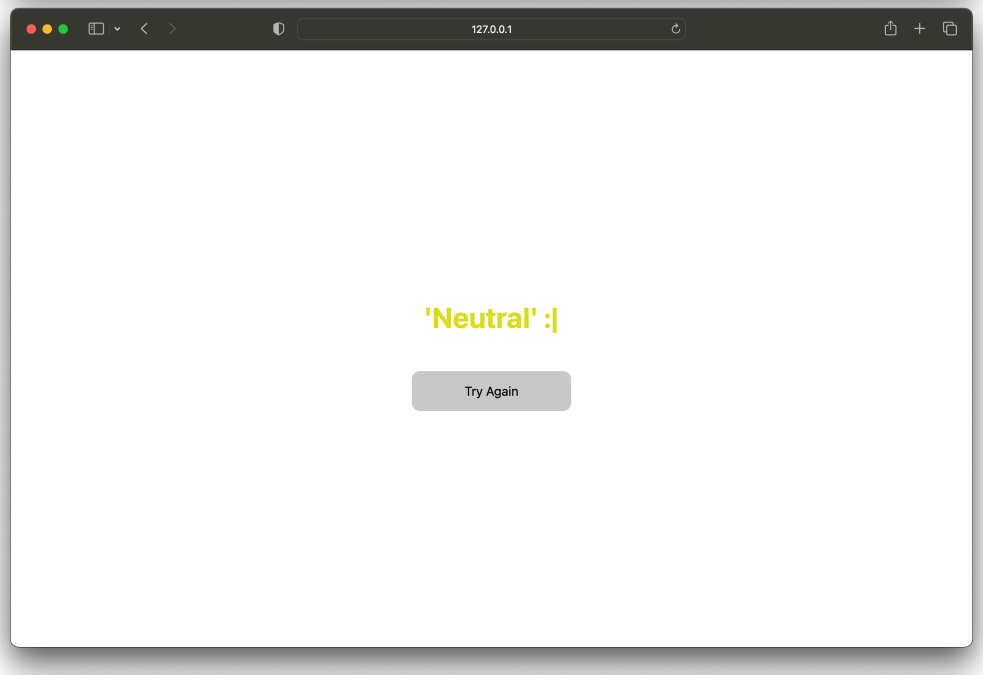
Machine Learning Models Training and Testing : Click Here for Code

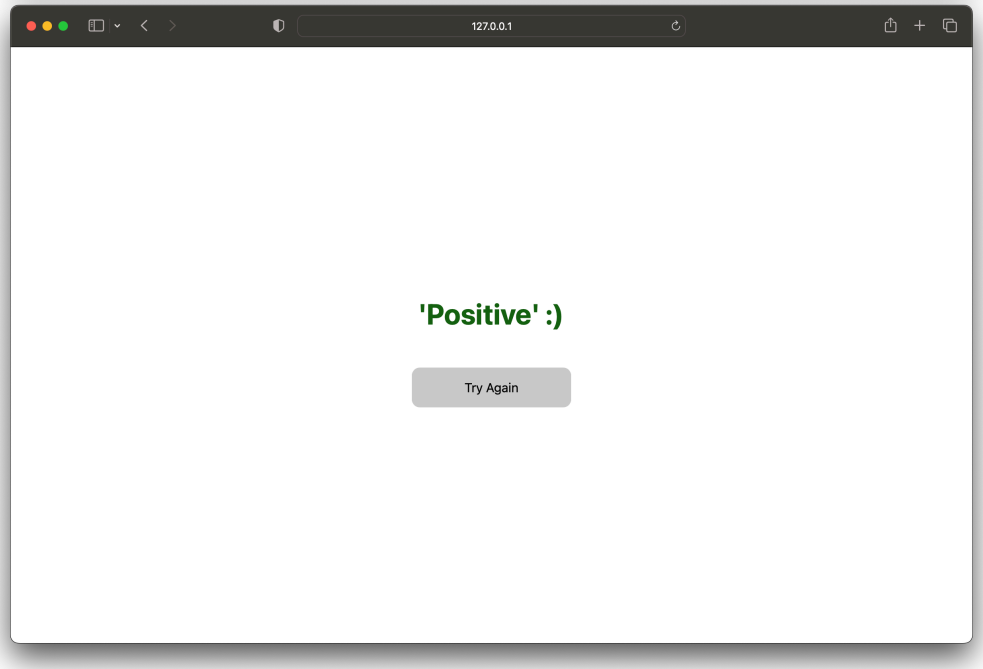
WebApp Demo Video : Click Here to View Demo Video

WebApp Deployment




127.0.0.1
'Negative' :(
Try Again


127.0.0.1
'Neutral' :|
Try Again


127.0.0.1
'Positive' :)
Try Again

CHAPTER 5

CONCLUSION & FUTURE WORK

5.1 Conclusion

In conclusion, the project successfully employed sentiment analysis on Amazon product reviews using five different machine learning models. After comparing the accuracy of each model, it was found that the SVM model performed the best, and you were able to further increase its accuracy through fine tuning.

Additionally, you developed a website that utilises this model to classify reviews as positive, negative, or neutral. This website provides a user-friendly interface that can aid individuals in making more informed purchasing decisions based on the sentiment of reviews.

Overall, your project showcases the effectiveness of machine learning in sentiment analysis and demonstrates its potential application in the field of e-commerce.

5.2 Future Work

In addition to the exciting possibility of building a website that performs comprehensive sentiment analysis on Amazon product reviews, there is a tremendous potential for expansion and improvement by incorporating cutting-edge techniques in natural language processing and deep learning.

To enhance the sentiment analysis capabilities of the website, future iterations could experiment with state-of-the-art deep learning models such as transformer models, such as BERT (Bidirectional Encoder Representations from Transformers) or GPT (Generative Pre-trained Transformer). These models have gained significant popularity and have demonstrated remarkable performance in various language processing tasks, including sentiment analysis, due to their ability to capture complex contextual information.

By leveraging these advanced models, the sentiment analysis process could be further refined to identify more nuanced sentiments, detect sarcasm, handle negations, and better understand

the overall tone and sentiment expressed in the reviews. These improvements would provide users with more accurate and comprehensive insights into the sentiment surrounding a product, enabling them to make more informed purchasing decisions.

In addition to sentiment analysis, the website could explore other text analysis techniques, such as aspect-based sentiment analysis, which aims to identify sentiments expressed towards specific aspects or features of a product. This approach could provide users with a deeper understanding of the strengths and weaknesses of a product, allowing them to assess its suitability based on their specific requirements.

As the project evolves, the website could also incorporate advanced web scraping techniques to extract additional information from Amazon product pages, such as product specifications, pricing trends, and customer ratings over time. This broader range of data would further enrich the analysis and provide users with a more comprehensive overview of a product's performance and popularity.

By combining web scraping, state-of-the-art deep learning models, and advanced sentiment analysis techniques, the website could become a powerful tool for consumers, assisting them in making informed decisions based on comprehensive insights and sentiment analysis. This integration of cutting-edge technologies would enhance the website's functionality, making it a valuable resource for users seeking trustworthy and comprehensive information before purchasing products. The possibilities for expansion and improvement are vast, and with ongoing advancements in artificial intelligence and natural language processing, the potential for creating an even more robust and user-friendly website is immense.

REFERNECS

1. Sayyed Johar, Samara Mubeen "Sentiment Analysis on Large Scale Amazon Product Reviews", International Journal of Scientific Research in Computer Science and Engineering Vol.8, Issue 1, pp.07-15, February (2020)

2. Shikha Maurya and Vibha Pratap "Sentiment Analysis on Amazon Product Reviews", 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)

3. Yuanhang Xiao and Hongyong Leng "Sentiment analysis of Amazon product reviews based on NLP", 2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)

4. Arwa S. M. AlQahtani "Product Sentiment Analysis for Amazon Reviews", International Journal of Computer Science & Information Technology (IJCSIT) Vol. 13, Issue 3, pp. 10-26 June (2021)

5. Ms. Jyoti Budhwar and Prof. Sukhdip Singh "Sentiment Analysis based Method for Amazon Product Reviews", International Journal of Engineering Research & Technology (IJERT) Vol. 9, Issue 8, pp. 54 - 57 April (2021)

6. Pankaj, Prashant Pandey, Muskan and Nitasha Soni "Sentiment Analysis on Customer Feedback Data: Amazon Product Reviews", 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con)

7. Joy Chandra Gope, Tanjim Tabassum, Mir Md. Mabrur, Keping Yu and Mohammad Arifuzzaman "Sentiment Analysis of Amazon Product Reviews Using Machine Learning and Deep Learning Models", 2022 International Conference on Advancement in Electronic Engineering (ICAEE)

8. Mandala Vishal Rao and Sindhu C "Detection of Sarcasm on Amazon Product Reviews using Machine Learning Algorithms under Sentiment Analysis", 2021 Sixth International Conference on Wireless Communication, Signal Processing and Networking (WiSPNET)

9. Rajat, Priyanka Jaroli, Naveen Kumar and Rajesh Kumar Kushal "Sentiment Analysis of Amazon Product Reviews Data Using Machine Learning Model", 2021 6th International

Conference on Innovative Technology in Intelligent System and Industrial Application (CITISIA)

10. Ansh Gupta, Aryan Rastogi and Avita Katal "A Comparative Study of Amazon Product Reviews Using Sentiment Analysis", 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)

11. Junjie Tang and Jiaxin Guo “An Analysis Method for Online Shopping Platform Comments Based on NLP-AHP: Taking Amazon as An Example”, 2022 3rd International Conference on Education, Knowledge and Information Management (ICEKIM)

12. Harsha Sinha and Arashdeep kaur “A Detailed Survey and Comparative Study of Sentiment Analysis Algorithm”, IEEE 2nd International Conference on Communication, Control and Intelligent System (CCIS)

13. Basu Dev Shivahare, Shiv Ranjan, Ashwin Murali Roa, Jeevesh Balaji, Dattattrey and Mohd. Arham “Survey paper : Study of Sentiment Analysis and Machine Translation using Natural Language Processing and its Application”, 2022 3rd international Conference of Intelligent Engineering and Management (ICIEM)

14. Lirong Yao and Yazhuo Guan “An Improved LSTM Structure for Natural Language Processing”, 2018 IEEE International Conferences of Safety Produce Informatization (IICSPI)

15. Kia Jiang and Xi Lu “Natural Language Processing and Its Applications in Machine Translation: A Diachronic Review”, 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI)

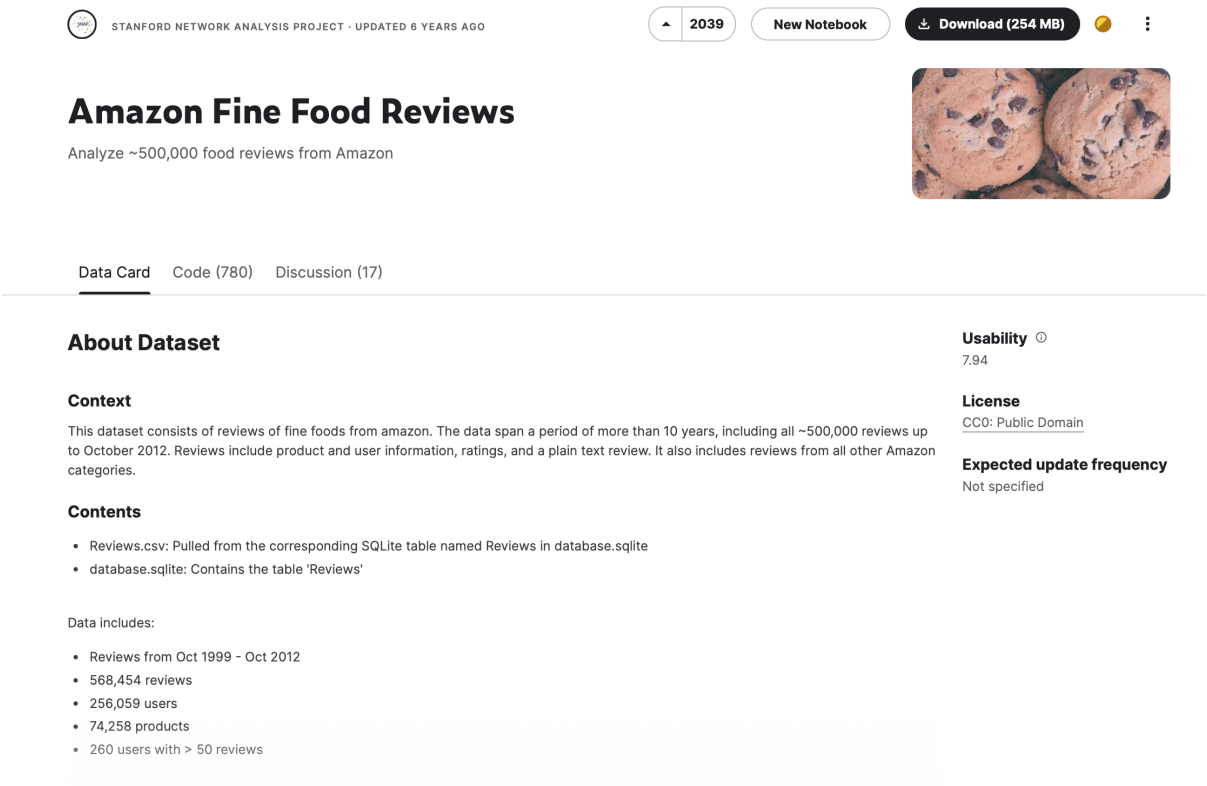
16. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa “Natural Language Processing (Almost) from Scratch”, Journal of Machine Learning Research, Vol.12, Issue 1, pp.2493–2537, August (2020)

17. K. Dave, S. Lawrence, and D. M. Pennock “Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews” 12th international conference on World Wide Web, pages 519–528. ACM, 2003.

18. S. Hota and S. Pathak “Knn Classifier Based Approach for Multi-Class Sentiment Analysis of Twitter Data”, International Journal of Engineering Technology, pages 1372–1375. SPC, 2018.

19. R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts "Recursive Deep Models for Semantic Compositionally over a Sentiment Treebank" 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642.

APPENDIX A



The Amazon Fine Foods Reviews dataset, available on Kaggle, is a comprehensive collection of customer reviews for various food products sold on Amazon. This dataset was made available by SNAP (Stanford Network Analysis Project) and contains detailed information that can be used for sentiment analysis, text mining, and recommendation systems.

The dataset consists of approximately 568,454 reviews collected between October 1999 and October 2012. Each review entry includes a variety of attributes, providing valuable insights into the customer's perspective. Some of the key features include:

1. Review Text: The actual text of the customer's review, which provides the main source of data for sentiment analysis and text mining tasks.

2. Summary: A concise summary of the customer's review, which can provide a quick overview of the sentiment or main focus of the review.

3. Score: The rating given by the customer on a scale of 1 to 5, where 1 represents the lowest rating and 5 represents the highest rating.

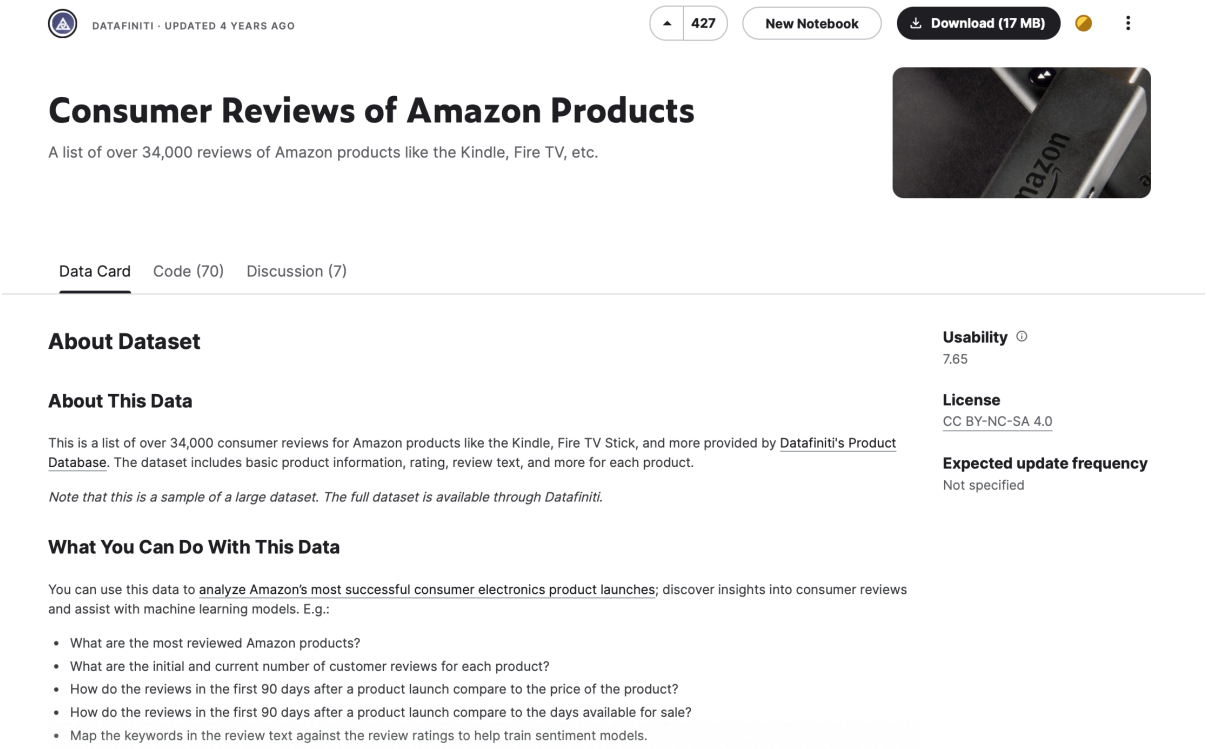
4. Helpful Votes: The number of users who found the review helpful, allowing for the identification of reviews that are considered more useful or influential.

5. Total Votes: The total number of votes received by the review, indicating the overall engagement and popularity of the review.

6. Time: The timestamp of when the review was posted, enabling temporal analysis and trend identification.

7. Product Information: Details about the specific food product being reviewed, including its name, brand, and category.

The dataset offers a wide range of opportunities for analysis and modelling. Researchers and data scientists can utilise this dataset to explore various aspects of customer sentiment, evaluate the effectiveness of different natural language processing techniques, build recommendation systems, or study trends and patterns in food-related reviews.



The "Consumer Reviews of Amazon Products" dataset available on Kaggle, provided by Datafiniti, is a comprehensive collection of consumer reviews related to various Amazon products. This dataset is designed to offer valuable insights into customer opinions and experiences with a wide range of products available on the Amazon platform.

The dataset consists of structured data, including information such as product titles, brand names, categories, prices, review dates, review ratings, and the textual content of the reviews themselves. It covers a diverse range of product categories, including electronics, books, movies, home appliances, clothing, and more. With over several thousand reviews, this dataset provides a significant volume of data for analysis and exploration.

Each review includes several attributes that can be leveraged for analysis purposes. The dataset captures the overall rating given by the customer, ranging from one to five stars, indicating their satisfaction with the product. Additionally, the dataset also provides the review text, allowing for sentiment analysis or natural language processing tasks.

Analysing the dataset can provide valuable insights into customer preferences, sentiment trends, and product performance across different categories. Researchers, data scientists, and analysts can utilise this dataset to understand consumer behavior, identify popular products, discover patterns in customer reviews, and develop models for sentiment analysis, recommendation systems, or product improvement strategies.

It is important to note that this dataset represents consumer reviews gathered from Amazon and may reflect individual opinions and experiences, which can vary in subjectivity and bias. As with any dataset, it is crucial to perform appropriate preprocessing, data cleaning, and validation before conducting any analysis or modelling tasks.

Overall, the "Consumer Reviews of Amazon Products" dataset on Kaggle provides a rich collection of customer reviews from Amazon, offering a valuable resource for studying consumer behavior, sentiment analysis, and product insights.