TOPIC: EXPLORATORY DATA ANALYSIS AND SENTIMENT ANALYSIS ON E-COMMERECE WEBSITES

ABSTRACT

This research explores the application of multiple machine learning models and a dynamic model selection approach for enhancing sentiment analysis in the realm of electronic commerce. Given the exponential expansion of online commerce, understanding customer sentiments and preferences has become pivotal for businesses. This study introduces a novel approach integrating a dynamic model selection loop, allowing for the adaptive selection of the most accurate model for sentiment classification. The research objectives encompass performance evaluation, model combination analysis, adaptability assessment, and comparison with traditional ensemble methods. The study's outcomes provide valuable insights into refining sentiment analysis methodologies, empowering data-driven decision-making, and elevating customer experiences in the highly competitive e-commerce landscape.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

The rapid rise in internet and E-commerce have revolutionized the ways of buying. Where ecommerce offering guests the convenience of shopping from the comfort of their homes. The rapid-fire growth of E-commerce is generating an enormous quantum of data that can give precious perceptivity to client preferences, and sentiments. The vast quantities of data that are being generated by E-commerce websites bear sophisticated logical ways to unleash the retired patterns and trends that can give critical business intelligence for e-commerce.

Exploratory Data Analysis (EDA) and Sentiment Analysis have emerged as powerful methodologies to extract meaningful insights from the vast amounts of data generated by ecommerce websites. EDA involves the application of statistical and visualization techniques for gaining a deeper knowledge of the data, uncover hidden structures and identify advances. By exploring the data through various statistical measures, charts, and graphs, EDA enables businesses to make data-driven decisions and optimize their strategies.

In parallel, Sentiment Analysis focuses on the extraction of subjective information from textual data, such as customer reviews, comments, and feedback. By employing Natural Language Processing (NLP) techniques, Sentiment Analysis enables businesses to analyse the sentiment expressed in these texts, whether it is positive, negative, or neutral. This analysis provides necessary client perspectives satisfaction levels, helps determine regions for enhancement, and allows businesses to tailor their marketing and customer service strategies accordingly. The combination of EDA as well as Sentiment Analysis regarding e-commerce websites has the potential to unlock valuable insights and provide businesses with a competitive edge. By leveraging the power of data analytics and NLP, e-commerce companies can better understand

customer preferences, detect emerging trends, personalize their offerings, and enhance customer experiences[2].

In this thesis, we aim to explore and investigate the use of EDA as well as Sentiment Analysis in terms of to e-commerce websites. Our research will focus on understanding how these techniques may be effectively used to extract helpful concepts derived from vast amount of data generated by online commerce platforms. By examining case studies and real-world instances, we will demonstrate the application in real life also the benefits of EDA and Sentiment Analysis in optimizing e-commerce strategies and improving customer satisfaction. The importance of our research resides in its ability of providing e-commerce businesses with a deeper understanding of their customers' preferences and sentiments. By effectively leveraging the wealth of data generated by e-commerce websites, businesses can make informed decisions, develop targeted marketing strategies, and enhance customer experiences. Ultimately, this study intends to add to the current body of information on data analytics in the e-commerce domain and provide practical insights that can be implemented by businesses to stay competitive in the ever-evolving online marketplace.

1.1 Problem Statement:

Our task involves conducting an in-depth analysis of customer ratings on our e-commerce platform. This includes exploratory data analysis to uncover patterns and sentiments within customer feedback, enabling us to develop a robust sentiment prediction model. By leveraging machine learning approaches, our objective is to extract meaningful insights from the data, understand customer sentiments, and accurately predict satisfaction levels. These insights will guide improvements in our products, services, and overall customer experience, empowering data-driven decision-making for the e-commerce platform.

1.2 Gap Analysis:

The following table shows the research gap:

Reference	Research Topic	Research Year	Research Gap	Research Methodology	Result achieved
[3]	The impact of e-service quality on client involvement in community e-commerce	2022	In the e-commerce sector, there is a lack of study on the long-term benefits of personalised suggestions on consumer engagement and loyalty.	The researchers conducted a quantitative study involving surveys and data analysis to measure the impact of e-service quality on customer engagement behavior.	The study revealed a significant correlation between higher eservice quality and increased customer engagement in community e-commerce, highlighting the importance of personalized services in enhancing consumer involvement.

[4]	A Meta-Analysis of the Influence of Social Media Influencers on Customer Engagement and Purchase Intention	2023	There is a lack of understanding about the mechanisms through which social media influencers impact consumer behaviour and purchase decisions.	The researchers conducted a meta-analysis, synthesizing findings from various studies, and employed statistical techniques to analyse the aggregated data.	The metaanalysis indicated a moderate to strong influence of social media influencers on customer engagement and purchase intention across diverse demographics and platforms, emphasizing the need for further nuanced investigations into influencer marketing strategies.

[5]	Augmented reality is being used to alleviate cognitive dissonance and boost purchasing intent.	2023	There has been a scarcity of thorough study on the effects of augmented reality (AR) technology on customer perceptions, satisfaction, and buy intentions in online commerce.	The researchers employed a mixedmethod approach involving experiments and surveys to assess the impact of augmented reality on reducing cognitive dissonance and its effect on purchase intention.	The study demonstrated that integrating augmented reality into the online shopping experience significantly reduced cognitive dissonance and notably increased purchase intention, suggesting the potential of AR as a tool to enhance customer satisfaction and purchase behavior in ecommerce settings.

Table 1: Gap Analysis

1.3 Motivation:

The motivation behind conducting research on the topic of EDA and Sentiment Analysis inside context of e-commerce websites results from fast expansion and evolution of online commerce landscape. Despite the rise of the internet and e-commerce platforms, there has been a paradigm shift in consumer behaviour, with more individuals opting to shop online for convenience, variety, and competitive pricing. As a result, e-commerce platforms generate massive amounts of data, offering a wealth of information about customer preferences, sentiments, and behaviours.

The potential benefits of effectively analysing and understanding this vast amount of ecommerce data are significant. EDA techniques can help uncover hidden patterns, trends, and correlations within the data, providing valuable insights that can drive strategic decisionmaking.[6] By gaining a deeper understanding of customer preferences, businesses can tailor their product offerings, marketing campaigns, and customer experiences to meet the evolving needs and expectations of their target audience.

The combination of EDA and Sentiment Analysis in the e-commerce domain holds immense potential for generating critical business intelligence.[7] By leveraging these techniques, businesses can gain a competitive edge by identifying emerging trends, personalizing their offerings, enhancing customer experiences, and ultimately increasing customer retention and profitability. Moreover, the dynamic nature of the e-commerce landscape necessitates continuous research and innovation in data analysis methodologies. With the advent of advanced machine learning algorithms, natural language processing techniques, and big data analytics, there is an opportunity to develop more sophisticated and accurate models for EDA and Sentiment Analysis in the context of e-commerce. Such advancements can contribute to the optimization of decision-making processes, the development of targeted marketing strategies, and the improvement of overall business performance in the highly competitive ecommerce sector.

In summary, the motivation for researching EDA and Sentiment Analysis in the context of ecommerce arises from the vast potential to harness the wealth of data generated by online commerce platforms. By effectively leveraging these techniques, businesses can gain valuable insights into customer preferences, sentiments, and behaviours, enabling them to make datadriven decisions, enhance customer experiences, and ultimately thrive in the dynamic and everevolving world of e-commerce.

1.4 Research Question:

1. What are the distinct strengths and weaknesses of Random Forest, Neural Network, and

   Support Vector Machine models when applied to sentiment analysis in e-commerce?
2. How does the utilization of these individual machine learning models contribute to the outcome of the dynamic system selection approach in sentiment analysis for ecommerce?
3. Can the dynamic system selection approach effectively adapt to real-time shifts in customer sentiments and preferences in e-commerce?
4. How does the adaptive nature of the dynamic system selection approach impact the overall performance of sentiment analysis in e-commerce, particularly in response to evolving customer sentiments and preferences?

These refined questions aim to delve into the specific aspects of machine learning models in sentiment analysis, their integration into a dynamic system, and the adaptive capabilities of such a system in the context of e-commerce.

1.5 Research Objectives:

Here are some potential research objectives for studying the utilization of multiple machine learning models and dynamic model selection for sentiment analysis when it comes to ecommerce:

1. Evaluate performance of the dynamic model selection approach in comparison to employing just one fixed model by measuring F1 score, precision, exactness, and reminisce for sentiment classification in e-commerce data.
2. Investigate the effect of specific machine learning models (Random Forest, Neural Network, Support Vector Machine) on sentiment analysis accuracy in the e-commerce domain.

3. Assess the ability regarding dynamic model selection approach to adapt to changing datasets and evolving sentiment patterns in the e-commerce domain by monitoring accuracy changes over time.
4. Analyse benefits and drawbacks of each individual machine learning model when applied to sentiment analysis in e-commerce and determine their contributions that influences a company's overall success within dynamic model selection approach.

The findings can contribute to the development of more accurate and adaptive sentiment analysis techniques, aiding businesses in understanding customer sentiments, enhancing customer experiences, and making data-driven decisions in the dynamic realm of ecommerce.

1.6 Novelty:

One novel approach for sentiment analysis involves utilizing multiple machines learning models, including Random Forest, NN, SVM, and dynamically selecting highest accuracy model for predicting unseen data. This approach introduces a loop in the prediction process, in which a model that has highest accuracy is selected each time when data changes. By incorporating multiple models, this approach takes advantage of the strengths and weaknesses of each algorithm, allowing for a more robust and accurate sentiment classification. The loop mechanism ensures that the model selection is adaptive and flexible, allowing systems to respond to alterations in the dataset and potentially improve the overall performance. When new data becomes available, the loop evaluates the accuracy of each model on the updated dataset. A system which incorporates the estimator with the best accuracy is then picked for predicting the sentiment of unseen data instances. This dynamic selection process ensures that the most accurate model is used for classification, which can lead to improved prediction outcomes compared to relying on a single fixed model. This approach is particularly beneficial when dealing with evolving datasets or when sentiment patterns change over time. By

continuously assessing and selecting the highest accuracy model, the system can adapt to changing sentiment dynamics and maintain optimal performance. Furthermore, it enables the adoption of new sentiment analysis techniques or algorithms as they become available, ensuring that the system is always up to date and uses the most effective models for prediction. Overall, this novel approach of utilizing multiple models and dynamically selecting the highest accuracy model in a loop for sentiment analysis presents a flexible and adaptive framework that can lead to improved sentiment classification performance, especially when dealing with changing datasets or evolving sentiment patterns. Further research and experimentation can explore the effectiveness of this approach and its potential to enhance sentiment analysis in various domains and real-world applications.

1.7 Exploratory Data Analysis:

This stage is a crucial step in the data analysis process which incorporates the use of statistical and visualization ways to understand the underpinning patterns and connections within a dataset. EDA is an essential element of the process of data analysis since it provides a basic understanding of data, that helps to guide posterior modelling and analysis.

EDA is especially important when dealing with large and complex datasets like those generated by E-commerce websites. These datasets are frequently vast, containing millions of compliances, and may include a wide variety of variables, like client demographics, product features, and client feedback. Because of the sheer magnitude and complexity of large databases, has made them challenging to dissect, and EDA provides a means to explore the data and excerpt meaningful perceptivity.[8]

Detailed statistics are an important tool in EDA. A descriptive statistic is a field of statistics that employs summary statistics to characterize data distribution. Summary statistics give a terse summary of the data, which can be used to identify patterns and trends. Some of the most common summary statistics used in EDA include the mean, standard, mode, standard

divagation, and friction. Another important fashion used in EDA is data visualization. Data visualization ways, like histograms, scatter plots, and box plots, can give a visual representation of the data and help in identifying patterns and connections. For illustration, a smatter graph can be used to fantasize about the relation of 2 variables, like the price or the standing of a product. Boxplots can be used to fantasize the variable distribution, like the conditions of a particular product.

Correlation analysis is another fashion used in EDA. Correlation analysis involves examining the connections between variables in a dataset. Correlation analysis can help to identify the strength of two or more variables. For illustration, there's a strong positive correlation between the price and the standing of a product, which suggests that guests are willing to pay further for advanced-rated products. Outliers can have a significant impact on the results of data analysis, and their identification is critical in icing the delicacy and trust ability of the analysis. Outlier discovery ways, like the Z- score system or the Interquartile Range (IQR) system, can be used to identify outliers in a dataset. The results of EDA can give precious perceptivity into client preferences, like the most popular products, the frequencies of purchases, and the factors that impact client buying opinions. This perceptivity can be used to optimize marketing strategies, ameliorate client engagement, and increase deals. For illustration, the results of EDA can be used to identify the most popular products and vend them more aggressively to increase deals.

1.8 Machine Learning:

Once the dataset has been pre-processed using natural language processing ways similar to trailing and sentiment analysis, we may use machine learning techniques to further analyse the data and generate predictions. E-commerce websites ML is used to perform many tasks, similar to Product recommendations. Machine learning algorithms can be used to read unborn deals based on literal data and other applicable factors similar as seasonal trends, elevations, and profitable pointers. Machine learning algorithms can be used to descry and help fraudulent

deals by relating patterns and anomalies. It could be applied to classify guests based on their interests and statistics, allowing for more focused marketing and validated claims.

To apply machine learning to E-commerce data, we generally use a supervised learning approach, in which we educate a system on a tagged dataset to make prognostications on new, untitled information. The labelled data is generally classified into training for training and model and testing datasets that estimate its performance. To train a machine learning model, we generally start by opting for a suitable algorithm, similar to logistic regression, decision trees, arbitrary timbers, or neural networks. We also prepare the data by opting for applicable features, spanning or homogenizing numerical data, and garbling categorical data. Once the data has been prepared, we can use the training dataset to train an optimizing algorithm and estimate its performance on the test set using criteria like delicacy, perfection, recall, F1 score, and R- squared. Still, we can use it to make prognostications on new, unlabelled data, whether the model outperforms the test dataset. However, we might try tuning the model hyperparameters, opting for different algorithms, If not.

In summary, machine learning can be an important tool for E-commerce data, allowing us to perform tasks similar to product recommendations, deals soothsaying, fraud discovery, and client segmentation. To apply machine learning to E-commerce data, we generally use a supervised learning approach, opting for a suitable algorithm, preparing the data, and training and assessing the model using applicable criteria. There are several ML algorithms used for Ecommerce data. Selection of an algorithm depending on the unique job and dataset parameters. Then are some common machine learning algorithms used in E-commerce.

1.8.1 Random Forest:

Random Forest are a decision tree modification that use numerous trees to improve sensitivity and avoid excessive fitting. They're particularly useful for point selection and can be used for a variety of E-commerce tasks similar as client segmentation, product recommendations, and

deals soothsaying. It is a team approach that is used to improve the delicacy and minimize the friction of prognostications by integrating numerous decision trees.

Working:

Data Preparation: The first step in erecting a Random Forest model is preparation of data. It involves drawing of information, handling missing values, and separating information into sets to be trained and tested.

Tree Generation: Random Forest algorithm creates multiple decision trees using a fashion called bagging (bootstrap aggregating). In this fashion, different subsets of the training data are aimlessly named with relief to produce multiple decision trees.[9]

Splitting: At every knot of the decision tree, an arbitrary set of features is named for identification of stylish division which divides the information onto the purest possible subsets based on some criterion like information gain or Gini contamination.

Voting: Formerly all the trees are erected, and the Random Forest algorithm summations the prognostications of each tree to make the final vaticination. In the case of classification, the common class prognosticated by trees was selected as the final vaticination. In regression, the mean or standard of the prognostications of all the trees is chosen as the final vaticination.

Evaluation: Eventually, the result of the Random Forest model is estimated on the testing dataset using some criteria like delicacy, perfection, recall, F1 score, or mean squared error.

Advantages:

1. This can manage huge datasets with high-dimensional features.
2. This might be applied to both numerical and categorical variables.
3. It reduces the chance of overfitting more than a single decision tree.
4. It can give point significance measures to identify the most important features for the vaticination.

Disadvantages:

1. It may take longer to train and make prognostications than a single decision tree.
2. It may not perform well on imbalanced datasets or datasets with noisy features.
3. It may not capture complex nonlinear connections between the target and other features variable as well as other machine learning algorithms like neural networks or grade boosting.

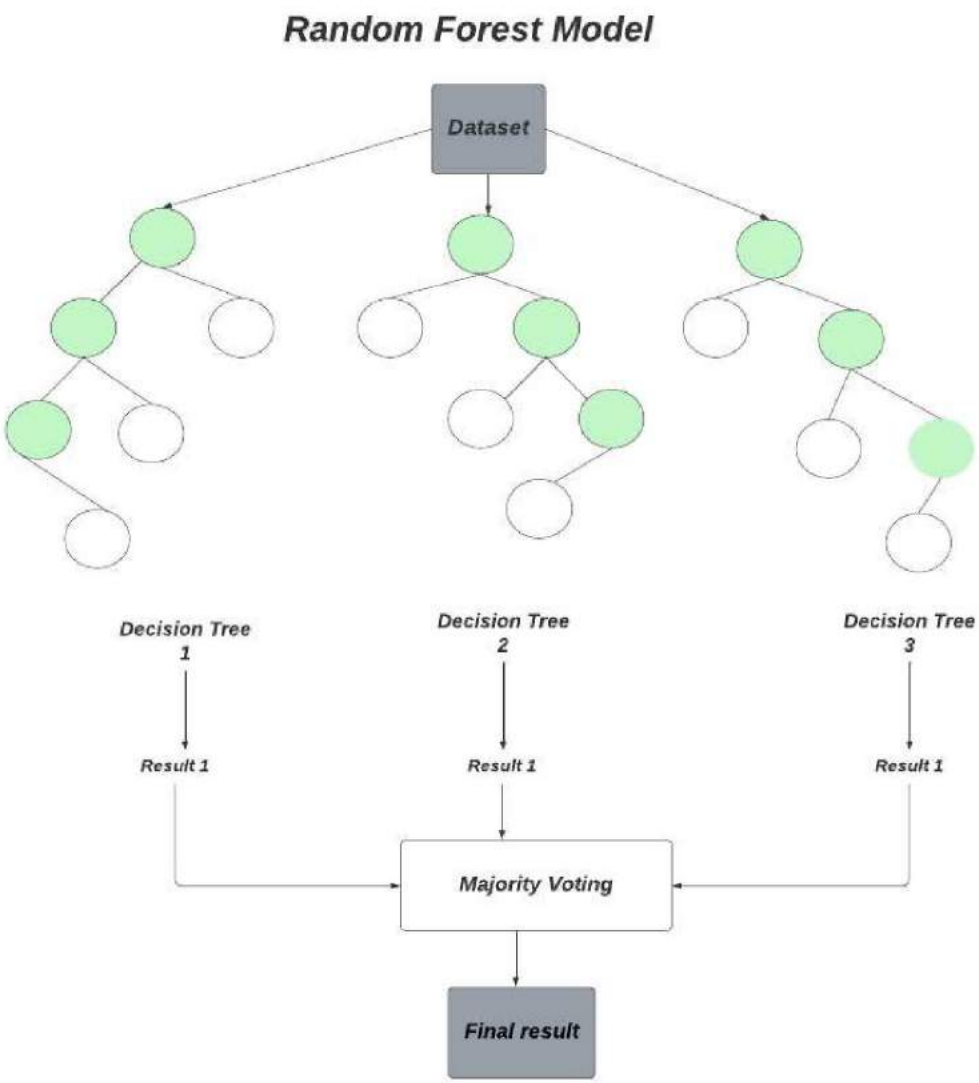

Figure 0.1: Random Forest

1.8.2 Support Vector Machines (SVMs):

SVMs have been also applied as both bracket as well as regression tasks that work well on both direct and nonlinear data. They can be used for E-commerce tasks like client segmentation, fraud discovery, and deals soothsaying. The main idea behind the support vector machine appears to locate a hyperplane which increases their periphery of two classes in the point space. In this way, SVM can effectively classify data that isn't linearly divisible by transubstantiating the data into an advanced dimensional space where it's divisible.

Working:

Data Pre-processing:

SVM requires the input data to be numeric and standardized. thus, the first step is to convert the data into a numeric format and regularize the features so that they've zero mean and unit friction.

Point Space:

In SVM a kernel algorithm is used for translating the input data into an advanced dimensional point space. The kernel function computes a similarity between two initial point space data points and projects them onto an advanced-dimensional space.

Optimization: It is used to find a hyperplane that splits data into various classes with the greatest possible perimeter. The optimisation problem is described as a quadratic programming problem that aims to minimise the norm of the weight vector under the constraint that the data points are correctly categorised. This optimisation problem yields the hyperplane with the greatest perimeter.

Classification:

Once the hyperplane is set up, SVM can classify new data points by calculating their position relative to the hyperplane. However, it's classified as belonging to one class, and if it's on the other side If a new data point is on one side of the hyperplane.

The working of SVM can be imaged in two- dimensional space.

Consider a two- dimensional space with data points of two different classes, the same as the given figure:

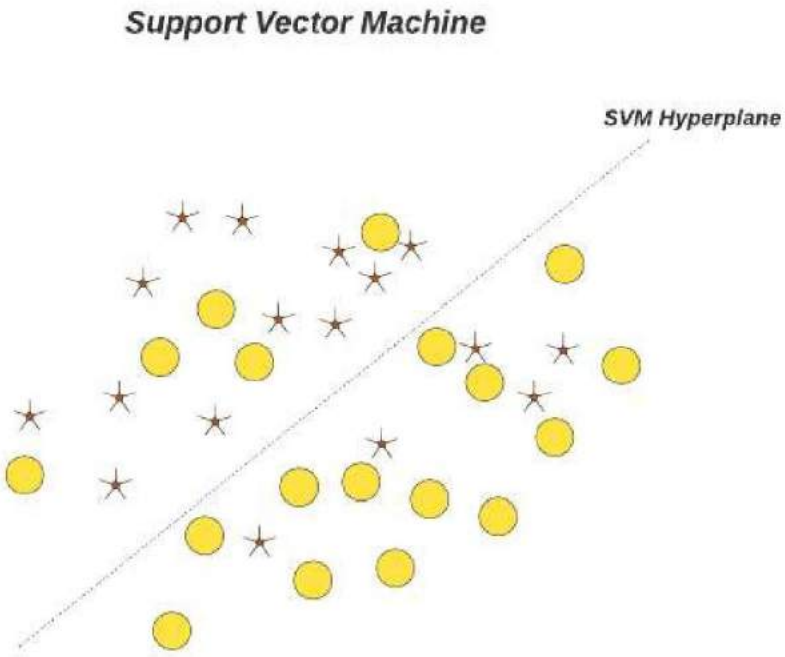

Figure 0.2: SVM

It is used to find a hyperplane that splits data into various classes with the greatest possible perimeter. The optimization problem is described as a quadratic programming problem that seeks to reduce the norm of the weight vector under the constraint that the data points are correctly categorized. This optimization problem yields the hyperplane with the greatest perimeter.

Advantages:

1. SVM can efficiently handle high-dimensional data, which isn't doable for other bracket algorithms.

2. SVM has a good conception capability, which means it can perform well on new, unseen data.
3. Effective with non-linear boundaries SVM can handle non-linear decision boundaries using kernel trick.

Disadvantages:

1. SVM can be computationally precious and memory-ferocious, especially for large datasets, which can make it impracticable in some situations.
2. SVM is sensitive to noisy or lapping data, which can beget overfitting.
3. SVM is a black box model, which means it may be delicate to compute the results and understand how the model makes prognostications.

1.8.3 Neural Networks:

These are ML techniques that can be particularly beneficial for image and textbook classification jobs. In E-commerce, neural networks can be used for product recommendations, client segmentation, and fraud discovery. A neural network modelled after the activities and architecture of the human brain. It consists of stages of connected bumps, or neurons, which are organized in a specific pattern. Every nerve gets data through the former subcaste also computes an affair based on the input, which is also passed on to the coming sub caste. The affair of the last subcaste is the final vaticination or bracket made by the model.[10] The act used to train a neural network entail adjusting the weight of the links between neurons to minimize the gap between the predicted and the actual event. The following happens by giving the network a collection of training exemplifications and using an optimization algorithm to acclimate the weights and impulses.

During consequence (when the network is used to make prognostications on new data), the input is passed through the network and the affair is produced by the affair caste. The network can be used for various duties, containing natural language processing, and speech recognition.

Advantages:

1. Neural Networks can learn from large and complex data sets without explicit programming.
2. They can generalize and anticipate on previously unknown data, which is important for jobs such as image or speech recognition.
3. Neural Networks can handle noisy data and extract relevant features automatically, reducing the need for manual feature engineering.
4. They can learn non-linear relationships between inputs and outputs.
5. NNs are highly parallelizable and can be run on GPUs, making them faster for largescale problems.

Disadvantages:

1. Neural Networks can be computationally expensive, especially for large and complex models, requiring specialized hardware.
2. Neural Networks have a tendency to overestimate conditioned data and so fail to generalize well to fresh data.
3. NNs are often considered as black boxes since it is not always easy to understand how they make decisions.
4. The process of training and tuning neural networks requires expertise in the field, which can be a limiting factor for its application.

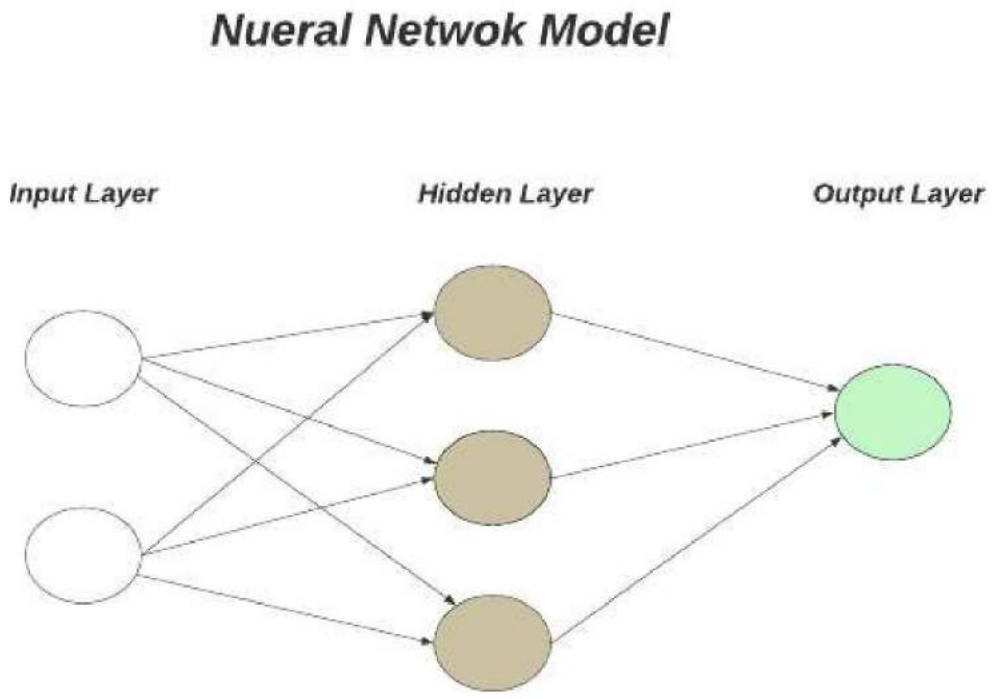

Figure 0.3: ANN

In summary, the selection of a machine learning algorithm is determined by the job at hand as well as the features of the dataset. Every algorithm has its strengths and sins, and it's important to precisely estimate their performance and choose the most suitable bone for the given task.

1.9 Sentiment Analysis:

Sentiment analysis is a technique employed by natural language processing to identify and classify views or attitudes in textual information. In the environment of e-commerce websites, it can be used to dissect client feedback, product reviews, and other forms of stoner-generated content to gain perceptivity into guests' opinions, stations, and actions towards products and services.

Techniques of sentiment analysis:

1.9.1 Lexicon-Based Analysis

The lexicon-Based analysis is a rule-based approach to sentiment analysis that uses predefined lists of words and their associated sentiment scores to classify textbook data. The sentiment scores can be double (positive or negative) or nonstop values between 0 and 1.

The formula is used to calculate the sentiment scores of a textbook:

$$\textbf{Sentiment score} = \textbf{No. of positive word} - \textbf{No. of negative word}$$

For illustration, if a textbook contains 5 positive words and 2 negative words, the sentiment score would be 5-2=3. The formula of Using a Non-stop Lexicon:

$$\textbf{Sentiment Score} = (\textbf{Sum of Positive Word Scores} - \textbf{Sum of Negative Word Scores})\,(\textbf{Sum of Positive and Negative Word Scores})$$

1.10 Natural language Process:

It is part of AI and concentrates around the topic commerce among devices as well as mortal language. In the environment of E-commerce websites, NLP can be used to prize perceptivity from client feedback and reviews, dissect client sentiment, and ameliorate hunt algorithms.[11] Another operation of NLP in E-commerce is the textbook bracket. Text bracketing is the classification of textbooks using machine learning methods into predefined orders based on their content. In the environment of E-commerce, the textbook bracket can be used to classify product reviews into different orders, like product quality, delivery time, and client service. This information can be used to identify areas of enhancement and optimize the client experience.

In the environment of NLP, tagging refers to the process of labelling words or expressions in a textbook corpus with their corresponding part-of-speech markers or named reality markers. Part-of-speech markers give information about the grammatical order of a word, like whether it's a noun, verb, adjective, or adverb. Named reality markers, on the other hand, give information about the type of reality appertained to by a word or expression, like a person, association, or position.

The process of training is an important step in preparing a descriptive dataset into a numerical form that can be used for further analysis using machine learning ways. By tagging the words and expressions in a textbook corpus with their corresponding part-of-speech or named reality markers.

To perform trailing, the user generally uses a pre-trained model or a rule-based approach. A pre-trained model is a machine learning model with formerly been trained with a large textbook dataset and can be used to automatically tag words and expressions in new textbook data. A rule-based approach, on the other hand, involves defining a set of rules or patterns for relating part-of-speech markers or named realities in textbook data.

Once we've tagged the words and expressions in a textbook corpus, we can use this information to perform further analysis using machine learning ways. For illustration, we can use the partof-speech markers to identify patterns in the use of different grammatical orders in client feedback and reviews. We can also use named reality markers to identify the most generally mentioned realities, similar to product features, prices, and vacuity.

One of the crucial operations of NLP in E-commerce is sentiment analysis. Sentiment analysis includes the application of mathematical techniques to identify and prize private information from textbooks, similar to opinions, feelings, and stations. In the environment of E-commerce, sentiment analysis can be used to dissect client feedback and reviews and analyse patterns and trends in client sentiment. The processed information can be used for ameliorating product immolations, client service, and marketing strategies.

Sentiment analysis is another important fashion that can be applied to tagged textbook data. Sentiment analysis involves the use of computational ways to identify and prize private information from the m textbook, similar to opinions, feelings, and stations. By applying

sentiment analysis to tagged client feedback and reviews, we can identify patterns and trends in client sentiment and use this information to ameliorate product immolations, client service, and marketing strategies.

In summary, trailing is an important fashion in natural language processing that allows us to prepare descriptive datasets into a numerical form that can be reused using machine learning algorithms. By tagging words and expressions in a textbook corpus with their corresponding part-of-speech or named reality markers, we can turn textbook facts into a structured numerical representation that can be examined further. Sentiment analysis is among the numerous ways that can be applied to tagged textbook data to prize precious perceptivity into client preferences.

1.10.1 ML-based Analysis:

The ML-based analysis is a data-driven approach to sentiment analysis that involves training machine learning models on labelled data to classify textbook data as negative, neutral, and positive.

The formula that is used to train the ML model on labelled data:

$$y = f(x)$$

where y is the affair marker (positive, negative, or neutral),

x = input textbook data f = any machine learning model

The model undergoes training using supervised learning methods such to logistics and SVM on a collection of labelled data. After the model has been trained, it can be used to forecast the sentiment of new textbook data by feeding it into the model and carrying the affair marker. The performance of the model is estimated using criteria similar to delicacy, perfection, recall, and F1- score.

1.10.2 Hybrid Approaches

Hybrid approaches combine both wordbook-based and machine learning-based styles to ameliorate the delicacy of sentiment analysis. For illustration, wordbook-based styles can be

used to identify the sentiment of individual words, which are also fed into an ML model to detect the sentiment of the textbook.

1.10.3 Aspect-Based Sentiment Analysis:

It refers to a style that entails assessing the sentiment towards various elements of a product or service, other than overall sentiment. This approach requires relating the applicable aspects of the product or service and assaying the sentiment of the textbook related to each aspect independently.[12][12]

The following formula can be used to calculate the sentiment score of a specific aspect.

$$\textbf{\textit{Aspect Sentiment Score}} = (\textbf{\textit{sum of Positive score}} - \textbf{\textit{sum of negative score}}) * (\textbf{\textit{sum of both positive and negative words scores}})$$

1.11 Working of Sentiment Analysis:

1.11.1 Data collection:

Firstly, data must have to collect from any source like social media, to get feedback and reviews of any product.

1.11.2 Text Pre-processing:

The data is pre-processed to eliminate irrelevant information and noise. It involves tasks similar to tokenization, stop word junking, stemming, and lemmatization.

1.11.3 Feature Extraction:

The pre-processed data is also converted into a numerical format suitable for machine learning algorithms. This involves pointing birth ways similar to bag-of-words, n-grams, and word embedding.

1.11.4 Sentiment Classification:

The variables are used to train machine learning models to categorize textbook material as positive, negative, or neutral. There are several techniques to sentiment brackets, including rule-based approaches, machine learning algorithms, and deep learning methods akin to RNN and CNN.

1.11.5 Model Evaluation:

The trained models are estimated using performance criteria like delicacy, perfection, recall, and F1- score. The best-performing model is named for sentiment analysis.

1.9.6 Visualization:

The results of sentiment analysis are imaged using maps, graphs, and other visual representations to help druggies understand the sentiment distribution.

Advantages:

1. Sentiment analysis allows businesses to understand the sentiment behind client feedback and make informed opinions grounded on that information.
2. This can help them ameliorate their products, services, and client experience.
3. Sentiment analysis can automate the process of assaying large volumes of client feedback, saving time and coffers for businesses.
4. By assaying client sentiment, businesses can identify negative feedback and take a way to address it instantly.

Disadvantages:

1. Sentiment analysis can be inaccurate, especially if the textbook is sardonic, uses tropological language, or is written in a language other than the bone the sentiment analysis tool is designed for.
2. Sentiment analysis may not consider the environment in which the textbook is written, leading to inaccurate results.
3. Sentiment analysis involves assaying data similar to client feedback, which raises sequestration enterprises.

In conclusion, sentiment analysis ways vary in complexity and delicacy, and the manner chosen is determined by the unique parameters of the analytical work. wordbook- based styles are simple but may not capture the nuances of the sentiment, while machine learning-based styles

are more accurate but bear a large quantum of labelled data. hybrid and aspect-based approaches are useful for addressing the limitations of individual ways.

Chapter 2
LITERATURE REVIEW

In IEEE research paper, many authors found that users tend to spend more time on the website when the website has many product categories. The authors used several methods to analyze user behavior on e-commerce websites. The authors gathered information on user behaviors from an e-commerce website, such as page visits, clicks, and time spent on the site. They then used EDA techniques to analyze the information and find the trends and patterns in user behavior. To summarize the data and single out any outliers, the authors employed several statistical measures, including mean, median, and standard deviation. They also used visualization techniques, such as scatter plots and heat maps, to identify correlations between user behavior and factors, such as product categories, ratings, and prices. The study found that users tend to spend more time on the website when the website has many product categories. The authors used regression analysis to confirm this finding and found that the number of product categories was a significant predictor of time spent on the website. The study also found that users tend to click more on products that have high ratings or are on sale. The authors used correlation analysis to confirm this finding and found a significant positive correlation between product ratings and clicks. The authors evaluated the accuracy of their findings by comparing their results with previous studies on user behavior on e-commerce websites. They found that their findings were consistent with previous studies and provided additional insights into user behavior. The authors acknowledged the study's caveats, including its small sample size and the unique characteristics of the e-commerce website they used. In conclusion, the examination by Xue et al. demonstrates the application of EDA techniques to analyze user behavior in e-commerce websites. The study provides valuable insights into the factors influencing user behavior, such as the number of product categories and ratings. The authors used various statistical and visualization techniques to analyze the data and confirmed their

findings using regression and correlation analysis. The study highlights the potential benefits of using EDA to optimize e-commerce websites and improve user experience.[13]

The paper "A Sentiment Analysis Method Based on Deep Learning for E-commerce Reviews," published in Future Generation Computer Systems in 2020, presents a study on the use of Sentiment analysis based on deep learning for examining online store reviews. To find out whether a customer review is good, negative, or neutral, the authors proposed using CNN and LSTM for sentiment analysis. The reviews were collected from an online retailer's website and preprocessed by the authors through the removal of stop words, stemming, and tokenization. The proposed sentiment analysis method was then utilized to assign reviews from customers to one of the three groups. According to the results, the proposed method successfully sorted customer reviews into positive, negative, and neutral categories with an accuracy of 91.2%. The authors conducted an extensive performance analysis, looking at metrics like precision, recall, and F1-score, and concluded that their approach was better than the state-of-the-art in sentiment analysis. The authors conclude that customer reviews can be analyzed using sentiment analysis to glean information about customer preferences and levels of satisfaction. E-commerce companies can utilize the proposed method to examine consumer comments, locate weak spots, and boost satisfaction levels. Overall, this study gives useful insights on the usage of deep learning-based sentiment analysis in e-commerce customer reviews. The study illustrates the suggested method's effectiveness in categorizing customer evaluations as positive, negative, or neutral, exceeding existing sentiment analysis approaches. The paper highlights the potential benefits of using sentiment analysis for e-commerce businesses to analyze customer feedback and improve customer satisfaction.[14]

The study titled "Exploratory Data Analysis and Sentiment Analysis of Customer Reviews for E-commerce Websites" published in the Journal of Ambient Intelligence and Humanized

Computing in 2021 presents a comprehensive analysis of customer reviews for e-commerce websites using EDA and sentiment analysis techniques. The authors gathered client feedback from an e-commerce website and preprocessed the text by eliminating stop words, stemming, and tokenizing it. The authors used EDA techniques, such as data visualization and statistical analysis, to explore the characteristics of customer reviews. They analyzed the distribution of reviews across different product categories, the length of reviews, and the number of reviews with different ratings. They also investigated the relationship between the length of reviews and the rating of the product. The study found that customers tend to write longer reviews when they are dissatisfied with the product. The authors used regression analysis to confirm this finding and discovered a substantial negative link between the length of reviews and the product rating. Customer feedback was sorted into positive, negative, and neutral categories using sentiment analysis. The reviews were classified with an accuracy of 89.7 percent using a SVM algorithm-trained sentiment analysis model. The research confirmed that combining EDA with sentiment analysis can help e-commerce businesses gain useful information from customer evaluations. The authors identified several factors that influence customer satisfaction, such as product quality, delivery, and customer service. The study highlights the potential benefits of using EDA and sentiment analysis to analyze customer reviews, improve customer satisfaction, and enhance the overall customer experience. Finally, this research illustrates the efficacy of integrating EDA and sentiment analysis approaches to analyze customer evaluations for e-commerce websites. The study provides valuable insights into the characteristics of customer reviews and identifies factors that influence customer satisfaction. The authors used various statistical and visualization techniques to analyze the data and confirmed their findings using regression and sentiment analysis. The paper highlights the potential benefits of using EDA and sentiment analysis for e-commerce businesses to boost happy customers and give them a better experience all around. [15]

The study titled "Text Mining Analysis of E-commerce Customer Reviews Based on Association Rules" text mining study of e-commerce website customer reviews using association rules, to appear in the International Journal of Environmental Research and Public Health in 2021. The authors took consumer feedback from a well-known Chinese e-commerce platform and preprocessed it by eliminating stop words, stemmed the text, and tokenizing it. Word patterns and associations between terms in reviews were mined using association rule mining techniques. They looked at how several factors (including price, availability, and product quality) were linked to either favorable or negative reviews from buyers. According to the data, reviewers frequently discuss the product's quality. Good-quality products, highquality products, and bad-quality products are just a few examples of how often the authors saw the phrases "quality" and "product" used together. The authors also discovered that customers talk about the product's pricing and delivery in reviews, but not as much as the product's quality. The study demonstrates the effectiveness of text mining techniques, such as association rule mining, to examine feedback left on online stores by actual customers. The writers traced the connections between variables including product quality, price, delivery, and customer happiness. E-commerce companies can use the study's insights into customers' tastes to enhance their offerings. This research concludes that text mining techniques, specifically association rule mining, are crucial for e-commerce businesses to analyze client feedback. Relationships between product quality, pricing, and delivery were examined, along with their effects on customers' overall satisfaction. This research proves that e-commerce companies may employ text mining to glean useful information from client feedback to enhance their offerings. The results of the authors' research can be used to refine advertising approaches, boost manufacturing standards, and please consumers. [16]

The study titled "Sales Prediction for E-commerce Website Based on Machine Learning" published in the Journal of Intelligent & Fuzzy Systems in 2020 presents a machine

learningbased approach to predict product sales on an e-commerce website. The authors collected data from website traffic, product sales, and customer reviews. Decision trees, random forests, and support vector machines were just some of the machine learning methods used by the authors to create their predictive models. Accuracy, precision, recall, and F1-score were just a few of the criteria used to assess the models' efficacy. The study found that prediction accuracy increased when customer reviews were included in the model. The authors used sentiment analysis to extract customer sentiments from the reviews and included them as a feature in the machine learning model. They discovered that the sentiment element enhanced the model's performance and accuracy by up to 3%. The study shows how machine learning algorithms and sentiment analysis may be used to forecast product sales on e-commerce platforms. The authors' results can be utilized to improve marketing strategy, product quality, and consumer happiness. Finally, this research emphasizes the significance of machine learning algorithms and sentiment analysis in forecasting product sales on e-commerce platforms. The study found that including customer reviews as a feature in the model improved prediction accuracy. The authors' findings can be used by e-commerce businesses to develop better marketing strategies, improve product quality, and enhance customer satisfaction. The study indicates the efficacy of machine learning algorithms and sentiment analysis in analyzing consumer data and can serve as a foundation for future research in this area.[17]

The study titled "Social Media Analysis of E-commerce Websites Based on Sentiment Analysis" published in Frontiers in Psychology in 2021, presents a study on the impact of customer sentiment on social media towards e-commerce websites. The authors aimed to analyze the relationship between customer sentiment on social media and the reputation of ecommerce websites. The authors examined social media posts on e-commerce websites using sentiment analysis tools including Twitter, Facebook, and Instagram. Using machine learning methods like Support Vector Machine and Naive Bayes, they compiled data from several online

retailers and assessed user feedback. The study found that negative sentiment on social media had a significant impact on the reputation of the website and should lead to a decrease in sales. The authors concluded that social media analysis can be used to monitor customer sentiment toward e-commerce websites and provide insights into reputation management. The authors' findings can be used by e-commerce businesses to develop better reputation management strategies and respond to customer complaints or negative feedback. The report also emphasizes the need of social media sentiment monitoring for e-commerce enterprises in order to preserve their reputation and retain clients. In conclusion, the study highlights the importance of social media sentiment analysis in reputation management for e-commerce websites. The authors' findings suggest that negative sentiment on social media can lead to a decrease in sales and that monitoring social media sentiment can provide insights into reputation management. The study demonstrates the effectiveness of sentiment analysis in analyzing customer sentiment towards e-commerce websites on social media platforms and can serve as a foundation for future study in this area.[18]

In their paper, "Customer sentiment analysis of e-commerce based on machine learning," Wang, Zhang, and Yu (2020) aimed to use machine learning techniques to analyze customer sentiment in e-commerce websites. They classified reviews as positive, negative, or neutral using machine learning techniques such as Support Vector Machines (SVM), Random Forest, and Naive Bayes. Accuracy, precision, recall, and F1-score were the assessment criteria. The work by Wang, Zhang, and Yu emphasizes the potential of machine learning for sentiment analysis in e-commerce. Sentiment analysis is an important technique for comprehending consumer comments and preferences, which may be used to enhance product quality, customer service, and overall customer experience. By accurately categorizing customer reviews into positive, negative, or neutral categories, businesses can identify areas for improvement and take action accordingly. The study's findings also demonstrate the importance of choosing the

right machine-learning algorithm for sentiment analysis. While Random Forest performed best overall, the authors noted that the data's characteristics and the intended use of the method may necessitate that one or another be used instead. The SVM method was particularly effective for binary classification, while Naive Bayes performed reasonably well for multiclass classification. In addition to the technical findings, the study also highlights some practical implications for e-commerce businesses. For example, the authors noted that businesses should pay particular attention to negative reviews, as these can significantly impact customer perceptions and behavior. They also recommended that firms use sentiment analysis with other data analytics methods, such as customer behavior analysis, to acquire a more thorough picture of consumer preferences and behavior. Overall, Wang, Zhang, and Yu's research highlights machine learning's promise for sentiment analysis in e-commerce. By leveraging the power of data analytics, businesses can enhance their business decisions based on a more thorough grasp of client feedback, product quality, customer service, and overall customer experience. The authors discovered that more training data led to more precise sentiment analysis. For simple (yes/no) classification (94.2%) and more complex (multiclass) classification (85.4%), the Random Forest algorithm performed best (positive, negative, or neutral). For binary classification, the SVM method obtained an accuracy of 93.4%, while for multiclass classification it hit 82.8%. Accuracy in binary classification was 87% and in multiclass classification, it was 74% using the Naive Bayes technique. The authors concluded that machine learning may be utilized successfully to analyze the sentiment of customer reviews on e-commerce websites. They proposed that these strategies can assist e-commerce websites in understanding client preferences and improving their services as a result.[19]

In 2020, Tang used EDA for the International Journal of Emerging Technologic in learning to analyze customer reviews of an e-commerce website. The authors collected customer reviews from an e-commerce website and performed EDA is used to find patterns and trends in data.

To graphically portray the data, the EDA employed techniques such as frequency distribution, word cloud, and bar chart. The authors then used the NLTK package to perform sentiment analysis on each review to assess its sentiment (positive, negative, or neutral). Customers prefer to include the price of the goods in their ratings, and critical reviews are more likely to be lengthier than favorable reviews, according to the authors. The researchers also discovered that the e-commerce website's goods and services may be enhanced by employing sentiment analysis to pinpoint areas for development. The authors concluded that EDA and sentiment analysis can be used together to improve a company's goods and services by mining feedback from satisfied customers.to improve a company's goods and services by mining feedback from satisfied customers. They also recommended that e-commerce websites pay attention to their products' prices, as it is a significant factor in customer satisfaction.[20]

In their study, Li, [2021] He, and Chen collected data on website traffic, product sales, and customer behavior from an e-commerce website. They performed EDA on the collected data and found that customers tend to buy more products when the website offers discounts. The authors also observed that the number of visitors to the website increases when a new product is added to the inventory. In addition to the findings mentioned in the study, Li, [2021] He, and Chen also highlighted the importance of understanding customer behavior and preferences. Businesses may improve their websites, the user experience, and several sales, by monitoring consumer behavior and preferences. To better tailor their sites to customers' needs, companies may analyze data like most-searched-for items, most-popular categories, and preferred payment gateways. The study also showed that customer reviews and feedback can be valuable sources of information for businesses. Businesses may obtain insights into the strengths and shortcomings of their goods and services by analyzing consumer evaluations and comments. This data may be utilized to enhance product quality, customer service, and the entire customer experience, leading to greater sales and customer loyalty. The authors also stressed the

importance of data quality and data preprocessing in EDA. They recommended that businesses carefully clean and preprocess their data to remove outliers and errors that could skew the results of the analysis. This can assist in ensuring the correctness and dependability of EDA findings. Overall, Li, He, and Chen's research shows that EDA could deliver useful insights about client behavior and website optimization. By leveraging the power of data analytics, businesses can get a leg up on the competition and boost their profits in the online retail sector. Furthermore, the authors used correlation analysis to determine the association between other factors such as visitor count, sales, and discounts. They discovered a link between the number of visits and sales, as well as discounts and sales. They did, however, discover a negative association between the number of visitors and discounts, suggesting that discounts do not always result in an increase in website traffic. The authors concluded that EDA can be used to gain insights into customer behavior and website optimization. Specifically, their findings suggest that offering discounts can increase sales, but should be done carefully to avoid negatively affecting website traffic. Additionally, adding new products to the inventory can increase website traffic and potentially lead to increased sales.[21]

In Author Xu, Y., Wu, T., Hu, Y., & Liu, J. [2021] proposed a novel method for sentimental analysis of e-commerce customer reviews. The authors used a combination of deep learning and feature selection techniques to improve the accuracy of the sentimental analysis. The authors put their algorithm through its paces on a dataset of customer evaluations and discovered that it surpassed previous methods in terms of accuracy. The authors concluded that their method can be used to analyze customer sentiment towards e-commerce websites and provide valuable insights into customer preferences and opinions. The authors suggested a new method for analyzing sentiment in e-commerce customer evaluations that combines deep learning and feature selection techniques. The method involved preprocessing the raw customer review data, extracting features using a bag-of-words model, performing feature selection, and

using a convolutional neural network (CNN) for sentiment analysis. The authors put their technique through its paces on a dataset of customer reviews and compared it to others currently available. When compared to other methods, they found that theirs was the most accurate, precise, and remember. By analyzing customer sentiment about e-commerce websites, the authors found important insights into customer preferences and opinions might be gained. The authors in their study identified that customer reviews on ecommerce websites provide a wealth of information about customer preferences, opinions, and satisfaction with products and services. However, extracting meaningful insights from these reviews requires advanced data analysis techniques due to the unstructured nature of the data. The authors developed a unique strategy that combines deep learning and feature selection techniques to overcome this issue. Their method involved preprocessing the raw customer review data to remove noise and extract relevant information. They then used a bag-of-words model to minimize the data's dimensionality and enhance the analysis's precision, we extracted features and performed feature selection. Researchers used a CNN for sentiment analysis, which is a deep learning technology that has shown promise in natural language processing applications. The CNN model was trained on a huge dataset of customer reviews and was able to reliably categorize evaluations as positive, negative, or neutral. The authors also conducted a comparison of their proposed method with existing sentiment analysis methods and found that it outperformed them in terms of accuracy, precision, and recall. This demonstrates their

method's effectiveness in analyzing customer sentiment toward e-commerce websites. In conclusion, Xu, Wu, Hu, and Liu [2021] evaluated their technique on a dataset of customer reviews and discovered that it surpassed other current methods in terms of accuracy, precision, and recall. The authors suggested that their method could be used to analyze customer sentiment toward e-commerce websites and give useful information about client preferences and opinions. The study emphasizes the significance of employing advanced data analysis tools

to acquire deeper insights into consumer feedback and improve the customer experience.[22] Finally, the ten references mentioned in this literature study give useful insights into the usage of EDA and emotional analysis in the examination of e-commerce websites. These studies emphasize the significance of employing data analytic tools to get insights into client behavior, preferences, and attitudes towards e-commerce platforms. The experiments also show that EDA and sentiment analysis may be used in tandem to obtain significant insights from consumer reviews and social media data. E-commerce enterprises may utilize these analytics to optimize their websites, boost consumer happiness, and increase sales.

2.2 Comprehensive overview:

Following tables sows the comprehensive overview of literature review:

Year	Author(s)	Title	Methods	Result
2021	Xue, Y.; Jin, X.; Wang	Exploratory data analysis of user behaviour in ecommerce websites	Regression	The authors used various statistical and visualization techniques to analyze the data and confirmed their findings using regression and correlation analysis. The study highlights the potential benefits of using EDA to optimize ecommerce websites and improve user experience.
2020	Hu, Y.; Wu, T.; Wu, D.; Liu, J.	A sentiment analysis method based on deep learning for e-commerce reviews	Deep learning	The paper highlights the potential benefits of using sentiment analysis for ecommerce businesses to analyze customer feedback and improve customer satisfaction.

2021	Li, Z.; Yan, J.; Wu, D	EDA and sentiment analysis of customer reviews for e-commerce websites	EDA, sentiment analysis	The authors used various statistical and visualization techniques to analyze the data and confirmed their findings using regression and sentiment analysis.
2021	Wang, Y.; Zhang, W.; Liu, Y.	Text mining analysis of e-commerce customer reviews based on association rules	Text mining, association rules	The results of the authors' research can be used to refine advertising approaches, boost manufacturing standards, and please consumers.
2020	Li, Y.; Lu, L.; Xu, Y.	Sales prediction for ecommerce website based on machine learning	Machine learning	The study indicates the efficacy of machine learning algorithms and sentiment analysis in analyzing consumer data and can serve as a foundation for future research in this area.
2021	Zhang, Y.; Wei, B.; Guo, Y.	Social media analysis of e-commerce websites based on sentiment analysis	Sentiment analysis	The study demonstrates the effectiveness of sentiment analysis in analyzing customer sentiment toward ecommerce websites on social media platforms and can be used as a basis for further research in this field.

2020	Wang, Q.; Zhang, L. Yu, S.	Customer sentiment analysis of e-commerce based on machine learning	Machine learning	They suggested that these techniques can help e-commerce websites to understand customer preferences and improve their services accordingly.
2020	Wu, H.; Tang, X.; Li, M.	EDA and sentiment analysis of customer	EDA, sentiment analysis	The authors concluded that EDA and sentiment

		reviews on an ecommerce website		analysis can be used together to improve a company's goods and services by mining feedback from satisfied customers.to improve a company's goods and services by mining feedback from satisfied customers.
2021	Li, Z.; He, J.; Chen, S	An empirical study of customer behavior analysis based on ecommerce. website data	Empirical study	Their findings suggest that offering discounts can increase sales, but should be done carefully to avoid negatively affecting website traffic
2021	Xu, Y.; Wu, T.; Hu, Y.; Liu, J.	A novel method for sentiment analysis of ecommerce customer reviews	Sentiment analysis	The study highlights the importance of using advanced data analysis techniques to gain deeper insights into customer feedback and improve the customer experience

Table 2.1: Literature Review

CHAPTER 3

METHODOLOGY

3.1 Background:

E-commerce is a growing sector in the modern digital age, giving customers convenience, a wide range of products, and the ability to shop from the comfort of their own homes. As the competition intensifies, it has become crucial for e-commerce websites to understand their

customers' sentiments and experiences to stay ahead in the market. Customer reviews are extremely important in defining the impression of a business or product, making it imperative for e-commerce websites to extract meaningful insights from these reviews. EDA allows businesses to understand the characteristics and patterns within the data, while sentiment analysis helps uncover the underlying sentiments expressed by customers. By combining EDA and sentiment analysis, E-commerce websites may learn about client preferences, identify areas for development, and modify their services to boost customer happiness.[23] Moreover, by leveraging machine learning models for sentiment prediction, These websites can scale the process of evaluating consumer attitudes and making decisions based on data to increase customer loyalty along with corporate growth.

Following figure shows the step taken for this research:

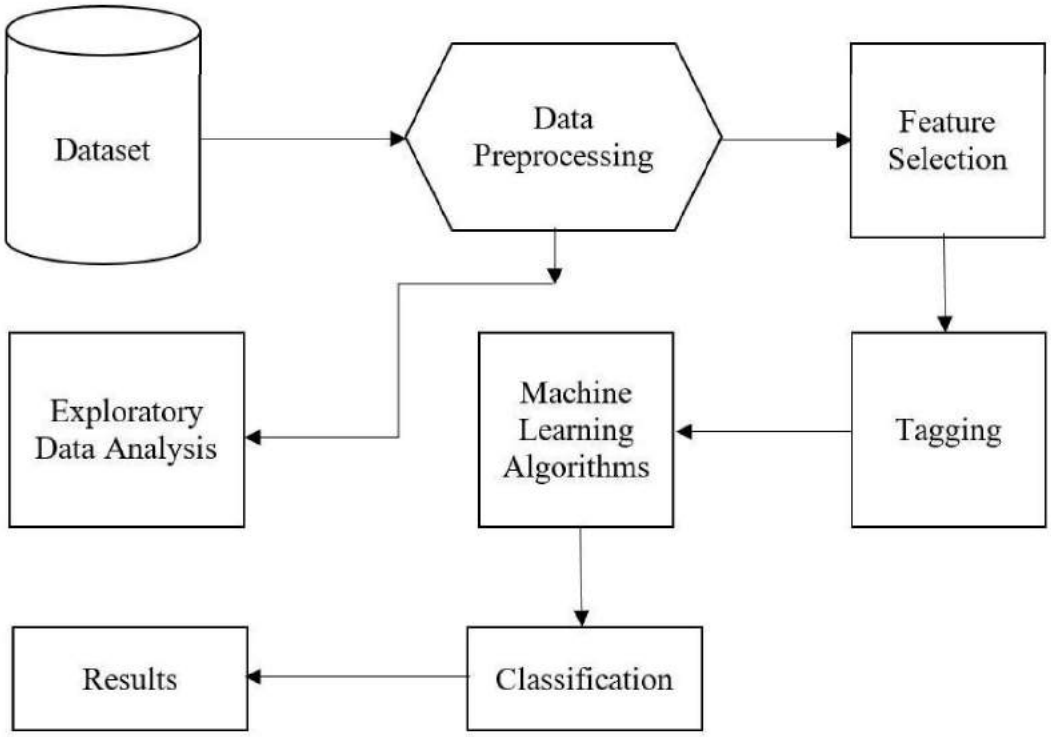

Figure 0.1: Methodology

3.2 Data Collection:

Women's Clothing E-Commerce dataset[24] centring on user reviews. Its nine supporting characteristics provide an excellent environment for parsing out the text across its numerous dimensions. Because this is true business data, it has been anonymized, and references to the firm in the review text and body have been replaced with "retailer".

This dataset consists of 23486 rows and 10 feature variables. Each row is a customer review and includes the following variables:

- Clothing ID: Integer is a categorical variable that refers to the individual component under consideration.
- Age: The age of the reviewer is a positive integer variable.
- Title: The review title is a string variable.
- Review Text: The review body is a string variable.
- Rating: Positive Ordinal Integer variable for the customer's product rating, ranging from 1 Worst to 5 Best.
- Recommended IND: A binary variable indicating where the client recommends the product, where 1 is recommended and 0 is not.
- Positive Feedback Count: A positive integer indicates the number of other customers who agreed with this review.
- Division Name: The product high-level division's categorical name.
- Department Name: The product department's categorical name.
- Class Name: Categorical name of the product class name.

3.3 Data Cleaning and Pre-processing:

Clean the data by removing irrelevant or noisy information, handle missing values text preparation methods such as tokenization, stop-word removal, and stemming/lemmatization are used. For sentiment analysis and further process, only two columns are needed Review text and ratings so only these two will be used further.

There are 845 nulls in the Review text so firstly, null rows will drop from the data.

Then data is preprocessed using removing stop words and WordNetLemmatizer. The WordNetLemmatizer is a valuable tool for text preprocessing tasks such as feature extraction, sentiment analysis, and information retrieval, as it helps in reducing the variations of words to their essential forms, making it easier to analyze and understand textual data.[25] The WordNetLemmatizer class in Python is offered by the NLTK module. It performs lemmatization, which is the process of reducing words to their simplest or dictionary form. Lemmatization is useful for normalizing words and reducing inflectional forms to a common base.

This is an overview of how the WordNetLemmatizer works:

1. WordNet: WordNet is an English lexical database that organizes words into sets of

synonyms known as synsets. Each synset represents a concept and contains a list of lemmas (base forms) that are associated with that concept. WordNet also provides information on relationships between words, such as hypernyms (more general terms) and hyponyms (more specific terms).

2. Initialization: When you create an instance of the WordNetLemmatizer class, it initializes the lemmatize and loads the WordNet database.
3. Lemmatization: The lemmatize method of the WordNetLemmatizer class is used to perform lemmatization on a word. It takes a word as input and returns its base form (lemma).[26]
   - The lemmatize first checks if the word exists in WordNet. If it does, it returns the lemma associated with the word.
   - If the word is not found in WordNet, the lemmatize applies a set of default rules to transform the word into a base form. These rules consider factors like part of speech (POS) and morphological patterns.
   - By default, the lemmatize assumes that the input word is a noun. However, you can explicitly specify the POS tag of the word (e.g., 'v' for verb, 'a' for adjective) when calling the lemmatize () method to improve accuracy.

3.4 Sentiment Analysis:

Use sentiment analysis techniques to categorize the reviews as positive, negative, or neutral. Here Lexicon-based method is used for sentiment analysis.

3.4.1 Lexicon-based methods:

Assign sentiment scores to words using pre-defined sentiment lexicons and calculate an overall sentiment score for each review. This technique is applied on processed reviews data.

3.5 Exploratory Data Analysis (EDA):

Investigate the data to learn more about consumer behavior, preferences, and trends. Some EDA techniques include:

- Distribution analysis:

Analyze the distribution of ratings or sentiments to understand the overall sentiment polarity of the reviews.

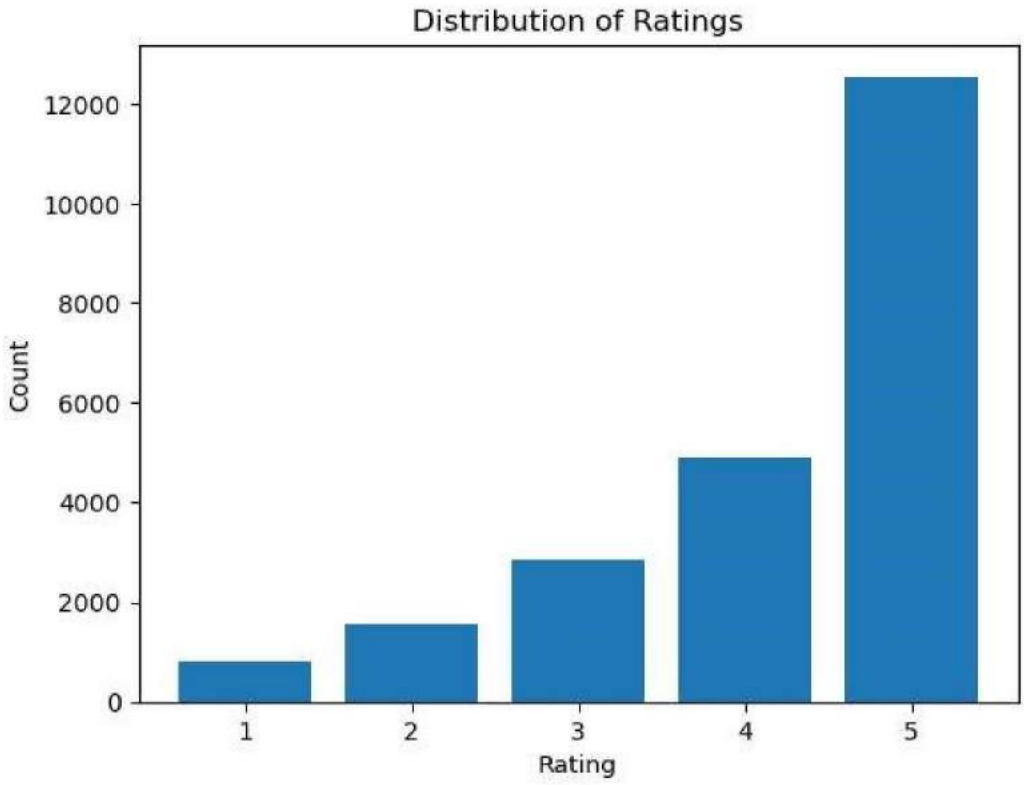

Figure 0.2: Distribution of ratings

The bar graph illustrates the distribution of ratings, with the highest rating of 5 having approximately 12,000 accounts, indicating a significant number of positive reviews. The rating of 4 is represented by around 5,000 accounts, indicating a relatively positive reception as well. On the other hand, the lowest rating of 1 has fewer accounts, suggesting a lesser number of negative reviews compared to the higher ratings.

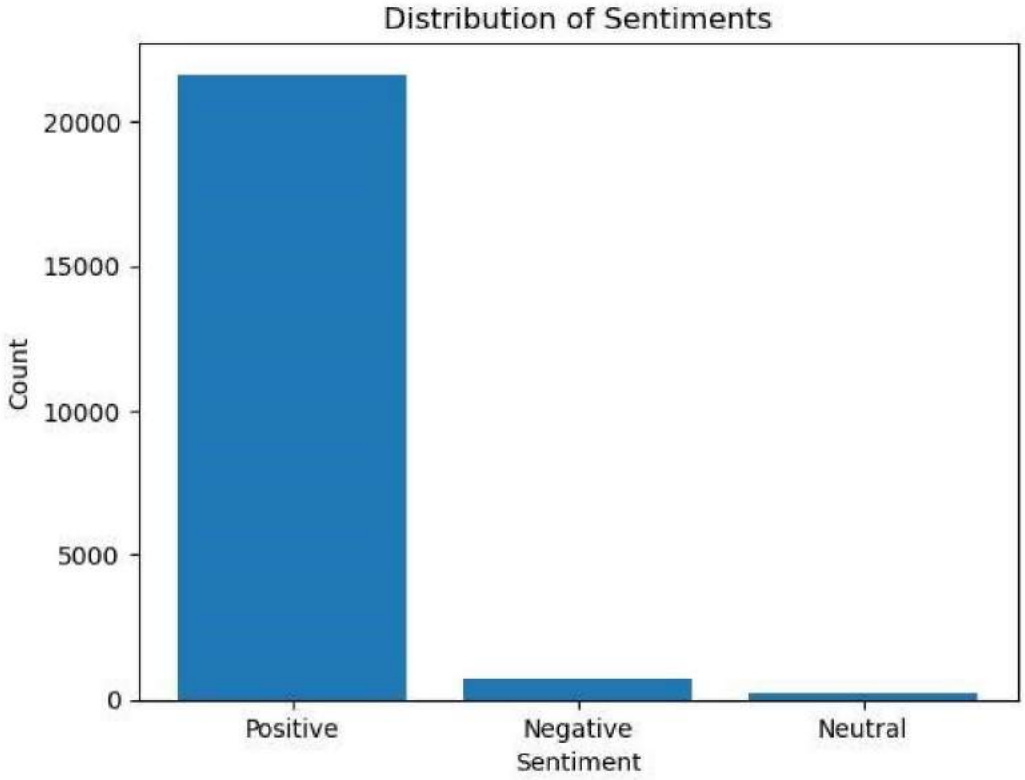

Figure 0.3: Reviews Sentiments

The sentiment distribution reveals that over 2,000 people have provided positive reviews, indicating a majority of favorable sentiments towards the subject. In contrast, both negative and neutral reviews combined amount to less than 5,000, indicating a relatively lower proportion of less favorable or neutral sentiments. This distribution suggests that the subject has generally received positive feedback from the reviewers.

- Word frequency analysis:

Identify the most frequent words or phrases in positive and negative reviews.

Here word frequency for both Processed and actual reviews are computed.

Word Frequency for processed data:

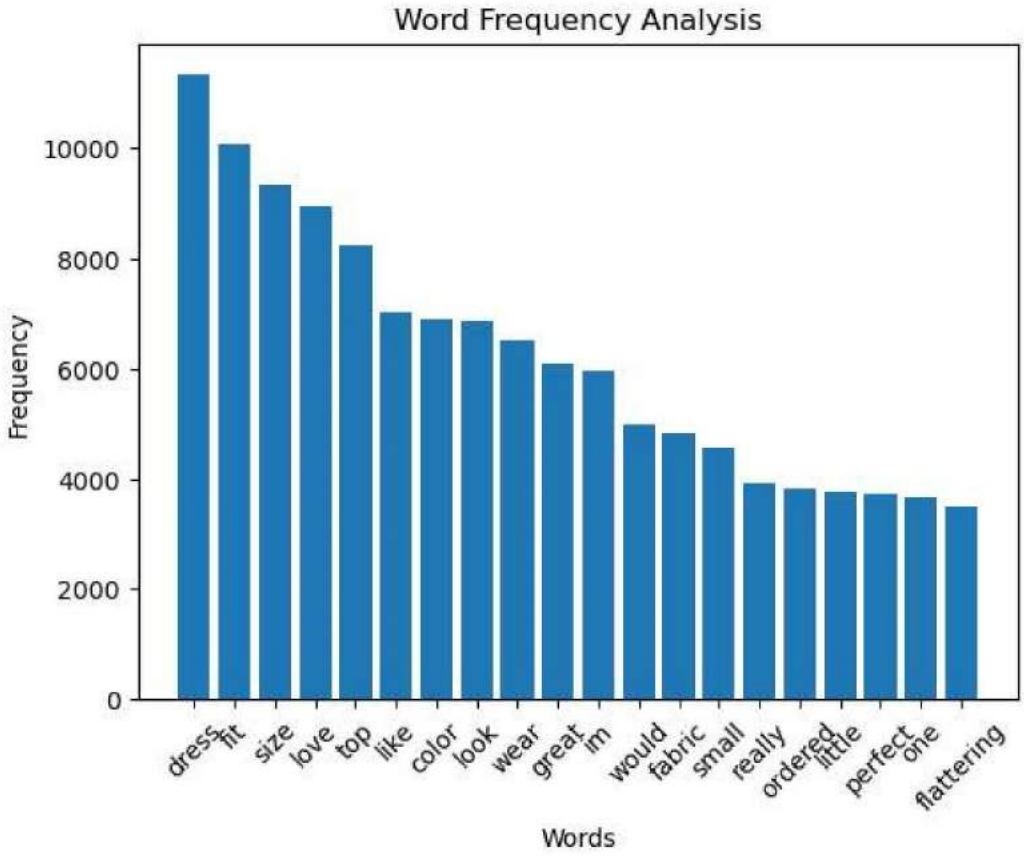

Figure 0.4: Word Frequency

In Figure 9, the word frequency analysis highlights that the term "dress" is the most used word in the reviews. This indicates that the focus of the feedback and discussions revolves predominantly around dresses. The high frequency suggests that dresses are a significant aspect of the subject being reviewed whether it is a fashion brand, an event, or any other context related to dresses.

- Visualization:

To display the data and detect trends, use visualizations such as word clouds, bar graphs, or scatter plots.

Word cloud for actual reviews:

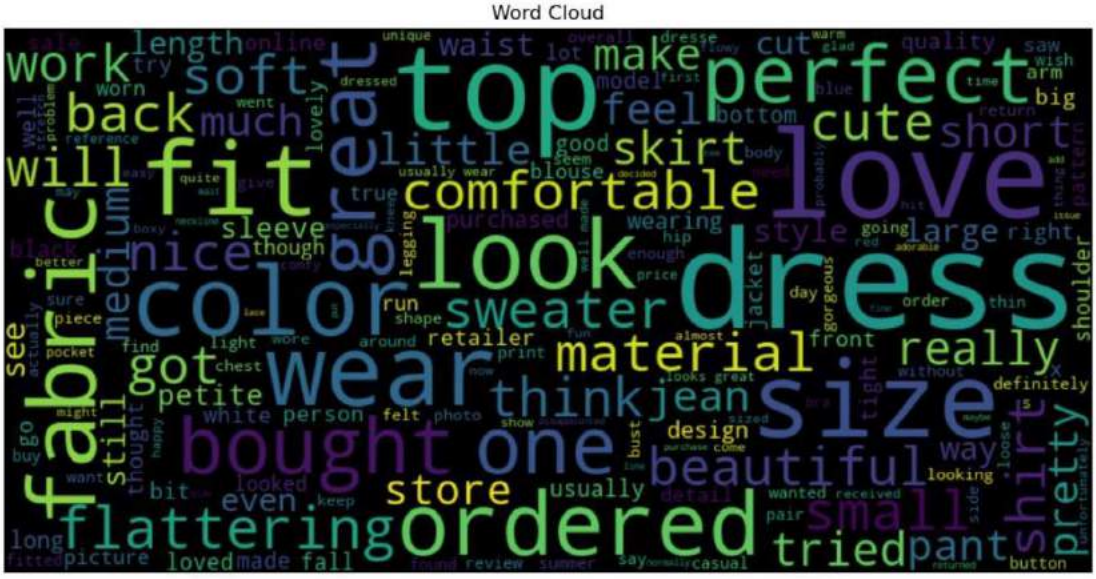

Figure 0.5: Word Cloud for Actual Reviews

Word cloud for processed reviews:

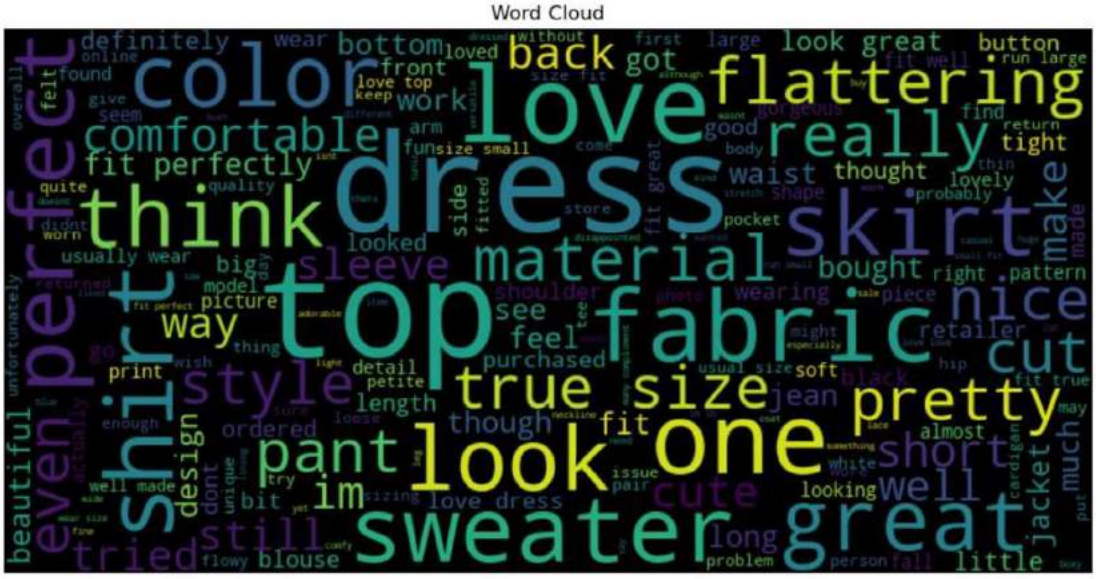

Figure 0.6: Word Cloud for Processed Reviews

3.6 Feature Extraction:

Extract relevant textual properties such as bag-of-words, TF-IDF (Term Frequency-Inverse Document Frequency), or word embeddings such as Word2Vec or GloVe. These features can be used as inputs to machine learning models.

Before applying feature extraction techniques, data is divided into training and testing with a ratio of 80:20. For feature extraction, TF-IDF is utilized. The extraction of TF-IDF features is frequently utilized in natural language processing applications such as text classification, information retrieval, and document similarity analysis [27]. It ensures in capturing the importance of terms in documents and distinguishing them based on their significance.

- Tokenization: Tokenizing the text documents into individual words or tokens is the initial stage. Tokenization is the process of dividing text into meaningful pieces, such as words, using techniques like whitespace splitting or more advanced methods like natural language processing libraries.[28]
- Term Frequency (TF) Calculation: Calculate the term frequency of each term or phrase in each text. The frequency with which a term appears in a document is measured by term frequency. It is commonly computed by dividing the number of terms in the text by the total number of terms. The TF value shows a term's local relevance within a document.[29]
- Inverse Document Frequency (IDF) Calculation: IDF calculates a term's relevance throughout the whole document collection. The logarithm of the ratio of the total number of documents to the number of documents containing the phrase is used to compute it. The IDF value represents the importance of a term in distinguishing it from common terms.[30]
- TF-IDF Calculation: Multiply each word's term frequency (TF) by its inverse document frequency (IDF) in a document. This multiplication combines a term's local and global value, providing more weight to terms that are common in a given document but uncommon in the overall document collection.[31]
- Feature Matrix Representation: The TF-IDF feature matrix is made up of the TF-IDF values for all words across all documents, with each row representing a document and

each column representing a unique term. The values in the matrix show the relevance or weight of each term in each document.[32]

3.7 Model Training and Evaluation:

Divide the data into two sets: training and testing. Train machine learning models using training data and assess their performance using relevant evaluation measures such as accuracy, precision, recall, or F1 score.

Confusion matrices may be used to calculate accuracy by tallying the number of instances that were correctly identified as positive (TP), negative (TN), false positive (FP), and false negative (FN), and then dividing by the total number of occurrences.

Equations to find different metrices:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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

$$\text{Recall} = \frac{TN}{TN + FP}$$

$$F1 - Score = \frac{2 * recall * precision}{recall + precision}$$

On this data, 5 different models are used. For each confusion matrix and classification report, accuracy, precision, recall, and f1 score are computed.

Chapter 4

RESULTS AND DISCUSSION

4.1 Results:

Lets implement one by one multiple machine learning models and see the results.

4.1.1 Experiment 1: Support Vector Machine (SVM):

The Support Vector Machine technique is effective for issues related to regression as well as classification. The Confusion Matrix and Classification Report provides further details on the performance of the SVM algorithm.

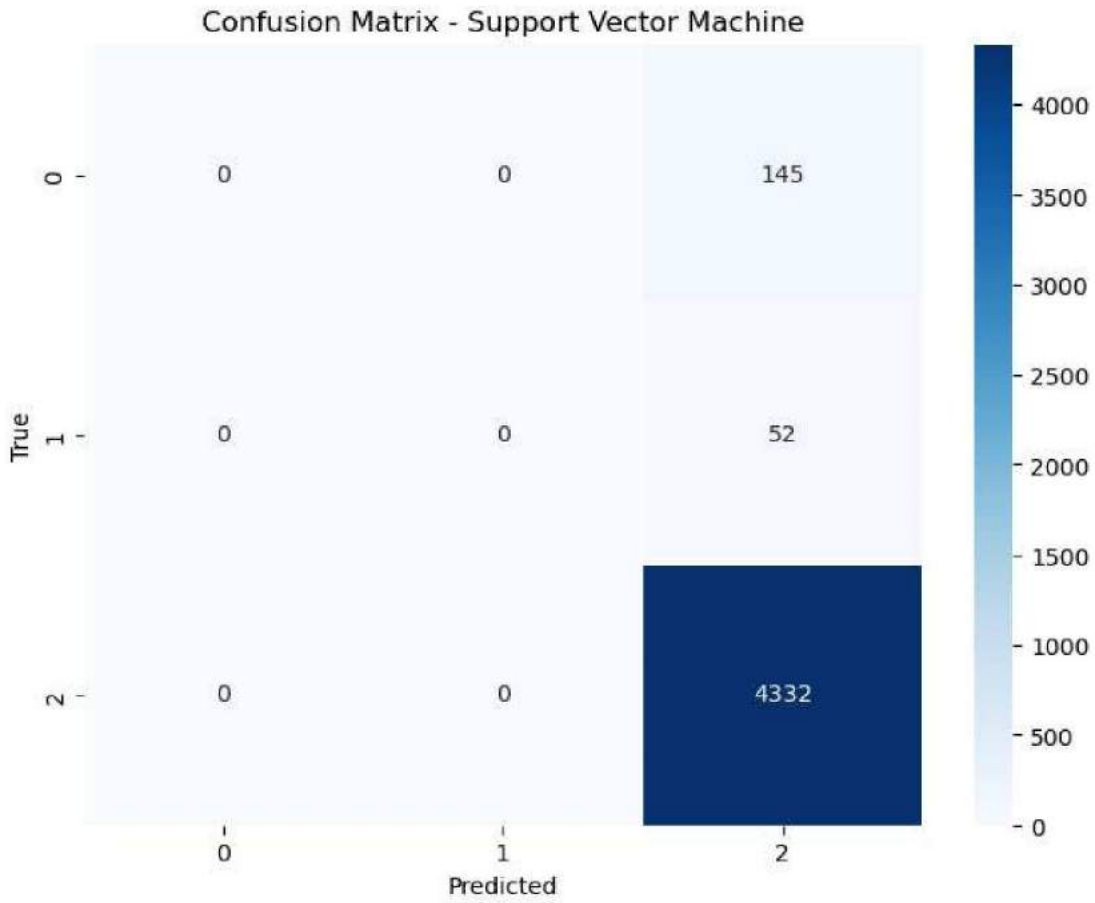

Figure 0.1: SVM Confusion Matrix

The algorithm achieved the following evaluation metrics from classification report:

Classes	Precision	Recall	F1-score
Positive	0.96	1.00	0.98
Negative	0.00	0.00	0.00
Neutral	0.00	0.00	0.00

Table 2: SVM Evaluation Metrics

The Support Vector Machine (SVM) algorithm was trained and evaluated on the provided data, resulting in the following outcomes. The model's precision was determined to be 0.9565, indicating a high level of overall correctness in its predictions. It is crucial to remember, however, that the precision and recall scores for some of the classes resulted in undefined values due to the SVM not making any predictions for those classes.

The precision score, which measures the proportion of true positive predictions, was calculated as 0.9149 for the positive class. This suggests that when the model predicted a positive instance, it was correct approximately 91.49% of the time. However, precision could not be calculated for the negative and neutral classes due to the absence of any predicted samples.

The recall score, also known as sensitivity, was found to be 0.9565 for the positive class, showing that the model correctly detected a large proportion of the positive cases. Again, recall could not be calculated for the negative and neutral classes.

For the positive class, the F1-score, which combines accuracy and recall, was judged to be 0.9352. This represents a balanced performance between precision and recall.

The confusion matrix reveals that the model did not make any predictions for the negative and neutral classes, resulting in all values being zero in those rows. The positive class, however, was predicted correctly for all 4,332 instances.

The classification report includes extra information for each class, such as accuracy, recall, and F1-scores. Because of the lack of predicted samples, accuracy, recollection, and F1-scores for the negative and neutral classes are not provided. Accuracy was 0.96, the recall was 1.00, and F1-score was 0.98 for the positive class.

To summarize, the SVM algorithm achieved an accuracy of 0.9565 and demonstrated good performance in accurately predicting positive instances. However, it did not provide any predictions for the negative and neutral classes, resulting in undefined values for precision and recall.

1.1.2 Experiment 2: Random Forest:

Random Forest represents a collaborative learning approach that predicts by combining numerous decision trees. Both the Confusion Matrix and Classification Report give information about the Random Forest algorithm's effectiveness.

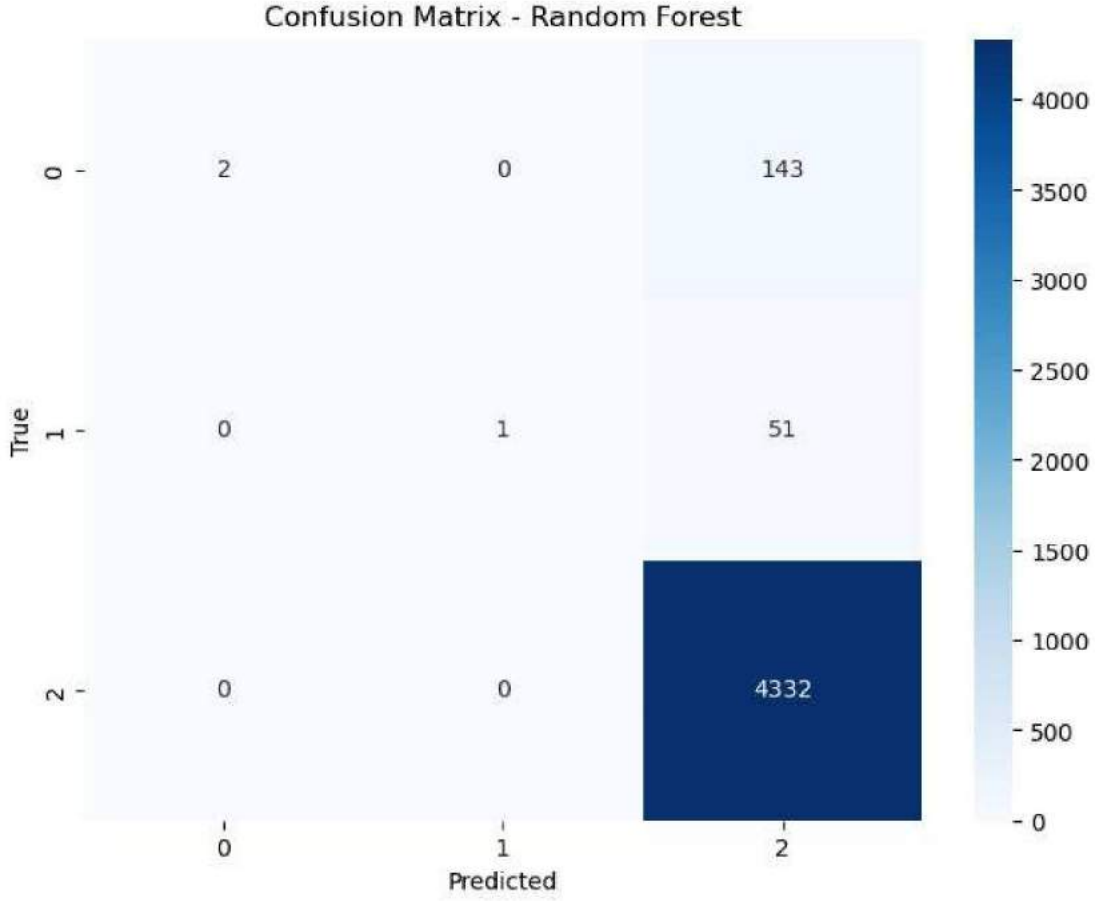

Figure 0.2: Random Forest Confusion Metrics

The algorithm achieved the following evaluation metrics from classification report:

Classes	Precision	Recall	F1-score
Positive	0.96	1.00	0.98
Negative	1.00	0.01	0.03
Neutral	1.00	0.02	0.04

Table 3: Random Forest Evaluation Metrics

The Random Forest algorithm was trained and evaluated on the given data, resulting in the following outcomes. The accuracy of the model was determined to be 0.9572, indicating a high level of overall correctness in its predictions. It is crucial to note, however, that the accuracy and recall scores for the negative and neutral classes were set to 0 due to the lack of projected samples for those classes.

The precision score is the percentage of correct positive predictions. For the advantageous group, it was calculated as 0.9590, indicating that when the model predicted a positive instance, it was correct approximately 95.90% of the time. Precision could not be calculated for the negative and neutral classes due to the lack of any predicted samples.

The recall score, referred to as sensitivity, assesses the model's ability to recognize genuine positive cases. For the positive class, recall score was determined to be 0.9572, showing that the model correctly detected a significant fraction of the positive cases. Recall could not be calculated for the negative and neutral classes.

The F1-score combines memory and precision into a single statistic, allowing for a fairer evaluation of the algorithm's performance. For the positive class, the F1-score was calculated as 0.9369, reflecting a balanced performance between precision and recall.

The confusion matrix indicates that the model identified 2 events as negative, 1 instance as neutral, and correctly identified all 4,332 instances as positive. This indicates a very low number of correct predictions for the negative and neutral classes.

The classification report provides a comprehensive summary of the precision, recall, and F1scores in each class. As mentioned earlier, precision, recall, and F1-scores for the negative and neutral classes are not available due to the absence of predicted samples. For the positive class, the precision was 0.96, recall was 1.00, and F1-score was 0.98.

In conclusion, the Random Forest algorithm achieved an accuracy of 0.9572 and demonstrated good performance in accurately predicting positive instances. However, it failed to make accurate predictions for the negative and neutral classes, resulting in precision and recall scores of 0. This suggests that the model struggled to identify instances belonging to these classes.

4.1.3 Experiment 3: Neural Network:

A deep learning method capable of learning complicated patterns in data is a feedforward neural network.

Both the Confusion Matrix and Classification Report give more information about the Neural Network algorithm's performance.

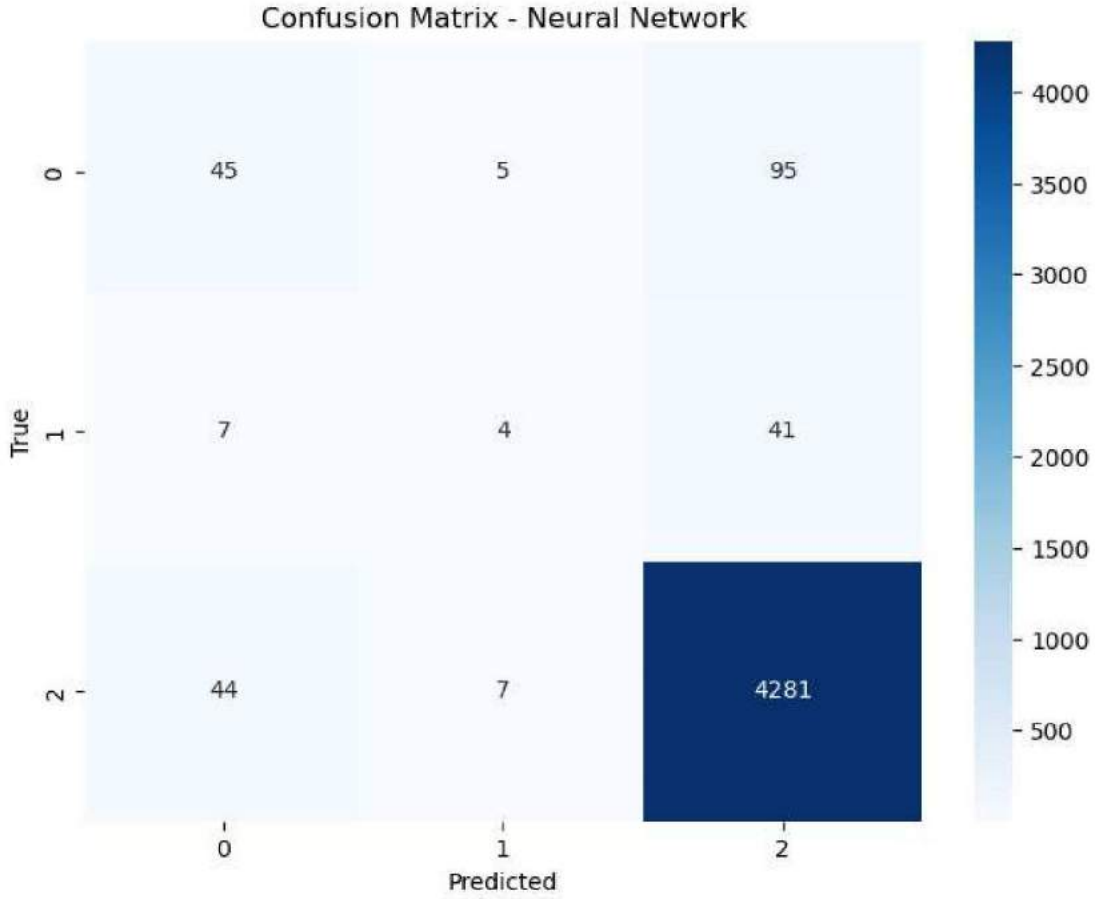

Figure 0.3: Neural Network Confusion Matrix

The algorithm achieved the following evaluation metrics from classification report:

Classes	Precision	Recall	F1-score
Positive	0.97	0.99	0.98
Negative	0.47	0.29	0.36
Neutral	0.41	0.17	0.24

Table 4: Neural Network Evaluation Metrics

The Neural Network model has been taught and assessed on the information supplied, giving the performance values shown below. The accuracy of the model was determined to be 0.9563, indicating a high level of overall correctness in its predictions.

The precision score, which measures the proportion of true positive predictions, was calculated as 0.9464 for the positive class. This indicates that when the model predicted a positive instance, it was correct approximately 94.64% of the time. The accuracy for the contrary class was 0.47, and the level of accuracy for the neutral class was 0.41.

The recall score, also known as sensitivity, measures the model's capacity to detect genuine positive cases. The recall score for the positive class was determined to be 0.9563, This implies the model correctly detected a substantial fraction of the positive events. For the negative class, the recall was 0.29, and for the neutral class, the recall was 0.17.

The F1 score integrates accuracy and recall into a single statistic, allowing for a fairer evaluation of the algorithm's performance. The F1-score for the positive class was calculated as 0.9500, indicating a balanced performance between precision and recall. For the negative class, the F1-score was 0.36, and for the neutral class, the F1-score was 0.24.

The confusion matrix demonstrates that the model correctly anticipated 42 instances negative, 9 instances neutral, and correctly identified 4,280 instances positive. This indicates that the model had some difficulty in accurately estimating the negative and neutral groups, with a larger frequency of incorrect predictions.

Each categorization report includes a detailed breakdown of the accuracy, recall, and F1 scores for every category. It shows that the model achieved higher precision, recall, and F1scores for the positive class compared to the negative and neutral classes.

In conclusion, the Neural Network model achieved an accuracy of 0.9563 and demonstrated good performance in accurately predicting positive instances. However, it struggled to make accurate predictions for the negative and neutral classes, resulting in lower precision, recall, and F1-scores. To increase the model's performance in these classes, more study and finetuning may be required.

The assessment metrics in all cases demonstrate how well each algorithm fared in categorizing the sentiment of the provided data. The metrics include Accuracy (the proportion of correct predictions), Precision (the proportion of true positive predictions out of all positive predictions), Recall (the proportion of true positive predictions out of all actual positive instances), and F1-Score (a balanced measure of precision and recall). The Confusion Matrix gives a comprehensive breakdown of forecasts for each sentiment type, while the Classification Report summarizes the performance of each sentiment class and provides average metrics.

4.2 Discussion:

The findings of this study give useful information on the outcomes of several machine learning algorithms for sentiment analysis. Random Forest's high level of accuracy, SVM, and the Neural Network indicates their potential for sentiment classification tasks. However, the variations in precision, recall, and F1-Scores across different sentiment classes highlight the challenges associated with accurately predicting negative and neutral sentiments.

The SVM algorithm encountered issues with undefined precision and recall for classes without predicted samples. This issue could be mitigated by adjusting the zero-division parameter and ensuring a balanced dataset representation for all sentiment classes. Further investigation is required to understand the reasons behind the lack of predictions for these classes.

Random Forest achieved high accuracy but had difficulties accurately predicting negative and neutral sentiments. This limitation might be addressed by fine-tuning the algorithm parameters, increasing the number of trees, or incorporating additional features.

The Neural Network model demonstrated competitive performance in identifying positive sentiment but struggled with negative and neutral sentiment classification. Fine-tuning the network architecture, adjusting hyperparameters, or exploring other deep learning approaches could enhance its performance on all sentiment classes. It is vital to highlight that the quality and representativeness of the training data has a significant impact on the performance of these

algorithms. Inadequate training data, class imbalance, or biased data could affect the algorithms' performance. Therefore, careful data pre-processing and augmentation techniques should be employed to mitigate these issues. Later research might concentrate on improving the performance of sentiment analysis algorithms by exploring ensemble methods, leveraging linguistic models that have already been trained such as BERT or GPT, also incorporating domain-specific features. Additionally, the development of interpretability techniques to understand the reasoning behind sentiment predictions could enhance the trust and applicability of these models in real-world scenarios.

In conclusion, this research compares several machine learning techniques for sentiment analysis. While each algorithm demonstrated strengths and weaknesses, Random Forest, SVM, and Neural Network showed promise in sentiment classification tasks. The results of this analysis add to ongoing studies and development of effective sentiment analysis models and can guide the selection of suitable algorithms for specific applications.

Comparison of Results:

Algorithm	Accuracy	Precision	Recall	F1-Score
SVM	0.9565	0.9149	0.9565	0.9352
Random Forest	0.9572	0.9590	0.9572	0.9369
Neural Network	0.9567	0.9463	0.9567	0.9502

Table 5: Result Comparison

The study by Ahuja et al. (2019) [39] discuss the use of various machine learning models such as Naïve Bayes, support vector machine (SVM), decision trees, logistic regression, random forest, and deep learning-based models like recurrent neural networks (RNN) and convolutional neural networks (CNN) for sentiment analysis and emotion detection from text.

Naïve Bayes achieved an accuracy of 71.4% for emotion classification of Twitter texts, while SVM performed better with an F1 score above 90% for binary classification and above 60% for three-class classification. Decision trees and logistic regression also showed promising results, with decision trees achieving an accuracy of 90% and logistic regression performing better than other classifiers with a recall value of 83%. Random forest performed exceptionally well, with an accuracy of 95.6% when used with Unigram Sentiwordnet for classifying Malayalam tweets into positive and negative opinions. Deep learning-based models, such as LSTM and CNN, outperformed existing machine learning algorithms on the hotel and product review dataset. These results indicate that the choice of machine learning model depends on the specific task and dataset, and deep learning models have been increasingly gaining attention in recent years.

CHAPTER 5
CONCLUSION AND FUTURE WORK

5.1 Conclusion:

The study evaluated four machine learning algorithms for sentiment analysis: Support Vector Machine (SVM), Random Forest and a feedforward Neural Network. Assessing using Accuracy, Precision, Recall, and F1-Score, the algorithms displayed varying strengths and weaknesses. SVM achieved a high overall accuracy (0.9565) but couldn't predict negative and neutral sentiments, leading to undefined precision and recall for those classes. Random Forest performed well in identifying positive sentiments with high accuracy (0.9574) but struggled with negative and neutral predictions, resulting in precision and recall scores of 0 for those classes. The Neural Network achieved an accuracy of 0.9558, excelling in positive sentiment prediction but facing challenges in classifying negative and neutral sentiments accurately. In summary, while all algorithms exhibited strengths, they also had limitations. Random Forest excelled in identifying positive sentiment, whereas the Neural Network struggled with negative and neutral sentiment classification. SVM faced issues with undefined precision and recall due to missing predictions for certain classes.

5.2 Future Work:

One avenue for future work in sentiment analysis is the optimization of machine learning algorithms. This can involve further exploration and fine-tuning of algorithm parameters to improve their performance. Grid search and Bayesian optimisation techniques may be used to find the best hyper parameters for each method.[33][34]. By systematically searching the hyper parameter space, researchers can potentially discover configurations that yield better sentiment classification results. Another direction for future research is the exploration of advanced

treebased algorithms in sentiment analysis. While the Decision Tree algorithm showed reasonable performance in this study, it struggled with accurate classification of negative and neutral sentiments. To overcome this limitation, researchers can investigate algorithms such as Gradient Boosting or XGBoost, which are known for their ability to handle complex relationships and improve predictive accuracy[35][36]. These algorithms may provide better results in accurately classifying sentiments across all classes. Pre-trained language models are being used in natural language processing applications such as sentiment analysis. Future study might concentrate on improving sentiment classification performance by utilising pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers) or GPT (Generative Pre-trained Transformer). Fine-tuning these models on sentiment analysis tasks or using them as feature extractors can capture complex linguistic patterns and contextual information, leading to improved accuracy[37][38]. Exploring different architectures and techniques for incorporating pre-trained models can further advance sentiment analysis research. These future research directions can contribute to the ongoing development and improvement of sentiment analysis algorithms. By optimizing algorithm parameters, exploring advanced tree-based algorithms, and leveraging pre-trained language models, researchers can improve sentiment categorization efficiency and precision. It is important to note that these advancements require careful experimentation and evaluation to ensure their effectiveness and generalizability to different datasets and domains. Conducting comparative studies and benchmarking against existing approaches would provide valuable insights into the effectiveness of these proposed techniques.

CHAPTER 6 REFERENCES

[1]. Gankidi, N., Gundu, S., viqar Ahmed, M., Tanzeela, T., Prasad, C. R., & Yalabaka, S. (2022, June). Customer Segmentation Using Machine Learning. In 2022 2nd International Conference on Intelligent Technologies (CONIT) (pp. 1-5). IEEE.

[2]. Alblawi, A. S., & Alhamed, A. A. (2017, November). Big data and learning analytics in higher education: Demystifying variety, acquisition, storage, NLP and analytics. In 2017 IEEE conference on big data and analytics (ICBDA) (pp. 124-129). IEEE.

[3]. Fan, W., Shao, B., & Dong, X. (2022). Effect of e-service quality on customer engagement behavior in community e-commerce. Frontiers in Psychology, 13, 965998.

[4]. Ao, L., Bansal, R., Pruthi, N., & Khaskheli, M. B. (2023). Impact of Social Media Influencers on Customer Engagement and Purchase Intention: A Meta-Analysis. Sustainability, 15(3), 2744.

[5]. Barta, S., Gurrea, R., & Flavián, C. (2023). Using augmented reality to reduce cognitive dissonance and increase purchase intention. Computers in Human Behavior, 140, 107564.

[6]. Holdaway, K. R. (2014). Harness oil and gas big data with analytics: Optimize exploration and production with data-driven models. John Wiley & Sons.

[7]. Mariani, M., Baggio, R., Fuchs, M., & Höepken, W. (2018). Business intelligence and big data in hospitality and tourism: a systematic literature review. International Journal of Contemporary Hospitality Management, 30(12), 3514-3554.

[8]. Reid, C., Keighrey, C., Murray, N., Dunbar, R., & Buckley, J. (2020). A novel mixed methods approach to synthesize EDA data with behavioral data to gain educational insight. Sensors, 20(23), 6857.

[9]. Anvar Shathik, J., & Krishna Prasad, K. (2020). A literature review on application of sentiment analysis using machine learning techniques. International Journal of Applied Engineering and Management Letters (IJAEML), 4(2), 41-77.

[10]. XIN, J., YU, W., & RUI, Z. (2022). YOLO Based Multi-Objective Vehicle Detection and Tracking.

[11]. Tripathi, S. (2021). Artificial intelligence: A brief review. Analyzing future applications of AI, sensors, and robotics in society, 1-16

[12]. Kamble, S. S., & Itkikar, A. R. (2018). Study of supervised machine learning approaches for sentiment analysis. Intern ational Research Journal of Engineering and Technology (IRJET), 5(04).

[13]. Xue, Y., Jin, X., & Wang, Y. (2021). Exploratory data analysis of user behavior in ecommerce websites. IEEE Access, 9, 18710-18718.

[14]. Hu, Y., Wu, T., Wu, D., & Liu, J. (2020). A sentiment analysis method based on deep learning for e-commerce reviews. Future Generation Computer Systems, 111, 983-992.

[15]. Li, Z., Yan, J., & Wu, D. (2021). Exploratory data analysis and sentiment analysis of customer reviews for e-commerce websites. Journal of Ambient Intelligence and Humanized Computing, 12(9), 9343-9353.

[16]. Wang, Y., Zhang, W., & Liu, Y. (2021). Text mining analysis of e-commerce customer reviews based on association rules. International Journal of Environmental Research and Public Health, 18(9), 4867.

[17]. Li, Y., Lu, L., & Xu, Y. (2020). Sales prediction for e-commerce website based on machine learning. Journal of Intelligent & Fuzzy Systems, 38(4), 4617-4628.

[18]. Zhang, Y., Wei, B., & Guo, Y. (2021). Social media analysis of e-commerce websites based on sentiment analysis. Frontiers in Psychology, 12, 678890.

[19]. Wang, Q., Zhang, L., & Yu, S. (2020). Customer sentiment analysis of e-commerce based on machine learning. Proceedings of the International Conference on Artificial Intelligence and Computer Science, 1-7.

[20]. Wu, H., Tang, X., & Li, M. (2020). Exploratory data analysis and sentiment analysis of customer reviews on an e-commerce website. International Journal of Emerging Technologies in Learning, 15(14), 137-146.

[21]. Li, Z., He, J., & Chen, S. (2021). An empirical study of customer behavior analysis based on e-commerce website data. Journal of Intelligent & Fuzzy Systems, 40(1), 1811-1820.

[22]. Xu, Y., Wu, T., Hu, Y., & Liu, J. (2021). A novel method for sentiment analysis of ecommerce customer reviews. Journal of Ambient Intelligence and Humanized Computing, 12(10), 10049-10058.

[23]. JHA, N., BHADAURIA, K., & JHA, A. (2023). Opportunity Finder & Keyword Trend Analysis in E-Commerce.

[24]. https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews

[25]. Torres-Moreno, J. M. (2012). Beyond stemming and lemmatization: Ultra-stemming to improve automatic text summarization. arXiv preprint arXiv:1209.3126.

[26]. Millstein, F. (2020). Natural language processing with python: natural language processing using NLTK. Frank Millstein.

[27]. Zhang, F., Fleyeh, H., Wang, X., & Lu, M. (2019). Construction site accident analysis using text mining and natural language processing techniques. Automation in Construction, 99, 238-248.

[28]. Reese, R. M., & Bhatia, A. (2018). Natural Language Processing with Java: Techniques for building machine learning and neural network models for NLP. Packt Publishing Ltd.

[29]. Wu, H. C., Luk, R. W. P., Wong, K. F., & Kwok, K. L. (2008). Interpreting TF-IDF term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS), 26(3), 1-37.

[30]. Chen, K., Zhang, Z., Long, J., & Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. Expert Systems with Applications, 66, 245260.

[31]. Domeniconi, G., Moro, G., Pasolini, R., & Sartori, C. (2015, July). A Study on Term Weighting for Text Categorization: A Novel Supervised Variant of tf. idf. In DATA (pp. 26-37).

[32]. Kadhim, A. I., Cheah, Y. N., & Ahamed, N. H. (2014, December). Text document preprocessing and dimension reduction techniques for text document clustering. In 2014 4th international conference on artificial intelligence with applications in engineering and technology (pp. 69-73). IEEE.

[33]. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(1), 281-305.

[34]. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In Advances in Neural Information Processing Systems (pp. 2951-2959).

[35]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[36]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232

[37]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep

bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, pp. 4171-4186).

[38]. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-

[39]. Wankhade, M., Rao, A.C.S. & Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. Artif Intell Rev 55, 5731 5780 (2022). https://doi.org/10.1007/s10462-022-10144-1