DEEP LEARNING DRIVEN SENTIMENT ANALYSIS OF E-COMMERCE CONSUMER IMPRESSIONS USING ADVANCED FETURE EXTRACTION TECHNIQUES

A THESIS SUBMITTED TO
THE FACULTY OF ARCHITECTURE AND ENGINEERING
OF
EPOKA UNIVERSITY

BY

MARJELA PRODA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JUNE 2024

Approval sheet of the Thesis

This is to certify that we have read this thesis entitled **"Deep Learning Driven Sentiment Analysis of E-commerce Consumer Impressions Using Advanced Feature Extraction Techniques"** and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Assoc. Prof. Dr. Arban Uka
Head of Department
Date: June 6, 2024

Examining Committee Members:

Assoc. Prof. Dr. Dimitrios Karras (Computer Engineering) _____

Prof. Dr. Betim Çiço (Computer Engineering) _____

Dr. Florenc Skuka (Computer Engineering) _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Surname: Marjela Proda

Signature: _____

ABSTRACT

DEEP LEARNING DRIVEN SENTIMENT ANALYSIS OF E-COMMERCE CONSUMER IMPRESSIONS USING ADVANCED FETURE EXTRACTION TECHNIQUES

Proda, Marjela

M.Sc., Department of Computer Engineering

Supervisor: Dr. Prof. Dr. Betim Çiço

E-commerce has emerged as one of the biggest players in the current digitized business environment, and this has led to the creation of large amounts of consumer data through consumer reviews and feedback. The objective of this master thesis is to identify the consumer impression in the e-commerce data by applying sophisticated feature extraction techniques and sentiment analysis based on deep learning approaches. This paper seeks to explore the elements of consumer sentiment as captured in online reviews, which is vital in increasing customer satisfaction and sales. The research problem seeks to establish the performance of different machine learning models in sentiment analysis of e-commerce reviews and feature extraction techniques such as TF-IDF and Word2Vec. The main goal is to identify which set of machine learning models and feature extraction methods gives the best accuracy and efficiency in sentiment analysis.

The methodology includes a systematic review of the literature in order to identify the current sentiment analysis methods and their uses in e-commerce. The analysis utilises a collection of Amazon product reviews, which is first cleaned, tokenized, and balanced before being used in the study. Thus, four machine learning models, including Support Vector Machine (SVM), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Bidirectional Encoder Representations from Transformers (BERT), are chosen for the comparison. These models are then

optimized and assessed with numerous evaluation metrics like accuracy, precision, recall, and F1 score.

Empirical Findings show that deep learning models especially BERT exhibit higher accuracy than traditional machine learning models in the sentiment analysis task because they can analyze the context and language features of the text. BERT provided the highest accuracy thus showing its effectiveness in handling the sentiment analysis of consumer reviews. The study also focuses on the significance of feature selection where TF-IDF and Word2Vec improve the results of the model.

The study outcome shows that the combination of the advanced feature extraction technique with the deep learning model is useful in developing a robust framework for sentiment analysis in the e-commerce context. This approach allows organizations to acquire a better understanding of customers' tendencies and issues, which helps in decision-making and improves customer engagement. Further research will focus on the development of the hybrid models and live sentiment analysis to improve the overall performance and usability of the proposed approach for dynamic e-commerce scenarios.

The study outcome shows that the combination of the advanced feature extraction technique with the deep learning model is useful in developing a robust framework for sentiment analysis in the e-commerce context. This approach allows organizations to acquire a better understanding of customers' tendencies and issues, which helps in decision-making and improves customer engagement. Further research will focus on the development of the hybrid models and live sentiment analysis to improve the overal

Keywords: *E-commerce, Sentiment Analysis, Consumer Reviews, Feature Extraction, Deep Learning, Machine Learning Models, TF-IDF, BERT*

ABSTRAKT

ANALIZA E PËRSHTYPJEVE TË KONSUMATORËVE TË E-COMMERCE E DREJTUAR NGA MODELET DEEP LEARNING DUKE PËRDORUR TEKNIKAT E AVANCUARA TË NXJERRJES SË KARAKTERISTIKAVE

Proda, Marjela

Master Shkencor, Departamenti i Inxhinierisë Kompjuterike

Udhëheqësi: Prof. Dr. Betim Çiço

E-commerce është shfaqur si një nga lojtarët më të mëdhenj në mjedisin aktual të biznesit të digjitalizuar, dhe kjo ka çuar në krijimin e sasive të mëdha të të dhënave të konsumatorëve nëpërmjet shqyrtimeve dhe feedback-ut të konsumatorëve. Qëllimi i kësaj teze master është identifikimi i përshtypjes së konsumatorit në të dhënat e e-commerce duke aplikuar teknika të sofistikuara të nxjerrjes së veçorive dhe analiza ndjenjash bazuar në qasjet e thella të të mësuarit. Ky shkrim kërkon të eksplorojë elementet e ndjenjës së konsumatorit siç janë kapur në shqyrtimet online, gjë që është jetike në rritjen e kënaqësisë së klientit dhe shitjeve.

Problemi i kërkimit kërkon të vendosë performancën e modeleve të ndryshme të mësimit të makinerive në analizën sentimentale të shqyrtimeve të e-commerce dhe teknikave të nxjerrjes së veçorive si TF-IDF dhe Word2Vec. Qëllimi kryesor është të identifikohet se cili grup i modeleve të mësimit të makinerive dhe metodave të nxjerrjes së veçorive jep saktësinë dhe efikasitetin më të mirë në analizën e ndjenjave.

Metodologjia përfshin një rishikim sistematik të literaturës në mënyrë që të identifikohen metodat aktuale të analizës së ndjenjave dhe përdorimet e tyre në tregtinë elektronike. Analiza përdor një koleksion të shqyrtimeve të produkteve Amazon, të

cilat së pari pastrohen, tokenizohen dhe ekuilibrohen para se të përdoren në studim. Kështu, për krahasimin janë zgjedhur katër modele të mësimit të makinerive, duke përfshirë Support Vector Machine (SVM), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) dhe Bidirectional Encoder Representations from Transformers (BERT). Këto modele pastaj optimizohen dhe vlerësohen me metrike të shumta vlerësimi si saktësia, saktësia, kujtesa dhe rezultati F1. Gjetjet empirike tregojnë se modelet e të mësuarit të thellë veçanërisht BERT shfaqin saktësi më të lartë se modelet tradicionale të mësimit të makinerisë në detyrën e analizës së ndjenjave, sepse mund të analizojnë kontekstin dhe veçoritë gjuhësore të tekstit. BERT dha saktësinë më të lartë duke treguar kështu efektivitetin e tij në trajtimin e analizës së ndjenjave të shqyrtimeve të konsumatorëve. Këto modele pastaj optimizohen dhe vlerësohen me metrike të shumta vlerësimi si saktësia, saktësia, kujtesa dhe rezultati F1. Gjetjet empirike tregojnë se modelet e të mësuarit të thellë veçanërisht BERT shfaqin saktësi më të lartë se modelet tradicionale të mësimit të makinerisë në detyrën e analizës së ndjenjave, sepse mund të analizojnë kontekstin dhe veçoritë gjuhësore të tekstit. BERT dha saktësinë më të lartë duke treguar kështu efektivitetin e tij në trajtimin e analizës së ndjenjave të shqyrtimeve të konsumatorëve.

Fjalët kyçe: *Tregtia Online, Analiza e Sentimentit, Përshtypjet e Konsumatorëve, Nxjerrja e Karakteristikave, Deep Learning, Modelet Machine Learning, TF-IDF, BERT*

ACKNOWLEDGEMENTS

Throughout my graduate journey, many individuals have played a crucial role in making this experience both valuable and memorable. First and foremost, I wish to express my deepest gratitude to Betim Çiço, my major professor and dissertation supervisor. His guidance and collaboration over the years have been profoundly intellectually rewarding and fulfilling.

I am especially thankful to my friends for their unwavering patience in helping me navigate numerous questions and issues.

Lastly, I owe my deepest appreciation to my family. Their endless patience and encouragement have been the foundation of my strength, for which I am eternally grateful.

.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Technology transition is unavoidable, and it will alter the way we live our daily lives. The importance of e-commerce is growing, particularly in a post-covid economy. Global e-commerce revenues are expected to reach US$4,117.00 billion by 2024, according to eMarketer Ethan Cramer-Flood (2020). Due to the continuous expansion of E-commerce, firms are creating huge amounts of data. The data is being received from various sources, such as social media, consumer reviews, feedback, and ratings. With the increasing number of consumers shifting towards online purchasing, businesses are quickly allocating resources towards digital transformation and leveraging Artificial Intelligence to enhance their products and meet customer expectations more effectively.

A typical consumer uses internet channels to buy more than one product in their lifetime. Out of these, the customers who are happy with the goods usually leave positive feedback that other people can use. The unsatisfied customer will react exactly the other way, stating their dissatisfaction along with criticism. It is a simple approach overall; whatever the customer feels is communicated directly and given a value between 1 and 5, with 1 being the lowest and 5 the highest. The difficulty occurs when the top rating given to a product was entirely based on personal preference or a specific component of that product, yet someone who purchased the same thing based on the same feedback may not have had the same experience. For example, one consumer may be more concerned with the product's pricing, while another is concerned with its location. Thus, the overall rating does not reflect the genuine level of consumer happiness.

A single review typically includes numerous components of the consumer's expressed sentiments. Positive thoughts are frequently expressed using negative phrases. Such reviews present a challenge for standard systems to classify as positive. According to Madhusudhan Aithal (2021), their experiment revealed that the majority of the negative words in their dataset were followed by a positive term. In addition, phrases containing positive words can often be interpreted as negative, depending on the classifier. Sentiment

analysis faces numerous obstacles due to variations in how consumers communicate their feelings. There will be instances of sarcasm, emoticons, and slang terms, among others. Consequently, it becomes a challenging task to examine the human sentiment conveyed through the text.

Extracting pertinent aspects from textual data, such as words, phrases, sentiment indicators, or grammatical structures, is crucial for gaining insights into the expressed sentiment. Proficient feature extraction necessitates an in-depth understanding of the language, context, and domain employed in the textual material. The primary objective of feature extraction is to generate a collection of characteristics that may be utilised to train a sentiment analysis model. An advanced and comprehensive set of features can greatly enhance the precision and efficiency of the sentiment analysis model, allowing it to accurately detect sentiment in previously unseen and unfamiliar text material. To summarise, sentiment analysis and feature extraction are intimately interconnected and crucial elements in the creation of accurate and dependable sentiment analysis models.

When faced with an abundance of data, businesses need to carefully analyse consumer feedback to improve the customer experience and drive sales. By analysing this data, e-commerce businesses can tailor their products and services to meet and exceed customer expectations, resulting in an enhanced customer experience. Having a clear understanding of customer sentiment, preferences, and pain areas is crucial for developing effective marketing campaigns that can be tailored to a specific audience. Companies in the e-commerce industry, such as Amazon, Zalando, eBay, and many others, experience a high volume of product sales. Customer reviews play a crucial role in influencing the purchasing decisions of other users, leading to shifts in their consumption choices. Direct customer experiences can serve as valuable insights for businesses to identify and resolve issues, ensuring the maintenance of their quality.

1.1 Business Application and background of sentiment analysis

A study conducted by Haque et al. (2018) found that a significant majority of online shoppers rely heavily on reviews, considering them to be as valuable as personal recommendations. The extensive wealth of human knowledge that is currently accessible has fueled the advancement of sentiment and opinion mining. With the progress in AI and

the emergence of ChatGPT, there has been a significant change in how businesses and consumers view AI. An important application involves recommendation systems that rely on sentiment analysis. In a recent study by Elzeheiry et al. (2023), the importance of sentiment analysis and word embeddings has been highlighted. Highlighting the utilisation of consumer-centric (reviewer) features to enhance accuracy, showcasing the effectiveness of feature extraction techniques such as Word2Vec or TF-IDF.

Managing and maintaining the reputation of a brand

Online marketplaces such as Amazon provide businesses with a platform to reach consumers and offer their products. When consumers begin sharing their feedback on their satisfaction levels, businesses and brands can gain valuable insights into the overall sentiment towards their products. This feedback allows them to understand the satisfaction levels both on a general level and at an individual level. By leveraging sentiment analysis, businesses can gain valuable insights into how their brand is perceived by clients and the general public. They have a keen eye for identifying patterns in sentiment, whether they are positive or negative shifts, recurring issues, or emerging concerns. By utilising this data, companies can gain a deeper understanding of client needs, identify potential areas for growth, and make informed decisions to enhance their brand's reputation. By leveraging this process, these brands can effectively enhance their reputation by highlighting their top-selling products and proactively addressing customer concerns.

When consumers are considering a purchase, they often find themselves more inclined to make the purchase after consulting multiple reviews (Andrienko et al., 2021). According to Chen (2022), there is an obvious connection between the reviewer who has thoroughly examined this product and other consumers who will make their purchasing decisions based on the reviewer's opinion. The author also discusses how brands can allocate their marketing budgets by having their products reviewed by influential individuals who have strong opinions.

Conducting market research and competitor analysis

Utilising sentiment analysis, brands can effectively pinpoint shifting market trends based on the direction of sentiment. With this in mind, these brands have the ability to fine-tune their products to better align with market demands. In addition, the available data can

be used to conduct a competitor brand analysis. This information is crucial for gaining a competitive advantage. For instance, Brand A has the capability to analyse the sentiment surrounding Brand B's recently released product line. They have a deep understanding of consumer sentiment, can easily identify gaps in the market, and have the ability to develop innovative products that outshine their competitors.

According to a study by Rambocas and Pacheco, (2018), it seems that marketers tend to heavily rely on a single source for reviews or comments. Exploring the wide range of venues available is crucial in order to effectively address biases. The researcher delves into the impact of brand-specific comment tracking costs on small-scale businesses. Therefore, it is advisable for businesses to consider using open-source tools such as Python NLTK.

Understanding consumer feedback and developing effective marketing strategies

The research conducted by Ghose and Ipeirotis, (2011), highlights the importance of the structure of consumer reviews. Product reviews are typically divided into two parts: subjective and objective. If the review was solely subjective or solely objective, it was found that product sales were associated. However, if it had a combination of both, it was often seen in a negative light.

In addition, Ghose and Ipeirotis, (2011) found that the readability of reviews had a direct impact on product sales. For products in the electronic category that had a combination of objective and less subjective components, their performance was quite impressive and accurately described. It is clear that consumer feedback plays a crucial role in boosting product sales. This information can assist businesses in developing effective marketing strategies, such as encouraging consumers to provide detailed and well-curated reviews. Applications can be deployed to process feedback automatically and group them into different categories based on their priority.

1.2 Objective

A significant proportion of the world's data was generated within a short timeframe. Over the course of several decades and as a result of ongoing digitization efforts, there has been a significant and rapid increase in the amount of data available. Consequently, there is a growing need to extract this data and gain a deeper comprehension of many topics. These

subjects pertain to distinct domains from which this data is derived. Contemporary organisations rely extensively on this data to operate their algorithms and generate significant outcomes. This accumulated knowledge possesses immense power. Particularly when an unlimited quantity of data is being generated continuously as we exist. Despite numerous advancements in AI technology, dominating the field of natural language remains an enormous challenge. Machine Learning (ML) approaches offer exceptional accuracy, but they necessitate specific requirements in terms of data accessibility, computational capacity, and training time for the models. Typically, the data needed is labelled data, which is challenging to get in real-world scenarios.

The focus of analysis is the sentiment expressed in consumer reviews posted online. Ecommerce websites offer users a forum to express their ideas through product reviews of items they have purchased. The reviews are accompanied with an overall rating, which is a numerical representation of the general user opinion towards the product. For instance, a rating that is less than or equal to 3 might be categorised as either bad or moderate, whereas ratings above 3 are deemed good. The rating linked to the review aids in comprehending the user's viewpoint and instructs the model in predicting and assigning scores to new and unobserved data.

The objective of this thesis is to investigate the historical progression and advancement of methodologies used in sentiment analysis. An extensive comparative analysis of machine learning methods will be conducted, followed by the application of these methodologies. The outcomes will be compared using several assessment metrics such as accuracy, precision, F1 score, and other relevant measures. Moreover, there will be an endeavour to comprehend feature extraction and evaluate its efficacy. The dataset utilised for this research is from Amazon (Amazon product customer reviews). In addition, this paper will examine the ongoing challenges encountered when utilising deep learning techniques for sentiment analysis and propose potential solutions to address them.

1.3 Outline of Thesis

This thesis will commence with exploring the existing research work on sentiment analysis. These studies primarily investigate algorithms, approaches, and models such as Support Vector Machines (SVM), Long Short-Term Memory (LSTM), and Bidirectional

Encoder Representations from Transformers (BERT), along with the dataset used and the evaluation metrics applied. Focusing on the historical background and the evolution of Natural Language Processing (NLP). To lay the foundation of this thesis, a thorough examination and comprehension of machine learning models and their development will be conducted. The literature review will analyse and assess the existing research in this topic and evaluate their findings. Moreover, it will assist the research in addressing and overcoming the obstacles encountered by other researchers.

After conducting a review of the existing literature, the theoretical section provides a foundation for comprehending sentiment analysis. It explores the development of sentiment analysis and several approaches to analysing the emotions expressed in text data produced by ordinary consumers. In addition, a comprehensive examination is conducted on several machine learning techniques used for sentiment analysis. Each method possesses a distinct approach to processing textual material, varying in levels of granularity. Following that, this research will examine assessment measures that will facilitate a comparison of the final outcome. However, it is essential to understand the functionality of each of these metrics in helping the research prior to that. Having a solid grasp of this theory will provide a strong foundation for implementing the subsequent approaches.

The methodology section will include the pre-processing of the dataset utilised in this thesis to ensure that it is in a format that is suitable for machine inputs. Preprocessing will begin with an exploratory analysis of the dataset, which will be followed by cleaning the dataset to eliminate null values, duplicates, and other anomalies that may have a negative impact on the modelling phase.

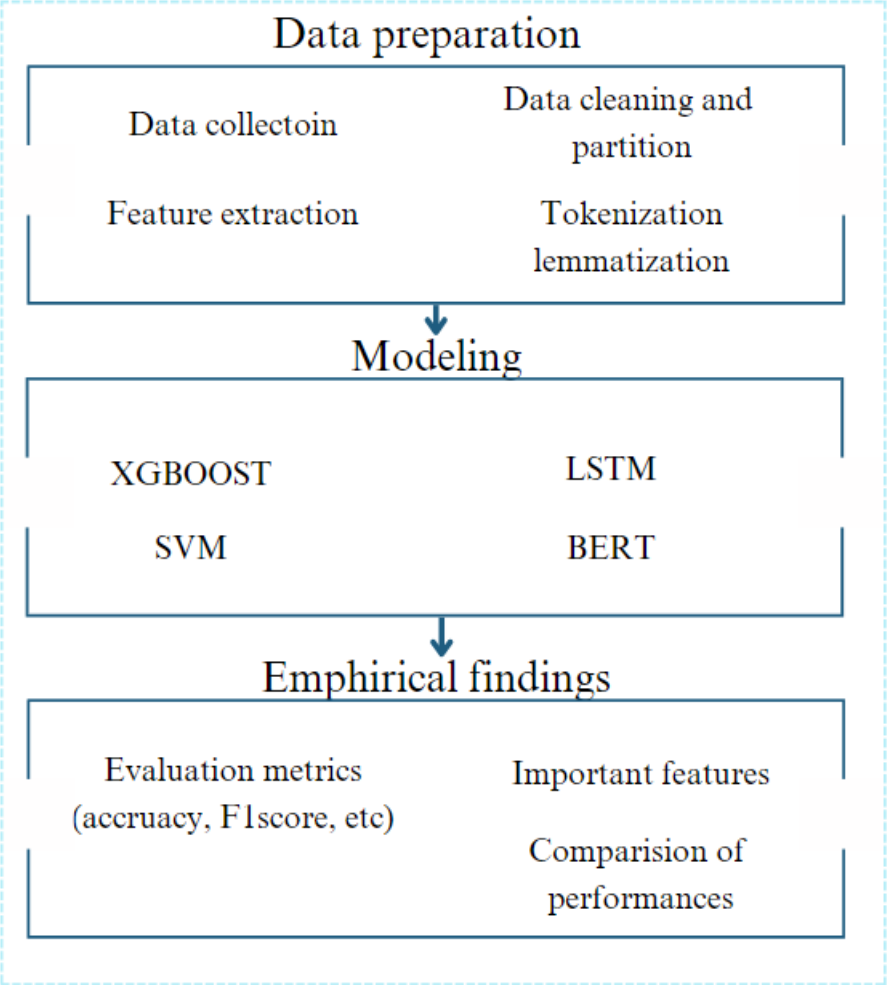

Figure 1. *Methodology Representation*

The experiment will involve running XGBOOST, SVM, LSTM, and BERT on the input data. Therefore, this leads to a thorough analysis to determine which study demonstrates greater efficiency and accuracy. Ultimately, the metrics obtained will assist the research in evaluating the efficiency of selected machine learning models in determining sentiment polarity.

The objective of this thesis is not to create a brand new model, nor is it the aim of this thesis to develop a model that achieves 100% accuracy in recognizing the polarity of hidden elements in customer reviews. Instead, can all of the hidden information in the customer reviews be accurately captured? This thesis aims to present a comprehensive analysis by delivering reliable results for real-world applications.

CHAPTER 2

LITERATURE REVIEW

The literature review for this thesis can be divided into six sections: an introduction to sentiment analysis, methods for sentiment analysis, the application of deep learning models in sentiment analysis, a review of relevant research comparing machine learning models, and finally, the challenges encountered in sentiment analysis. The fundamental analysis part will primarily address reasons for the necessity of sentiment analysis and its role in comprehending various methodologies employed in sentiment analysis. The section on traditional approaches will explore the techniques that were in use before the development of deep learning. Sentiment analysis with machine learning involves utilising machine learning techniques to get insights into human emotions from textual data. The final section will discuss the existing research work that has been conducted on comparing the results of various machine learning models. The findings and the theory will be examined and elaborated.

2.1 Overview of Sentiment analysis

The Sentiment analysis is a difficult undertaking that involves recognising and extracting opinions, emotions, and attitudes from textual data. The first paper explores different techniques employed in sentiment analysis, such as lexicon-based methods, machine learning-based methods, and hybrid methods. It also delves into the topic of subjectivity analysis, which explores the distinction between subjective and objective tests. Understanding and analysing sentiment can be a challenging task due to the diverse ways in which people express their emotions. Dealing with unlabeled data, navigating the complexities of sarcasm and irony, and working with specialised languages are a few of the obstacles that come with sentiment analysis. It is important for the sentiment analysis models to have the ability to identify and understand various contexts. Often, a word or phrase can have different interpretations based on the context it is used in (Liu).

In their paper, Pang and Lee (2008) provide a comprehensive overview of opinion mining. They employ various machine learning techniques, such as supervised and unsupervised learning, to classify sentiment. In addition, the study delves into the difficulties posed by irony and subjectivity in sentiment analysis. This opinion classification is crucial for product-based brands. They can use it for their marketing research, addressing customer issues, and further consolidating their brand's position.

In a study conducted by Jemimah Ojima Abah (2021), sentiment analysis was performed on reviews of electronic products from Amazon. The author employed a CNN and LSTM model for sentiment classification. At first, a dataset with an imbalance was utilised for sentiment analysis, leading to biassed outcomes and overfitting. Consequently, to eliminate the bias, an up sampling was carried out for the minority class. In addition, the author found that using an imbalanced dataset for training the model resulted in a significant rate of type 1 errors, specifically false positives. However, when models were trained using the unsampled data, there was a significant issue with the type 2 error rate, resulting in a high number of false negatives. Lastly, after conducting hyperparameter tuning, it was noted that the LSTM model demonstrated improvements, while the CNN model performance remained the same.

Dave et al. (2003) examines the subject of sentiment analysis and opinion mining, which can enhance data-driven decision-making processes and facilitate comprehension of client preferences. The authors talk about the difficulties encountered in the research, such as inconsistent rating and short review lengths. Ambivalence was one of the challenges where reviewers used negative words but ultimately tend to express their satisfaction. Therefore, conventional methods such as Support Vector Machines (SVM) encounter challenges when attempting to achieve detailed classifications.

In Jonathon Read's (2005) research titled "Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification," the author investigates the use of emoticons as a beneficial asset for sentiment analysis. The study examines the potential benefits of integrating emoticons to decrease reliance on manually labelled data and enhance the effectiveness of machine learning methods for sentiment analysis. The paper commences by emphasising the difficulties associated with sentiment classification, such as the limited availability of labelled data and the necessity for efficient feature selection. Next, it presents the concept of utilising emoticons, commonly employed in text to convey

emotions and feelings, as a valuable resource for sentiment research. The study conducted by Read explores the application of machine learning techniques to sentiment classification problems. The study investigates the impact of include emoticons as features in the classification process. The author conducts a comparative analysis of various classifiers and assesses the influence of incorporating emoticons on the accuracy of sentiment categorization. The experimental results indicate that the inclusion of emoticons as features can improve the effectiveness of sentiment classification models, especially when there is a lack of labelled training data. The paper continues by examining the impact of these findings and emphasising the potential of emoticons as a beneficial tool for sentiment research tasks (Read, 2005).

Tripathy and Rath (2017) utilised three datasets that have distinct JSON formats in their study. The data lacked labels, and manual labelling was not a viable choice. The data was pre-processed and active learners were utilised for labelling. The rating system employed consisted of a 5-star scale, where 5 stars indicated the most positive review, 3 stars represented a neutral evaluation, and 1 or 0 stars denoted the most negative assessment. The data pre-processing included the process of tokenizing the data, eliminating stop words, and performing Part of Speech tagging. The feature extraction technique included the utilisation of Bag of Words, Term Frequency–Inverse Document Frequency (TF-IDF), and Chi-Square. The research involved the utilisation of many machine learning algorithms, including Naïve Bayesian, Support Vector Machine Classifier (SVC), Stochastic Gradient Descent (SGD), Linear Regression (LR), Random Forest, and Decision Tree. The investigation revealed that the Support Vector Machine Classifier had an accuracy rate of 94.02%.

2.2 Techniques of sentiment analysis

Supervised learning is a machine learning technique in which the algorithm is trained using a labelled dataset, where each input point is associated with a pre-assigned output value. For sentiment analysis of Twitter data, the algorithm is trained using a dataset of tweets that have been manually categorised as positive, negative, or neutral. The algorithm subsequently learns the ability to identify patterns within the data linked to each sentiment label, and use these patterns to categorise incoming tweets as positive, negative, or neutral. Commonly employed supervised learning methods for sentiment analysis include logistic regression, support vector machines, naive Bayes, and decision trees. In sentiment analysis,

recent advancements in deep learning have demonstrated promising outcomes, particularly with the utilisation of recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Feature engineering approaches can be utilised to extract pertinent information from twitter text, hence enhancing the precision of sentiment analysis (Hochreiter and Schmidhuber, 1997).

Unsupervised learning involves an algorithm that does not receive labelled input, but instead must independently identify patterns and structures within the data. This can be advantageous in situations when there is a limited availability or high cost associated with obtaining labelled data. Unsupervised learning algorithms may exhibit inferior performance compared to supervised learning algorithms in sentiment analysis due to their lack of training on labelled data containing known sentiment values. The author highlights that certain unsupervised learning methods, like as clustering and topic modelling, and might be employed for sentiment analysis by categorising tweets according to similarities in their content or subject matter. Nevertheless, the precision of these methods can be influenced by the data's quality and the choice of suitable features for analysis. In summary, the author proposes that supervised learning algorithms are generally superior for sentiment analysis of Twitter data, although unsupervised learning techniques may be worth investigating in specific circumstances.

Various methodologies exist for conducting sentiment analysis. Kaur and Sidhu, (2018) examine these methods in their research article. The study explores the utilisation of sentiment analysis in many sectors of the industry, including social media, marketing, and customer services. The authors thereafter present a summary and constraints of certain traditional methods, including lexicon-based techniques and machine learning-based approaches. Moreover, they conclude the discussion by examining the deep learning methodologies that have demonstrated encouraging outcomes in sentiment analysis due to their capacity to automatically learn characteristics from unprocessed data.

The study discovered that the suggested Recurrent Neural Network (RNN) model had superior results compared to alternative machine learning models like Naive Bayes and Support Vector Machines (SVM). The RNN model achieved an accuracy over 90% in sentiment categorization of Amazon product evaluations. The study also highlighted the significance of word embeddings and hyperparameter tuning in enhancing the performance of the RNN model (Iqbal et al., 2022).

There are three different analytical approaches for conducting sentiment analysis:

Lexicon-based approach

A lexicon is a compilation of words and their corresponding definitions. Within this framework, the lexicon consists of words that are linked to a singular polarity. It has the potential to be either positive or negative, as stated by Hota et al. (2021). The lexicon-based approach in sentiment analysis utilizes a pre-prepared sentiment lexicon to assess the polarity of a document. As an expert in analysing systems, one approach involves assigning a positive polarity score to words like "happy" and a negative polarity score to words like "sad" in a lexicon-based system. When examining a sentence that includes both words, the lexicon-based approach considers the overall sentiment conveyed by the words in order to calculate the sentiment score of the sentence.

A widely-used formula for calculating the sentiment score (StSc). This formula can also be applied to a dictionary-based approach, as mentioned by Aline Bessa, (2022).

$$StSc = \frac{number\ of\ positive\ words - number\ of\ negative\ words}{total\ number\ of\ words}$$

Equation 1. Dictionary-based method used to determine the sentiment score.

The lexicon-based approach can be categorised into two main groups: dictionary-based and corpus-based.

a. Dictionary-based

In the dictionary-based approach, pre-defined dictionaries contain lists of words and their corresponding polarity scores. As an example, this approach to sentiment analysis could utilise a lexicon that includes a collection of positive words, such as "good", "great", and "beautiful", as well as a set of negative words, like "horrible", "bad", and "unhappy". Every word in the lexicon is assigned a value that represents its polarity, ranging from -1 to 1.

This process utilises the polarity information found in the dictionary. When examining text using the dictionary-based method, the process entails calculating the overall sentiment of the text. For instance, if the text includes several words with positive connotations and only one word with a negative connotation, it is given a positive sentiment score. In the same way, when the number of negatives outweighs the positives, a negative sentiment score is given.

In general, the dictionary-based approach is a straightforward and efficient method for conducting sentiment analysis, particularly when working with smaller datasets. This approach has its limitations when it comes to capturing the complex nature of language, and it may not be as advanced as some of the machine learning methods. Given their dependence on pre-defined lexicons, dictionary-based approaches may encounter challenges when it comes to capturing the sentiment of content that is specific or includes humour or sarcasm.

b. Corpus-based

When analysing the Corpus-based approach, a significant amount of text is examined to identify the shared patterns and connections between words. For instance, utilising a corpus of product reviews to construct a sentiment lexicon through the analysis of the language employed in the reviews. This approach offers a distinct advantage compared to the dictionary-based method, as it has the capability to acquire orientations specific to a particular domain.

This approach necessitates a greater amount of computational power to analyse extensive collections of data. Another possibility is to gather the categorization of emotions for a vast collection of data. While this approach requires the creation of a lexicon dictionary, it has been proven to be effective in sentiment analysis, as shown by Mohammad and Turney (2013).

Machine learning-based approach

There are three distinct methodologies to machine learning that can be used for sentiment analysis. Supervised, unsupervised methods that have been previously discussed, and also and semi-supervised learning method. Semi-supervised learning combines elements from both supervised and unsupervised approaches.

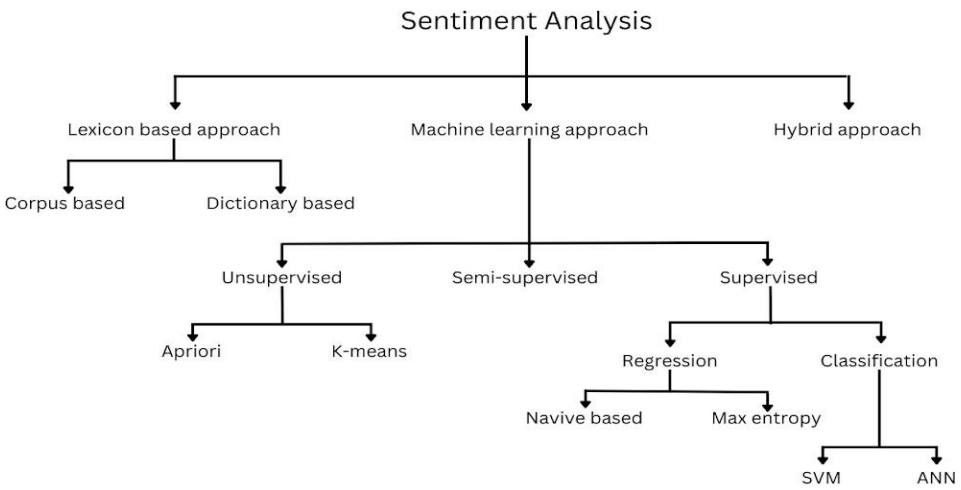

Figure 2. *Outline of Sentiment Analysis Methods*

2.3 Overview of Feature Extraction

Feature extraction involves converting textual data into numerical representations, specifically through text vectorization. There are different techniques available for feature extraction using genism and sklearn, such as countvectorizer, TF-IDF vectorizer, and Word embeddings. In this paper, we will implement the TF IDF feature extraction method to extract features from the textual data. Our modelling process benefits greatly from incorporating semantic information from the text.

CountVectorizer is a versatile and widely applicable method for extracting features. This approach is known as the 'Bag of Words' method, as it solely focuses on the frequency of words in a document. Grammatical conventions and word order are completely disregarded. The arrangement or structure of the document's words is not considered (turbolab, 2021).

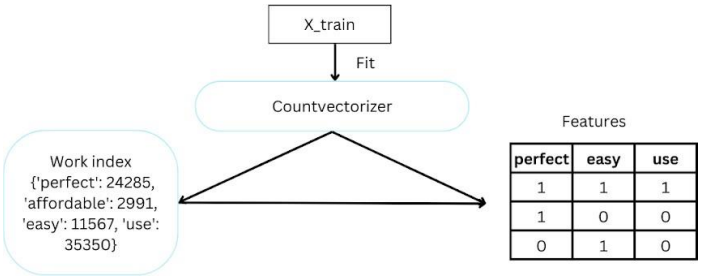

Figure 3. *Countvectorizer feature extraction (turbolab, 2021)*

The TF-IDF Vectorizer calculates the weight for each word in the document by taking into account two factors: term frequency and inverse document frequency. Term Frequency (TF) quantifies the occurrence of a term within a document. Highlighting the significance of that particular term in the document. This implies that the frequency of the term directly correlates with its significance. Term Frequency can be calculated using various formulas, including raw term frequency, binary representation, or logarithmic scaling. This approach is comparable to the Bag of Words approach. Inverse Document Frequency (IDF) quantifies the significance of a term throughout the entire corpus. It prioritises terms or words that have a lower frequency of occurrence in the document. The calculation of Inverse Document Frequency involves taking the logarithm of the ratio between the total number of documents and the number of documents that contain the specific term (Haque et al., 2018).

$$w_{x,y} = tf_{x,y} x \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Equation 2. TF-IDF (turbolab, 2021)

In addition, the values are combined to compute the TF-IDF weight for each word in the document. The weight of a term within the corpus and the document is directly proportional to its importance, with higher weights indicating greater significance.

In the experiment conducted by Liu et al. (2018), the author utilised TF IDF to extract semantic features from the text data. For the purpose of extracting features and clustering, a dataset containing two main topics was utilised, with 2500 instances for each topic. TF-IDF and Word2vec were utilised to create a virtual word vector. In addition, the classification was performed using the K-nearest neighbour (KNN) algorithm. Using TF-IDF, the classification accuracy for topic 1 was an impressive 98%, while for topic 2 it reached 82%. On the other hand, the method suggested utilising Word2vec experienced a significant boost for topic 101% and topic 2 at 82.2%. After careful analysis, it was determined that there was no notable enhancement in the accuracy.

Word embeddings are commonly utilised in Natural Language Processing to represent words as vectors in a high-dimensional space. The design of these vectors ensures that words with similar meanings are positioned close to each other. It assists in capturing both the semantic relationship and the context information. With the expertise of a data scientist, machine learning models can gain a comprehensive understanding of words in a highly effective manner.

Word2vec, introduced by Mikolov et al. in 2013, is the go-to algorithm for learning Word embeddings from a vast amount of data. The algorithm acquires these embeddings through the training of a substantial volume of textual data on neural networks (Mikolov et al., 2013). There are two main architectures used in Word2Vec. CBOW and Skip-gram are two popular models used in natural language processing. In Continuous Bag-of-Words, the model makes predictions about the current word based on the surrounding context. However, in Skip-gram, the model is designed to predict the context words based on a given target word. Both architectures prioritise the acquisition of word embedding that contain valuable insights and capture the semantic relationships between words.

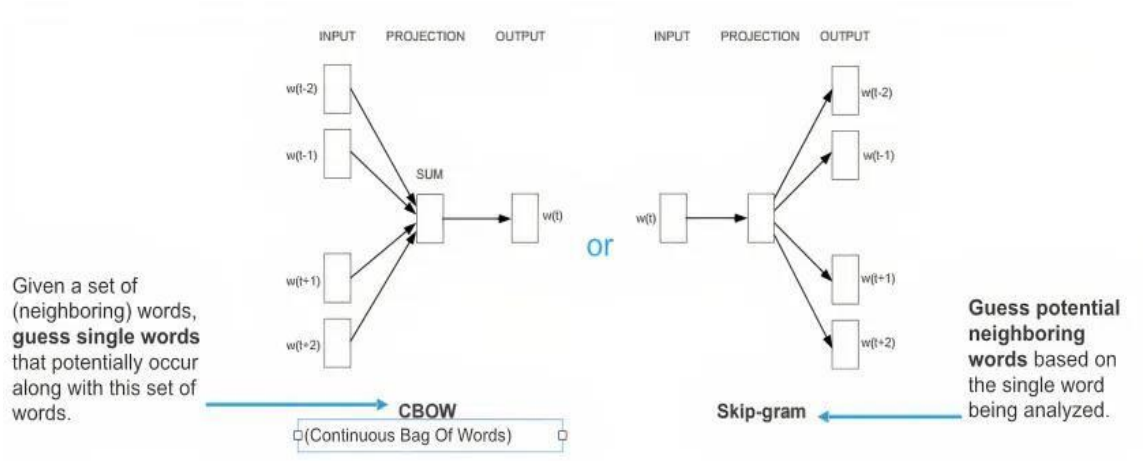

Figure 4. *Feature Extraction by Word2Vec (turbolab, 2021)*

In a study conducted by Read, (2005), it was demonstrated that the process of feature extraction is influenced by the specific feature being analysed. When using the bag of features approach, it's important to note that a classifier trained on a movie review corpus may not necessarily perform as well on other datasets, such as automobile or food reviews. In a study conducted by Turney (2002), it was found that the term "unigram unpredictable"

had a positive sentiment when applied to the movie dataset. However, when used with the automobile dataset, it yielded different results. It's important to note that certain feature extraction methods may be influenced by the specific features they are designed for, and may not perform as well when applied to different types of data.

2.4 Utilizing Deep learning models for sentiment analysis

This section will explore the previous research on the models that will be utilised in the research of this thesis. These algorithms are part of the deep learning family of machine learning models. This will provide insights into the expectations surrounding deep learning models and any potential challenges encountered by researchers.

Convolutional neural network

"Convolutional Neural Networks for Sentence Classification" by Kim, (2014) is a highly regarded paper that delves into the use of Convolutional Neural Networks (CNNs) for sentence classification. This paper highlights the effectiveness of CNN for sentiment analysis. The dataset utilised in this paper is sourced from IMDb movie reviews and Stanford Sentiment TreeBank. The author utilises the CNN architecture for sentence classification. The model utilises a convolutional layer and max-pooling to extract local features from various n-grams (word sequences of length n) in the sentences. Various filters with different window sizes are used to capture features at different scales. The resulting feature maps are then passed through fully connected layers to perform classification. An array of hyperparameters, including filter sizes, pooling strategies, and activation functions, were employed to fine-tune the performance of the CNN model during the experiment. In addition, they compare the results of their CNN model with various traditional methods and showcase the exceptional performance of CNNs in sentence classification tasks. Ultimately, the experiment showcased in the paper highlights the impressive performance of CNN in sentence classification tasks, surpassing traditional methods. The CNN model presented in this paper achieved an impressive accuracy of 88.89% when applied to the IMDb movie reviews dataset.

Recurrent neural network

Neural networks have been utilised since the 1990s. An artificial neural network closely resembles a biological neural network. Recurrent neural networks are specifically designed to acquire knowledge from patterns through time. A Recurrent Neural Network (RNN) is a type of neural network that incorporates a feedback mechanism (Fausett, 1994).

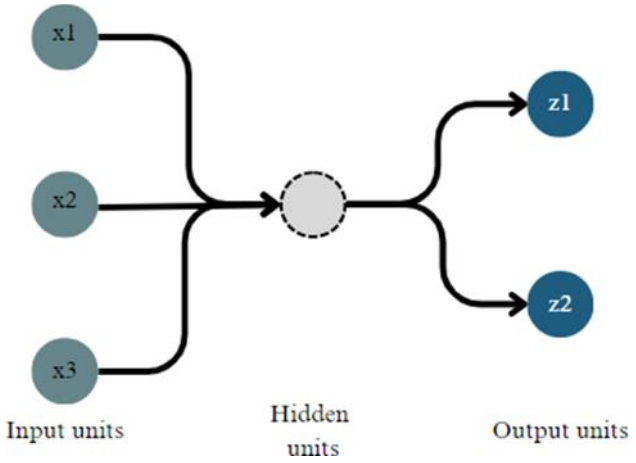

Figure 5. *Basic neural network*

Recurrent neural networks (RNN) are so named because they perform the same operation for each element in a sequence, and the outcome is always dependent on prior computations (Javaid Nabi, 2019).

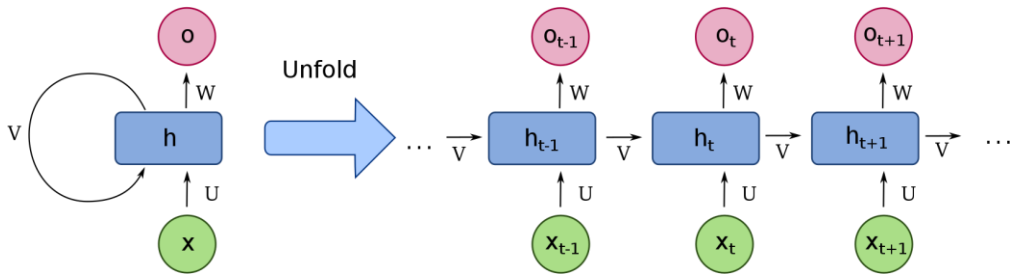

Figure 6. *Recurrent Neural Network (Wikipedia, 2023)*

Check For calculating the current state:

$\mathbf{h_t = f(h_{t-1}, x_t)}$

Where,

h_t= **current state**

h_{t-1}= **previous state**

x_t= **input state**

To apply the activation tanh function,

$\mathbf{h_t = \tanh (w_{hh}h_{t-1} + w_{xh}x_t)}$

There exist various categories of Recurrent Neural Networks:

- The Vanilla RNN is the fundamental model of the Recurrent Neural Network (RNN). It is often referred to as the Elman network, named after Jeffrey Elman, who first proposed the concept of the Simple Recurrent Neural Network (SRNN).
- The Long Short-Term Memory (LSTM) model has a gating mechanism and memory unit (cell).
- The Gated Recurrent Unit (GRU) is a type of recurrent neural network (RNN) that is similar to the Long Short-Term Memory (LSTM) model but offers improved computational efficiency.
- Bidirectional RNN, useful for situations that demand a thorough comprehension. They encompass both backward and forward directions, thereby preserving both historical and prospective information.
- Hierarchical RNN: Employ many layers of RNN to handle hierarchical data, such as document categorization.

Furthermore, according to Chung et al. (2014), the effective applications have been achieved using several iterations of RNN rather than the basic form. The Long Short-Term Memory (LSTM) model, which was introduced by (Chung et al., in 2014) demonstrates superior performance compared to the Vanilla RNN. Similarly, the advanced recurrent network GRU has obtained comparable outcomes (Cho et al., 2014). (Werbos, 1990) highlights the significance of back propagation across time as a solution to the issue of vanishing gradient in recurrent neural networks (RNNs). The author explores a complex forward feeding neural network that will progressively unfold the RNN over time.

Transformer based model

Vaswani et al., (2017) introduced a model based on transformers. The transformer based model demonstrated superior performance compared to the convolutional and recurrent models. The attention mechanism is a crucial component in sequence understanding, as it helps the model determine the significance of each term or word. Just like a human, one would naturally try to focus on the important part of the sentence. A transformer is made up of an encoder and a decoder.

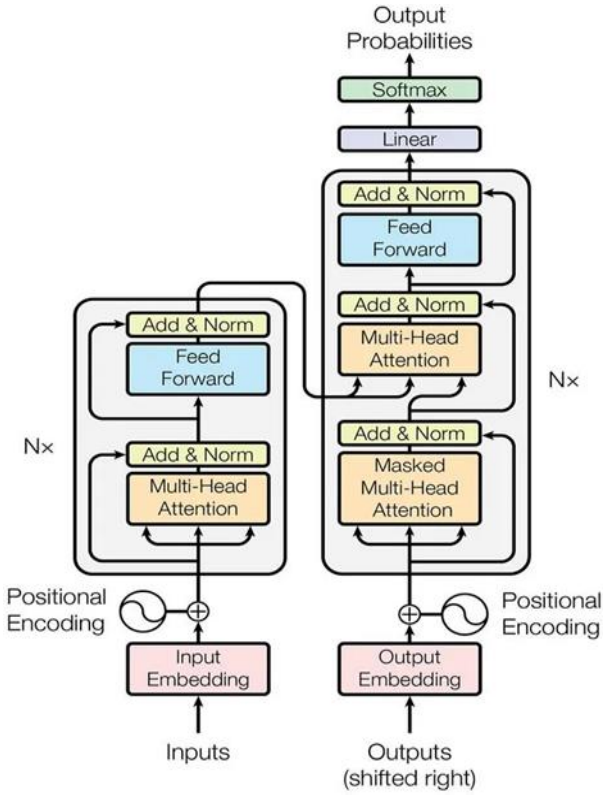

Figure 7. *Transformer detailed architecture (Vaswani et al., 2017)*

This thesis will also explore the use of a transformer-based model known as BERT (Bidirectional Encoder Representations from Transformers) to determine if it surpasses the performance of other selected models.

2.5 Overview of related research studies comparing machine learning models

Multiple studies have been conducted to determine the machine learning algorithm that can produce the most optimal outcome. This process is depended upon several elements. Firstly, the key factors to consider are the dataset, the programming language, and the accessibility of resources. Tan et al. (2022) conducted a comparative analysis of game reviews utilising XGBoost, SVC, Multinomial Naïve Bayes, and Multi-layer Perceptron Classifier. These models belong to a family of conventional machine learning models and are specifically supervised learning models. It was noted that on a dataset with imbalanced and oversampled data, the Support Vector Classifier (SVC) outperformed other models with accuracies of 72% and 82% respectively. Moreover, following the optimisation of hyperparameters, the model successfully attained an accuracy rate of 89%. This demonstrates that hyperparameter adjustment is crucial for determining the optimal parameters for a specific model.

In a study conducted by SEPIDEH PAKNEJAD, (2018), the author employed TF-IDF feature extraction methods on a dataset of around 200,000 Amazon reviews. The author employed Support Vector Machines (SVM) and Naive Bayes algorithms for the purpose of classification. Based on the results, SVM demonstrated superior performance compared to Naive Bayes, achieving an impressive accuracy score of 93% on the reviews. On the other hand, the models showed stronger performance when it came to summarising reviews. It's worth noting that the lack of data could pose a potential limitation.

In their study, Srinivas et al. (2021) presented the findings of their experiment, which examined the performance of an LSTM, CNN, and a single layer neural network. The authors used 1.6 million tweets for text data and categorised them into positive and negative groups. Based on the observations, it was found that LSTM had the highest accuracy rate of 87%. It is evident that RNN based models have the potential to outperform other models

when it comes to sentiment analysis classification.

You can find a comparison of CNN, BERT, and a combination of both in the experiment conducted by Li et al., (2021). The author utilised a large dataset of Weibo data, which had been properly labelled. The author experimented with various combinations of BERT, CNN, and LSTM layers, but for the purpose of this thesis, the focus will solely be on the results obtained from BERT and CNN. A tokenizer was used to preprocess the data. The accuracy of BERT was significantly higher at 84.4% compared to CNN, which achieved 73.5%. Highlighting the potential for improved performance with a transformer-based model.

2.6 Challenges encountered in Sentiment analysis

This section will address the difficulties associated with sentiment analysis. These challenges vary in terms of the techniques employed and the accessibility of data. The following points outline the difficulties encountered when applying sentiment analysis.

Handling irony and sarcasm

Sentiment analysis frequently faces challenges in identifying the sentiment conveyed by sarcastic or ironic phrases. Creating methods to manage such occurrences would be a significant contribution to the field. Sarcasm involves the use of satirical remarks to emotionally hurt or mock someone. It is accomplished by utilising information that is entirely contradictory to the given context (polar opposed). Despite being one of the more difficult aspects of natural language processing, it is becoming more and more popular because of its benefits (Eke et al., 2020).

Multilingual sentiment analysis

Sentiment analysis often prioritises the examination of data written in the English language. Nevertheless, there is an increasing demand for sentiment analysis in more languages, specifically within the realm of social media. Creating techniques to do sentiment analysis in several languages would be a great addition to the profession. For example, the English language has regional variances in countries such as India, Australia, USA, and UK. In Indian English, the term "thong" refers to undergarments, similar to British English.

However, in Australian English, it refers to flip flops. This disparity alone can have significant ramifications for the processing of textual material. This can result in redundancies and challenges in categorization. An illustrative instance is the variation in the spellings of "color" and "colour" (Wankhade et al., 2022).

Sentiment analysis for specific domains

Sentiment analysis is typically trained using broad datasets, but its effectiveness might be compromised when applied to specific sectors such as healthcare, politics, and finance. It would be beneficial to the discipline to create sentiment analysis models that are domain-specific.

Ambiguity

Dealing with ambiguity poses a major obstacle in sentiment analysis. Understanding the various meanings of words can be a challenging task, especially when it comes to determining the intended sentiment.

Detecting Emotion

While sentiment analysis usually concentrates around categorising sentiment as positive, negative, or neutral, it's important to recognise that emotions are much more intricate. For instance, a statement may have a negative connotation, yet the emotions conveyed could encompass anger, sadness, or frustration.

2.7 Summary

The literature review section provides a comprehensive understanding by providing an overview of sentiment analysis. There are various techniques used in sentiment analysis. Since the development of the internet and the increasing freedom for people to express their emotions, there has been a growing demand for improved methods to assist businesses in driving change. Sentiment analysis has emerged as a valuable tool in this regard. In addition, we explored techniques for sentiment analysis to gain a deeper understanding of machine learning and deep learning approaches. The section provides a detailed explanation of the work conducted in the field, where various machine learning models were employed to

analyse a specific corpus and determine the most effective ones. The more data you have, the more accurate and insightful the results will be. However, there are certain challenges that have been discussed above.

This table provides an overview of the findings in the literature review section.

Table 1. *Different machine learning models employed in this research*

Approach	Characteristics	pros	Cons
Lexicon based approach	A straightforward rule-basedapproach. Assigning negative or positivepolarity to keyword in order to calculate the result.	Easy to use Interpretable	Does not consider sentence structure,thus a constraint on contextual understanding.
CNN	Able to learn hierarchical representations and patternsfrom text using convolutionlayer	Captures local features. Can handletextual data with variable lengths.	Contextual understanding is limited as it does not understand theoverall context. Fixed sized input is required for CNN, thus unableto handle varyinglength text effectively.
LSTM	Versatile in data handling. Utilizes Back propagation through time. Gating mechanism andmemory cells.	Captures dependencies between the data	Problem of vanishing gradient. Sensitive to hyper parameters.
BERT	A transformer-based model. Utilizes mechanism of self-attention and generate wordembeddings.	Transfer learning. Bidirectional attention. Contextual word embedding generation	Higher trainingand memory requirement. Pre-trained models might not be precise (fine- tuned) for domain specific tasks
SVM	Utilizes a hyperplane toseparate classes. A simple linear binary classification technique.	Easy to interpret. Memory efficient.	Limited to binary classification. Sensitive to outliers and noisydata.

After conducting extensive research and gaining a deep understanding of sentiment analysis, the models selected for this research are CNN (convolutional Neural Network), LSTM (Long Short Term Memory), and BERT (Bidirectional Encoder Representations from Transformers). These models fall under the category of deep learning models. In order to facilitate comparison, this research will also incorporate a traditional SVM (Support Vector Machine). This aids in gaining a deeper understanding of the distinctions associated with each approach.

2.8 State of Art

In this section, we present a brief state of the art in sentiment analysis applied to e-commerce reviews, taking into consideration the literature more closely related to this topic and the most recent works. The specific papers which have been considered to be most relevant to this thesis in methods and goals are the following: A short description of each paper and state the research contribution to the field in regards to innovation to the existing knowledge.

1. A Comprehensive Study on Sentiment Analysis for Amazon Product Reviews Using Machine Learning Algorithms by Ni et al. (2019)

Summary: In the following paper, the author aims to pursue various avenues of machine learning to conduct sentiment analysis for the reviews of the Amazon products. The works, as stated earlier, contrast typical machine learning algorithms like the SVM, Naïve Bayes, and Decision Trees against deep learning algorithms such as LSTM and CNN. Concerning feature extractor, we compare between the Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec techniques to determine which type would yield the most improved performances.

Contribution: The paper specifically explains how the new generation of deep learning models outperform the previous models of the machine learning in dealing with text data of large scale. It also calls for the need to select the most suitable feature extraction method for enhanced sentiment classification.

2. Devlin et al. (2019) present the BERT model for general, pre-COVID-19 sentiment analysis of Amazon product reviews.

Summary: This paper presents an implementation of BERT, a transformer-based model, for the purpose of classifying the given text's sentiment as positive or negative, on a collection of Amazon product reviews. BERT, in particular, is a bidirectional model, which is helpful in capturing the context for left and right in defining sentiment in textual data. The authors explain how BERT works using the results from a large, real-world dataset of Amazon reviews and establishing its superiority over standard benchmarks such as LSTM and CNN.

Contribution: When it comes to SA, BERT brings a significant breakthrough and successfully breaks a new record. It provides evidence that BERT has a comparatively better performance over others in terms of accuracy and F1 score besides. It demonstrates relevancy and applicability of transformer-based models in natural language processing tasks.

3. Life cycle costing analysis of solar PV systems in Ghana by Arthur et al. (2022)

Summary: This paper is focused on evaluating a number of algorithms using Machine learning strategies for sentiment analysis of e-commerce reviews. To test the models, the authors used balanced and imbalanced dataset for calculations such as SVM, Random Forest, XGBoost. Optimization of the model also comprises hyperparameter tuning in the given research efforts.

Contribution: Overall, the paper offers meaningful and informative information that can be useful for identifying the advantages and drawbacks of various machine learning models for sentiment analysis. It is quite insightful in discussing about the need to maintain balance in data portions and the right hyperparameters that need to be set in order to create good classification.

4. Published in 2022, Iqbal et al. exposes prospect views on Learning Technologies for Sentiment Analysis of Product Reviews.

Summary: From the following methodology, this paper aims at investigating the application of Deep Learning Method using LSTM and CNN techniques on product reviews.

The authors include an experimental evaluation of their models using a large dataset from Amazon reviews and a comparison of LSTM and CNN models. This paper also analyzes the effect of various word embeddings to the model on performance.

Contribution: The analysis proves that the models with the LSTM and CNN algorithms, when properly matched with an appropriate word embedding, can perform successful sentiment analysis. It also gives a comprehensive review on what was done in the area of training deep learning models for sentiment analysis, and some of the difficulties that are likely to be encountered.

5. Title: Sentiment Analysis of E-commerce Product Reviews Using Transformer-Based Models Author(s): Li Dongyun, Ling Zhaohui, Tan Wei, Zhong Xiaosong, Du Zhixin Year: 2021

Summary: This study is specifically concerned with using transformer models, BERT in particular, and products derived from it for the sentiment analysis of e-commerce product reviews. In order to test their efficiency the authors apply these models on a large set of reviews which are labeled and compare it to the previous models which are the traditional models as well as the deep learning models.

Contribution: The paper also reveals that transformer-based models have the ability to provide efficient and real-time sentiment analysis. Using the arguments also support the conclusion that BERT and its variants surpass most of the models in terms of accuracy as well as the computational cost. The transfer learning is also another aspect which has been stressed in the study regarding to the enhancement of the models perceptivity in particular domains.

CHAPTER 3

RESEARCH QUESTIONS

Every day, more people are drawn to the growing business. The buyers then make a huge amount of data. This data can be unstructured when it comes from customer reviews and notes. However, a sentiment analysis can be used to get the customer's mood from this data. Stephanie Chevalier's, (2022) study for Statista shows that the e-commerce market will reach $8.1 trillion by 2026. In this way, it becomes even more important to understand how the customer feels.

Machine learning development is experiencing rapid growth, particularly in the realm of large language models (LLMs). Given the rapid pace of development, there is now an abundance of various versions of a single model that can be implemented. We have a range of machine learning models available, including SVM, XG-Boost, and Random Forest. In addition, there are several deep learning models such as LSTM (Long term short memory) and BERT (Bidirectional Encoder Representations from Transformers) that are particularly effective for sentiment analysis.

RQ1: *What is the comparative performance and effectiveness of various machine learning models for sentiment analysis on e-commerce review data?*

RQ2: *Which feature extraction method is more suitable for sentiment analysis of Amazon reviews, TF-IDF or Word2Vec?*

This research aims to investigate various deep learning models and analyse their outcomes. By examining different machine learning models, this study enhances prediction accuracy for the Amazon customer review dataset.

CHAPTER 4

METHODOLOGY

This section covers various aspects related to this research. Understanding the deep learning models that will be used, in addition to traditional machine learning models, is where it all starts. Understanding how they work with the selected dataset is crucial for establishing a solid foundation. Next, the section will discuss various evaluation metrics necessary for making a final comparison during the empirical findings and determining the performance of the models.

This research process utilised a thorough methodology of Knowledge Discovery in Database (KDD). This methodology consists of specific steps that provide guidance for the research and development process. Ultimately producing the desired outcome.

1. Problem Understanding: This stage requires a thorough comprehension of the problem domain and a precise definition of the specific problem to be addressed. One of the important steps is to identify the project objectives, requirements, and any constraints that should be taken into account.
2. Data Selection: In the data selection stage, we meticulously choose relevant data from a variety of sources, taking into account the specific requirements of the problem at hand. Our main objective is to determine the most suitable scope and level of detail for the data.
3. Data Preprocessing: This stage involves preparing the chosen data for analysis by cleaning and transforming it. Steps such as eliminating duplicates, addressing missing values, and normalising the data are carried out to guarantee its suitability and quality for subsequent analysis.
4. Data Transformation: In this step, the data undergoes a process of transformation and enhancement to facilitate the identification of significant patterns. Various techniques, such as aggregating data, reducing its dimensionality, and creating new features are used to manipulate and transform data into a suitable format for analysis.

5. Data Mining: This stage focuses on applying a range of algorithms and techniques to extract patterns and valuable knowledge from the transformed data. Statistical analysis, machine learning, classification, clustering, and association rule mining are some of the techniques employed to reveal concealed insights.
6. Evaluation: In this stage, the patterns and knowledge that have been discovered are evaluated to determine their quality, significance, and usefulness. Various evaluation metrics and validation techniques are used to assess the effectiveness and reliability of the knowledge that has been discovered.
7. Knowledge Presentation: The last step is all about conveying the knowledge that has been found to the relevant stakeholders in a clear and effective manner. One way to effectively communicate the insights and findings is by utilizing visualization techniques, reports, or interactive interfaces. These methods help present the information in a clear and understandable manner.

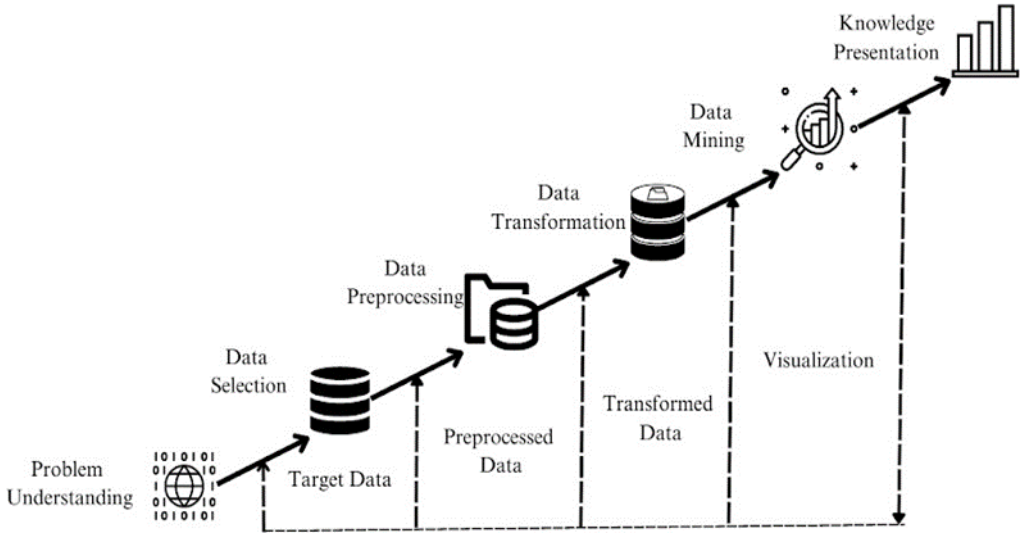

Figure 8. *Knowledge Discovery in Database Process*

4.1 Overview of algorithms used in this research

This section will discuss the algorithms chosen for the research. At first, we will focus on understanding and designing the models. This process will assist in establishing a strong foundation for understanding the operation and expectations of these models. In addition,

we will delve into specific examples that demonstrate how the model handles textual data and the various features it recognises in order to facilitate the sentiment analysis process. We will delve into the intricacies of a traditional machine learning model SVM, as well as deep learning models such as CNN, LSTM, and BERT.

4.1.1 SVM

SVM is a member of the supervised machine learning algorithms family. This approach utilises a hyperplane to separate the data for classification purposes. When separating the two classes, this hyperplane takes into account factors like the variance between the classes and the variance within each class. These criteria are maximised and minimised, as explained by Thompson et al., (1974).

In their work, Cortes and Vapnik, (1995) describe a hyperplane as a linear function that separates two classes. The author explains that support vectors are necessary for determining the margin of a hyperplane. Based on a study by SEPIDEH PAKNEJAD, (2018), it was found that SVM achieved an accuracy of 93% in binary classification tasks for sentiment analysis. This suggests that SVM has great potential in this area. The researchers used the dataset of Amazon reviews for their study, which is similar to this experiment. Therefore, it is possible to obtain different scores, but it is clear that SVM is a more suitable choice compared to other traditional approaches.

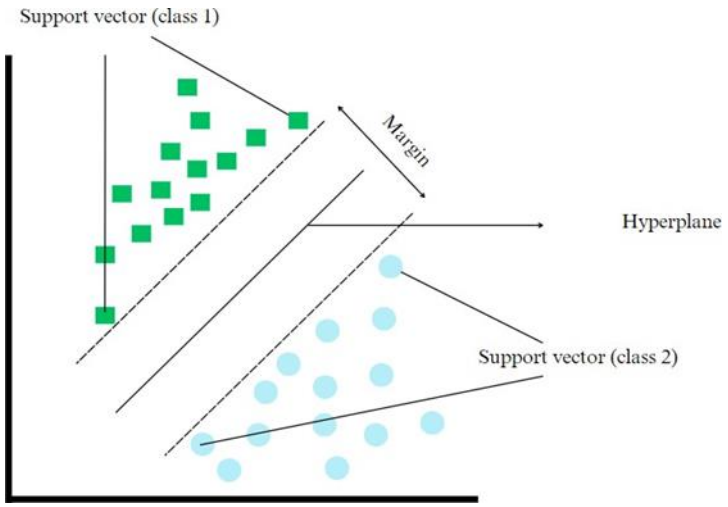

Figure 9. *Representation of SVM Hyperplane*

a. **Hyperplane:** A plane that separates two classes in a given space.
b. **Margin**: A line that runs parallel to the hyperplane on each side. Classifies data into two distinct categories, positive and negative.
c. **Support Vectors:** Points that are closest to the margin, either positive or negative.

4.1.2 SVM

Long short-term memory (LSTM) is a specific sort of recurrent neural network (RNN) designed to solve the issue of the vanishing gradient problem commonly faced by conventional RNNs. The issue of vanishing gradient occurs when back propagation is utilised. The neural network's weight is adjusted using gradient values. Vanishing gradient refers to the phenomenon where the gradient diminishes or becomes smaller as it propagates through time. The contribution from a gradient diminishes significantly when its values approach zero. RNN (Recurrent Neural Networks) encounter difficulties with short-term memory. If the sequence is excessively long, RNN may encounter difficulties in effectively transferring information from earlier phases to subsequent ones. Consequently, when a paragraph of text is being evaluated for predictions, an RNN has the potential to disregard information from the beginning (Phi, 2018). The Long Short-Term Memory (LSTM) model has memory cells, which are utilised to selectively retain data. Therefore, they are able to uphold long-term dependence.

The LSTM architecture includes memory cells by incorporating three types of gates: the input gate, the output gate, and the forget gate. These gates control the transmission of information within the network by selectively altering the memory cell and impacting the output according to the current input and the data stored in the memory cell.

1. Forget Gate

The forget gate, the determination of which information to retain and which to discard is made by this gate. The sigmoid function is applied to the information from the previous hidden state and the current input state of two states. The sigmoid function compresses these values inside the interval of 0 and 1. The value closer to 0 is discarded while the value closer to 1 is retained.

2. Input gate

This gate is crucial for updating the state of the cell. The current input and the prior hidden state are both processed by a sigmoid function. The values are transformed into a range of 0 to 1 in order to determine which values should be kept and rejected. In addition, the current input and previous hidden state are both processed by a hyperbolic tangent (tanh) function. This process applies a transformation to the values within the range of -1 and 1, which aids in the regulation of the network. These two numbers are multiplied together in a plot-wise manner, and the resulting output from the sigmoid function determines which information from the tanh output should be retained.

3. Cellular condition

The cell state is a crucial component of LSTM. The forget vector is initially multiplied element-wise with the cell state. When multiplied by values that are near to 0, this could lead to the cell state losing its values. Next, employing a pointwise addition operation, we modify the cell state by incorporating the most recent values that are considered relevant by the neural network. This is achieved by utilising the output obtained from the input gate (Michael Phi, 2018).

4. Output Gate

Finally, there is an output gate. It determines the values for the subsequent concealed state. The current input and previous hidden state are both subjected to a sigmoid function. The recently produced cell state is transmitted using a hyperbolic tangent function. The result of these two is thereafter subjected to a point wise multiplication. In this context, the sigmoid output determines whether information from the tanh function should be retained as the concealed state for the subsequent phase.

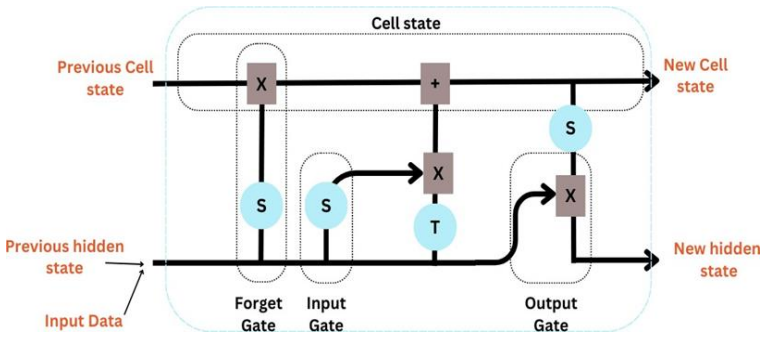

Figure 10. *Architecture if LSTM (Michael Phi, 2018)*

S = Sigmoid (Sigmoid activation)

T = Tanh (Tanh activation)

X = Plotwise multiplication

+ = Plotwise addition

The Sigmoid activation function is equivalent to the tanh activation function, as it is used in Recurrent Neural Networks (RNNs). The sigmoid function compresses the data within the range of 0 and 1. It is important to either erase or update the data. When a number is multiplied by 0, the result is always 0. Similarly, when a number is multiplied by 1, the result remains the same. The output corresponding to the value 0 is considered as 'forgotten', whereas the output corresponding to the value 1 is considered as 'kept'. Consequently, insignificant data is disregarded while significant data is retained. The hyperbolic tangent (tanh) activation function restricts the values to the interval between -1 and 1 (Phi, 2018).

The ability of LSTM's to selectively store, delete, and update information within its memory cell enables it to efficiently record distant connections in sequential data, hence resolving the problem of disappearing gradients. LSTM networks are well-suited for applications that involve sequences, such as speech recognition, time series prediction, and natural language processing. In summary, LSTM is a specific variant of recurrent neural network that utilises memory cells and adaptive gating mechanisms to surpass the limitations of conventional RNNs.

The distinctive architecture of LSTM allows it to effectively capture long-term dependencies in sequential data, making it highly suitable for a wide range of applications (Hochreiter and Schmidhuber, 1997). Simply put, the way LSTM works is that, for instance, a customer intends to purchase a guitar from Amazon. This customer will first need to review the large number of feedback available for the specific guitar in question. When a customer encounters a review that says, "Amazing! The guitar is tuned to perfection. I just got it, and will definitely be spreading the word to more people. Customers tend to remember the highlighted parts of a product review rather than the entire content. Customers often share the same information when asked for their review. That's essentially what LSTM achieves.

4.1.3 CNN

A Convolutional Neural Network (CNN) is a particular kind of deep learning model. This software is specifically designed to efficiently process images or sequences of grid-like data. Additionally, it can also be used to handle vectors that are stacked on top of each other to create an "image". When it comes to NLP, the input is a sentence or document that is represented as a matrix. Each row in the matrix corresponds to a word represented as a vector. These vectors are usually word embeddings, which are condensed representations of words in a lower-dimensional space. CNN has the ability to learn and extract hierarchical representations of features, capturing temporal and spatial patterns. This can be achieved by utilising the convolutional layer. The convolutional layers are essential components of a CNN. These layers are made up of kernels, which are a set of filters that can learn and perform operations on the input data. Every filter has the task of identifying particular patterns or characteristics, such as edges, shapes, or textures (Denny Britz, 2015).

Once the convolutional layers have done their job, the pooling layers step in. Just like a data scientist, the pooling layer efficiently reduces the spatial dimensions by subsampling their input. One commonly used type of pooling is max pooling, which involves performing a max operation. When max pooling is used on a region, it hones in on the most significant feature of that region, while disregarding the rest. It captures the crucial data on whether the feature is present in a sentence. By adopting this approach, it disregards the less significant elements and also enhances computational speed. Max pooling diminishes spatial information, resulting in the loss of precise feature location and prominence. However, it is aware of its presence in that area. Just like a data scientist, the model can easily recognise even the slightest change to a feature, regardless of its position. This assists CNN in managing variations in the data. According to Denny Britz (2015), max pooling ultimately leads to improved performance by helping the network handle variations, reducing computation load, and effectively retaining important features.

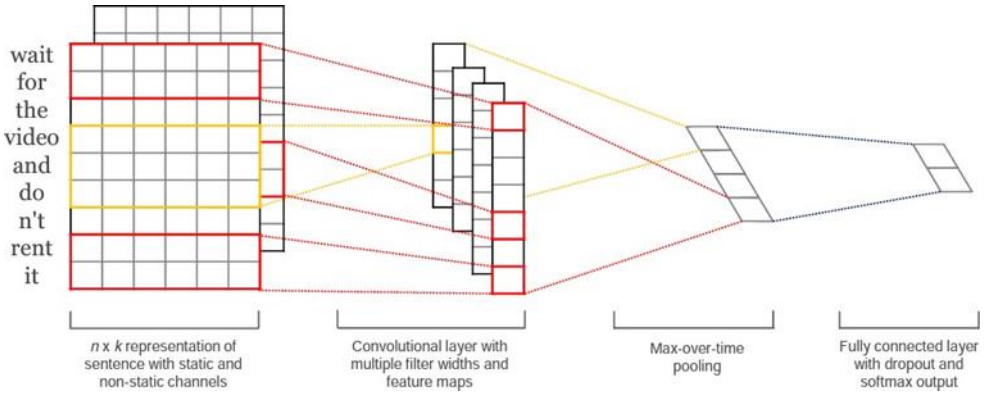

Figure 11. *Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification*

Non-linear relationships can exhibit complicated patterns and dependencies in the data that are beyond the scope of linear transformations. Therefore, in order to effectively capture these, CNN utilises activation functions. The functions bring in a level of non-linearity that enables the network to capture intricate relationships. There are several activation functions that are commonly used, such as ReLU, Sigmoid, and SoftMax.

4.1.4 BERT

BERT is an acronym for Bidirectional Encoder Representations from Transformers, which was published by researchers from Google AI Language. A common approach used by traditional models is to process information either from right to left or left to right. BERT is a combination of both, hence the name bidirectional. The model is trained in both directions at the same time. BERT utilises a specific model known as a Transformer, which has the ability to selectively focus on various sections of the input text. Training BERT bidirectionally allows it to gain a deeper understanding of the context and progression of language, surpassing models that only analyse a single direction (Devlin et al., 2019).

Google AI Language researchers have introduced two strategies for training to address the limitations of the single-directional approach: Masked LM (MLM) and Next Sentence Prediction (NSP). Masked LM is a technique that involves randomly hiding or masking a word in a sentence. The model is then trained to figure out the missing word based on the context of the surrounding sentence. Take the sentence **"The dog is brown and the dog is hungry"** as an example. The model will replace the word "brown" with a mask, resulting in

"The dog is *[MASK]* and the dog is hungry". Through training, the model will learn to predict the missing word based on the surrounding context. With this process, BERT can understand the connection between words and accurately predict missing words. With this approach, BERT gains the ability to grasp the underlying meaning of a sentence, even when certain words are concealed or deleted.

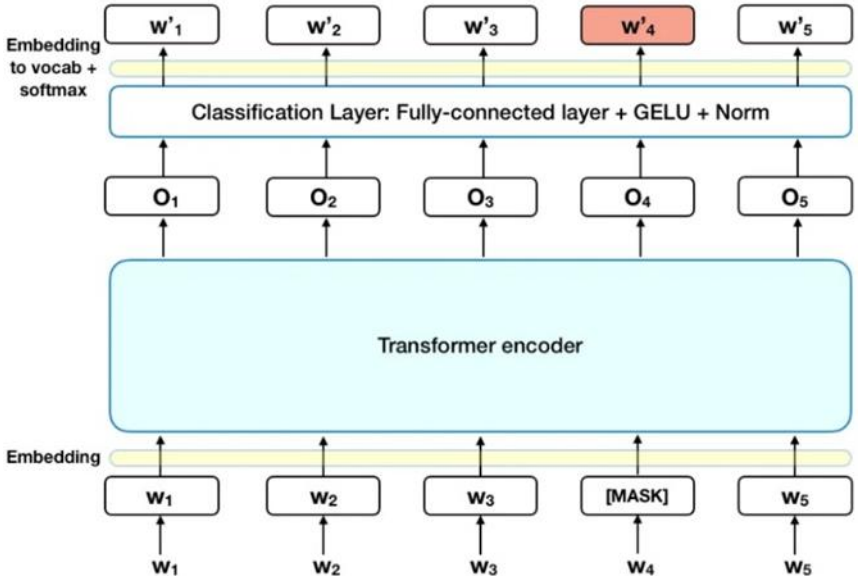

Figure 12. *Architecture of BERT (Rani Horev, 2018)*

NSP improves BERT's comprehension of sentence dependencies. It requires training the model to assess whether two sentences in a pair follow a logical sequence or not. During the training process, BERT is provided with pairs of sentences and tasked with determining the logical coherence between them. For example, let's take a sentence pair: "The dog is hungry. The weather outside is rainy. It is important for the BERT model to understand that the first and second sentences are unrelated. Training BERT on these tasks enhances its understanding of sentence relationships, information flow, and sentence structure. NSP is highly beneficial for tasks such as document summarization, questioning, and answering, as it requires a deep understanding of the relationships between words and sentences.

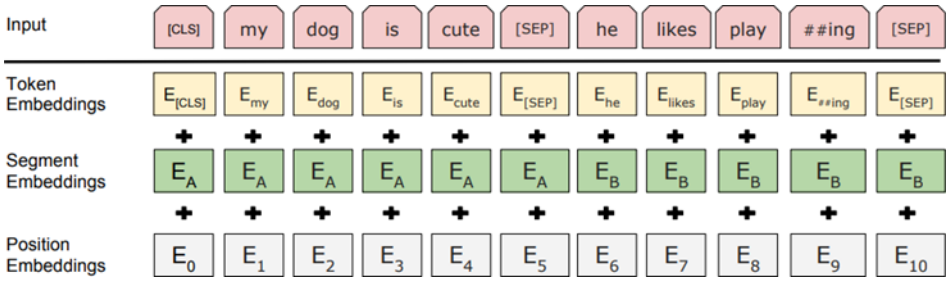

***Figure 13**. BERT input representation. (Devlin et al., 2019)*

4.2 Evaluation Metrics Overview

Within this section, you will find a comprehensive overview of the evaluation metrics utilised to assess the performance and various aspects of the models employed for sentiment analysis. These metrics are essential for measuring the effectiveness and accuracy of predicting sentiment labels for the reviews data.

Accuracy

Accuracy is a quantitative measure used to assess the overall correctness of predictions. It is calculated by determining the proportion of correctly predicted instances out of the total number of instances. The technique is frequently employed for datasets that have an equal distribution of classes, but its dependability may decrease when working with datasets that have imbalanced classes. The calculation of accuracy is determined by equation 3, as shown below.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + False\ Negative}$$

Equation 3. Accuracy

Precision

Precision is an evaluation metric that measures the proportion of positive instances precisely predicted in relation to the total number of positive instances predicted. The primary focus is on reducing the occurrence of false positives, which renders it advantageous

in situations where the expense of such errors is substantial. The precision is computed as shown in Equation 4.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

Equation 4. Precision

Recall

Recall is an evaluation metric that compares the proportion of accurately predicted positive instances to the total number of actual positive instances. It is alternatively referred to as sensitivity or true positive rate. Its importance is underscored in scenarios where the financial impact of false negatives is considerable, as its primary objective is to minimise the occurrence of overlooked positive cases. Recall is computed as follows via equation 5.

$$Recall = \frac{TruePositive}{TruePositive + False\ Negative}$$

Equation 5. Recall

F1 Score

The F1 Score is a metric used to evaluate the balance between precision and recall. The calculation involves taking the harmonic mean of precision and recall. The F1 Score is especially useful in situations where there is a disparity between the classes, as it takes into account both precision and recall during the evaluation. The calculation of recall is determined using equation 6, as shown below.

$$F1Score = 2 \text{ x } \frac{Precision\ x\ Recall}{Precision + \ Recall}$$

Equation 6. F1 Score

Confusion matrix

A confusion matrix is a tool used to evaluate the performance of a machine learning model in predicting various classes. Assessing and comprehending the performance of a machine learning model is beneficial in a classification challenge. This matrix typically consists of four categories:

- True Positives (TP): These are instances where the model is capable of correctly predicting the positive class.
- True Negatives (TN): These are instances where the model is capable of correctly predicting the negative class.
- False Positives (FP): These are instances where the model incorrectly predicts the positive class whereas the real result is negative. It's also known as a Type I error or false alarm.
- False Negatives (FN): These are instances where the model incorrectly predicts the negative class whereas the real result is positive. It's also known as a Type II error or miss.

	(Predicted) Negative	*(Predicted) Positive*
(Actual) Negative	*TN*	*FP*
(Actual) Positive	*FN*	*TP*

***Figure 14**. Confusion matrix*

Through the analysis of these categories, we can assess the model's effectiveness and compute a range of metrics including accuracy, precision, recall, and F1 Score. The confusion matrix provides insights into the model's ability to accurately classify distinct classes and make accurate predictions.

4.3 Data Selection

In the data selection phase, careful consideration was given in order to select an appropriate dataset. The data used for the research purpose is sourced from Amazon. The

dataset selection process was driven by its alignment with the research objectives, considering the significance of Amazon product reviews as a valuable resource for sentiment analysis. The dataset consisted of a diverse range of reviews covering different products belonging to the Music Category of products. Furthermore, the availability of sentiment labels was taken into consideration, as labelled data plays a vital role in training and evaluating sentiment analysis models. Stringent measures were implemented to ensure the dataset's integrity, involving thorough data cleaning, and preprocessing to eliminate any noise or irrelevant information.

During the data selection phase, we took great care to choose an appropriate dataset. The data used for the research purpose is obtained from Amazon. The dataset selection process was guided by its alignment with the research objectives, taking into account the importance of Amazon product reviews as a valuable resource for sentiment analysis. The dataset included a wide variety of reviews that encompassed various products within the Music Category. In addition, the presence of sentiment labels was considered, as labelled data is crucial for training and evaluating sentiment analysis models. Stringent measures were taken to ensure the integrity of the dataset, including extensive data cleaning and preprocessing to remove any noise or irrelevant information.

Table 2. *Dataset attributes (amazon e-commerce consumer reviews) (Ni et al., 2019).*

Attribute name	Attribute description
reviewerID	ID of the reviewer, e.g. A2SUAM1J3GNN3B
asin	ID of the product, e.g. 0000013714
reviewerName	Name of the reviewer
vote	Helpful votes of the review
style	A dictionary of the product metadata, e.g., "Format" is"Hardcover"
reviewText	Text of the review
overall	Rating of the product
summary	Summary of the review
unixReviewTime	Time of the review (Unix time)
reviewTime	Time of the review (raw)
image	Images that users post after they have received the product

The reviews vary from May 1996 to Oct 2018. This table displays the attributes of the dataset. The dataset is publicly available and was provided by Ni et al., (2019).

4.4 Summary

Based on the methodology section, it is evident how the chosen algorithm will operate with sample data. After that, there will be a discussion of evaluation metrics that will be used to rank the models. Lastly, the chosen dataset will undergo pre-processing, which will be further elaborated in the upcoming sections. The process and result will be discussed in more detail in the upcoming section on Empirical Findings.

CHAPTER 5

EXPERIMENTAL SETUP AND EMPIRICAL FINDINGS

This section highlights the setup utilised in this research. Providing an overview of the technical design and project architecture utilised in the implementation of this research project. First, this section discusses the resources used for this research. After this step, the results will be shown for the models that were chosen.

In addition, this section will discuss the outcomes achieved by various models. These results, along with the evaluation metrics, will assist in determining the most optimal model. Eventually, it will become clear which feature extraction method works best when combined with the chosen machine learning models.

The figure below illustrates the fundamental sequence of events that comprise the research experiment.

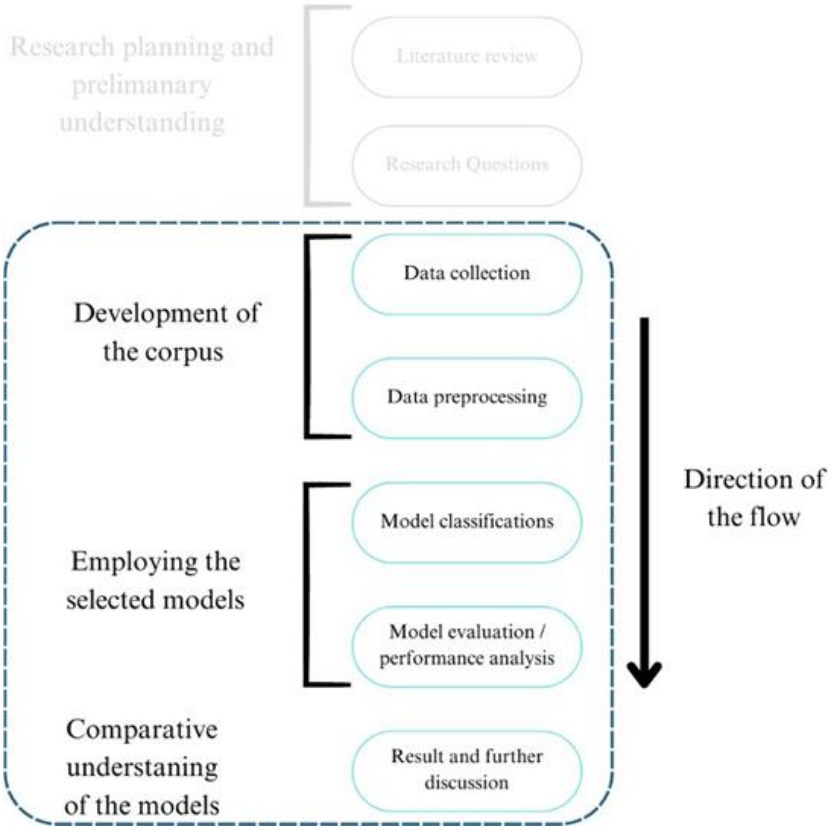

Figure 15. *Basic guideline for the flow of the research experiment*

Data Collection: Raw data is acquired from an e-commerce platform, such as Amazon reviews.

Data Preprocessing: The text data is processed by cleaning it, tokenizing it, and removing any stop words and special characters.

Model Selection: In the model setup section, we evaluate various machine learning and deep learning models, such as SVM, LSTM, CNN, and BERT.

Model Evaluation: When evaluating models, it is important to consider performance metrics like accuracy, precision, recall, F1 score, and AUC ROC score. These metrics help determine the effectiveness of each model.

5.1 Design Specification

5.1.1 Overview of Architecture

The research process involves gathering data from Amazon Customer Reviews. This data is readily accessible to the public and is of significant magnitude. Considering the constraints of the research process, a subset of the data is randomly selected due to limited resources and the intensive nature of data processing.

Here is a broad overview of the architecture that will be used to support the implementation of this research. There are three levels of division.

Presentation level: At the presentation level, we focus on creating visually appealing and informative representations of the project's results and insights.

Logic/Machine Learning level: At the logic level, we have a powerful deep learning model that will be utilised for sentiment analysis.

Data level: At the data level, the final layer consists of the raw data necessary for research experiments. Various tools and technologies are employed to retrieve, manipulate, and store this data.

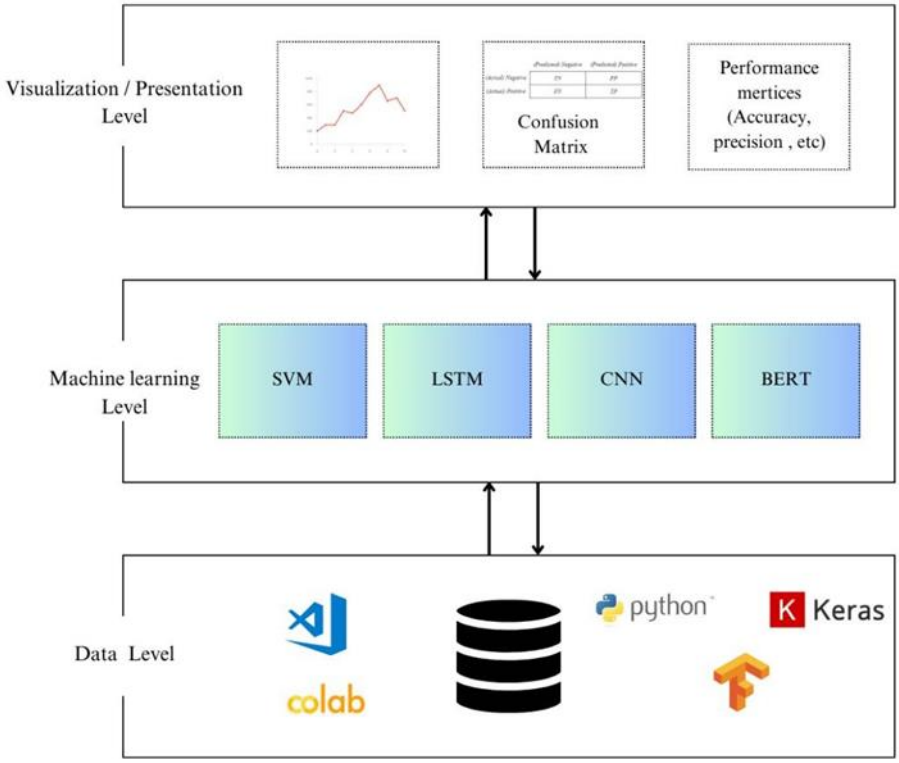

Figure 16. *Overview of the project architecture*

5.1.2 System configuration

For this research, the system configuration required utilising Google Colab, a cloud-based Jupyter Notebook environment, to carry out the experiments. Python was chosen as the programming language for this project, along with various Machine Learning (ML) libraries and deep learning frameworks suitable for Natural Language Processing (NLP) tasks.

Table 3. *System configuration*

Operating System	Windows 11
Memory	8 GB
CPU	Intel® Core™ i5-10210U processor
Software	Jupyter Notebook, Google Colaboratory
Python Version	3.10.12
Keras Version	2.13.1
TensorFlow Version	2.13.0

In this study, deep learning methods like LSTM, CNNs, and transformers were employed for the NLP tasks. BERT, a prominent model, was also utilised. The deep learning models were implemented using the TensorFlow and Keras libraries. The experiments were carried out on a Google Colab virtual machine, which granted access to high-performance GPUs for efficient computation. Using Google Colab, Python, machine learning libraries, and deep learning methods, this research aimed to leverage the convenience and computational power of these tools to conduct thorough NLP analysis, including precise sentiment analysis and other related tasks.

A possible alternative approach to KDD (Knowledge Discovery in Database) is the CRISP-DM (Cross-Industry Standard Process for Data Mining) method. CRISP-DM is characterised by its meticulous attention to detail, iterative nature, and systematic approach, in contrast to KDD, which offers a broader and more general framework. KDD adopts a holistic approach by including the full knowledge discovery process, which encompasses not just data mining but also elements such as data selection, preprocessing, transformation, and knowledge display.

During this stage of the research, the data that has been pre-processed will be subjected to several machine learning models. The train dataset will be used to train the models, while the test dataset will be used to evaluate the model performance.

5.2 Data preprocessing

It is the subsequent phase in the KDD approach that comes after the data selection process. At this point, we do a first examination of the chosen data. The subsequent subsections provide an in-depth exploration of this level.

5.2.1 Data exploratory analysis

Null values

At this stage, an initial analysis is performed to identify the existence of null or missing values in the dataset. Performing this verification is essential to verify the existence of any empty values in the target attribute that necessitate attention, in order to effectively

address them during the cleaning process. Table 4 displays the existence of null values in 5 characteristics.

Table 4. *Null values in the dataset per column*

Attribute name	Null values
reviewerID	0
asin	0
reviewerName	150
vote	1268582
style	839992
reviewText	855
overall	0
summary	380
unixReviewTime	0
reviewTime	0
image	1482365

Rating count plot

This stage provides additional insight into the distribution of user ratings throughout the dataset. This provides an insight into the overall distribution of the reviews. It can be noted that the 5 rating exhibits the greatest concentration near 1,000,000. Therefore, it is inevitable that the dataset predominantly include positive reviews. Additionally, this information will be valuable for generating a well-rounded dataset for training and testing purposes.

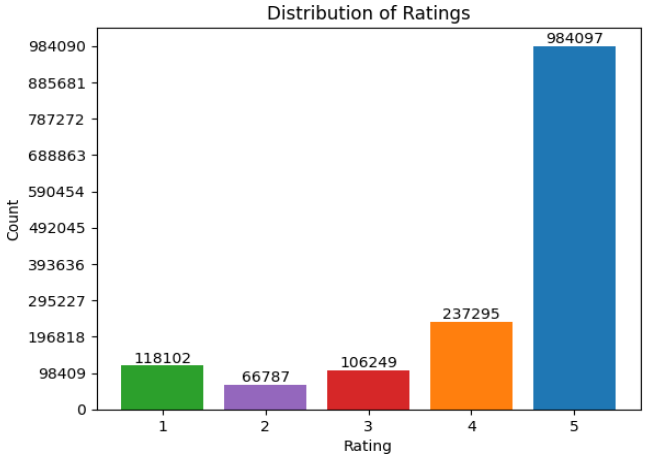

Figure 17. *Distribution of Ratings*

Word cloud

A word cloud provides insight into the frequency of words within a given corpus. This visualisation presents a concise overview of the most commonly used words, with the magnitude of each word directly corresponding to its frequency in the corpus. As previously stated, the primary contribution comes from reviews with a rating of 5. There is a high likelihood that there will be a greater number of positive words displayed in a bigger font size compared to negative terms.

1. Positive reviews only word cloud.

Figure 18*. Word cloud of positive reviews*

2. Negative reviews word cloud

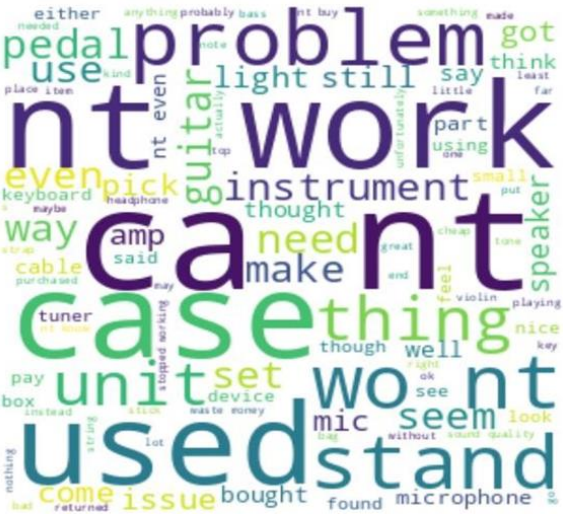

Figure 19. *Word cloud of negative reviews*

5.2.2 Data cleansing

Data cleaning, sometimes referred to as data cleaning or data scrubbing, is the systematic procedure of detecting and rectifying or eliminating errors, discrepancies, and irregularities in a collection of data. The process entails a sequence of measures to guarantee the accuracy, comprehensiveness, and dependability of the data used for analysis. Below are few typical stages involved in the data cleansing procedure:

1. Dealing with missing values

Addressing missing values in a dataset is crucial to mitigate the potential impact of incomplete data on the analytical process. There are several approaches to address these missing values. One approach is to substitute these absent values via mean, median, or mode imputation. Two such examples of imputation approaches are SMOTE and ROSE. In addition, if it is not feasible to impute the missing values, one alternative is to remove the column containing those values. An advisable approach is to eliminate columns that have above 55% of missing values. Inputting a substantial amount of data can add bias and potentially alter the initial distribution of the data.

Referring to figure 5, which provides the specifics of the quantity of missing values in the dataset. Consequently, the columns pertaining to Vote, Style, and Image can be eliminated from the dataset. Moreover, regarding the task of assigning values to the remaining columns that have missing values. In this scenario, imputation is not a viable choice as their data consists of either a distinct identifier or string data. Therefore, they can be maintained in their current state.

However, it is necessary to eliminate the missing values for the reviewText property from the dataset. Due to the difficulty of imputing string data, this study strategy will involve dropping rows with incomplete data. This will ensure a consistent column for sentiment analysis. Although there are missing values in the reviewer Name and summary columns, they are acceptable for this research as they do not directly affect the analysis. Consequently, these columns can be removed from the dataset.

Table 5. *Null values after cleaning the dataset*

Attribute name	**Null values**
reviewerID	0
asin	0
reviewText	0
overall	0
unixReviewTime	0
reviewTime	0

2. Dealing with Duplicates

Replicating values might significantly contribute to inconsistency in the sentiment analysis procedure. Therefore, it is imperative to eliminate duplicate data prior to proceeding. Replication can result in a partiality in the portrayal of the emotional attitude. The presence of excessive redundant data can significantly impact the training process of the sentiment model. Furthermore, there is a possibility of overfitting. Given the model's extensive knowledge of a particular instance of the data.

Performance matrices can be artificially increased when duplicate data is present. The model's ability to predict the scores of duplicated data will be enhanced, resulting in an artificial increase in accuracy.

3. Handling Outliners

Outliers are data points that exhibit considerable deviation from the rest of the dataset. Based on the analytic objectives and the nature of the data, a decision can be made regarding whether or not to eliminate the outliers. This study approach necessitated the use of two main columns.

The dataset contains two main attributes: the first one is the 'overall' ratings, which represent the overall evaluation of the items. The second property is the 'reviewText', which contains the data needed for sentiment analysis in the research.

During the data purification process, the initial analysis reveals that there are no outliers requiring treatment.

***Table 6**. Number reviews per rating (1-5)*

Ratings	Value count
5	960812
4	230948
3	65005
2	103417
1	114729

4. Cleaning review text

A crucial endeavor is undertaken to cleanse the wording of the Reviews. Users often exhibit a lack of organization or planning when expressing themselves on the online platform. Some users possess proficiency in various languages and distinct dialects. In addition, textual data acquired from other sources, such as e-commerce website pages or any type of social media, may include several HTML tags. The presence of these tags is inconsequential for sentiment analysis as they include formatting information. Hence, it is

crucial to remove these HTML tags from the text in order to concentrate exclusively on the content of the review throughout the analysis.

Punctuation marks, such as commas, periods, and exclamation marks, are often considered as noise because they often do not convey any emotional or sentiment-related information. In addition, any excessive spaces, line breaks, or tabs can be eliminated as they do not have any impact on sentiment analysis.

To further standardise the data, the accented characters are also eliminated. Finally, all the strings are transformed to lowercase in order to enhance the level of data normalisation. Therefore, it would guarantee that words like 'amazing' and 'Amazing' receive equal consideration.

5.2.3 Data balancing

Imbalanced datasets present a clear obstacle in sentiment analysis. Therefore, ensuring data balancing is essential in order to tackle this difficulty. Typically, this imbalance occurs in the dataset when one class is much more prevalent than the other, resulting in bias in the model's predictions. This part will discuss the significance, benefits, and drawbacks of data balancing, as well as the procedures involved in balancing the dataset utilised for this thesis.

As depicted in Figure 20, it is evident that there exists a significant disparity between the two classes. Consequently, the process of data balancing is implemented.

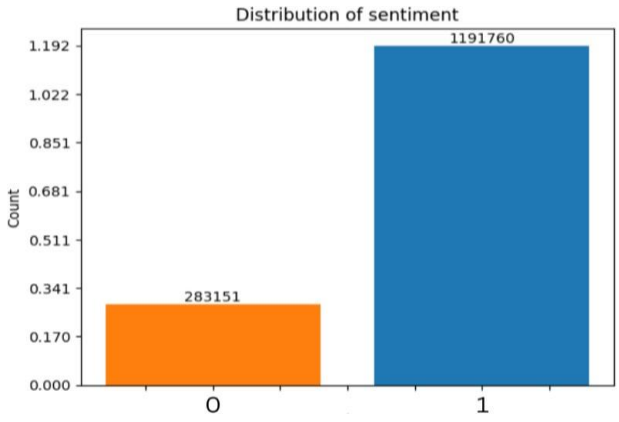

Figure 20. *Distribution of sentiment Negative (0) and Positive (1) in the dataset*

The significance of data balancing lies in its ability to address biases that may be created towards the majority classes. This model has a tendency to inadequately depict and exhibit subpar performance when it comes to the minority classes. In order to generate impartial predictions, the model must acquire knowledge from all the classes in an equitable manner. Several methodologies exist for data balancing, such as:

1. Oversampling refers to the process of increasing the number of samples in a dataset in order to balance the representation of different classes or categories.
    a. Random oversampling: The minority class is augmented by duplicating the minority instances in a random manner.
    b. SMOTE (Synthetic Minority Over-Sampling Technique): This method functions by artificially augmenting the number of instances. It does it by interpolating between the existing instances.

2. Subsampling

This strategy involves randomly removing instances from the majority class in order to achieve balance within the dataset. This approach can also be accomplished by determining the centroid of the dominant classes and thereafter eliminating those occurrences.

3. Combustion Sampling

This strategy utilises a combination of oversampling and undersampling techniques.

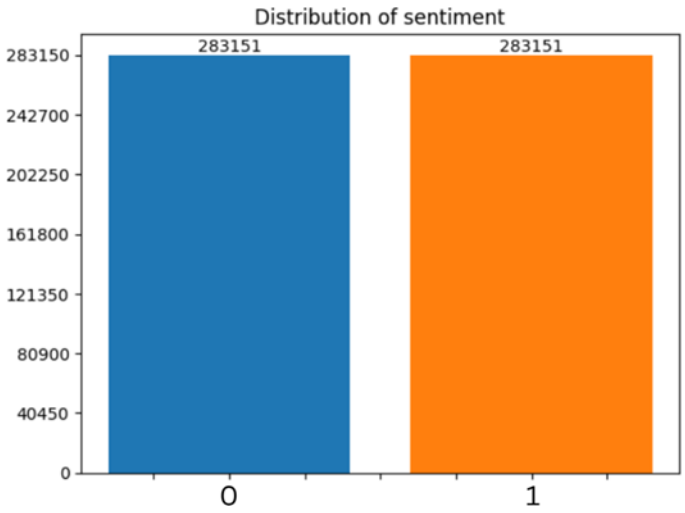

Figure 21. *Distribution of Negative (0) and Positive (1) reviews after undersampling of the dataset.*

The undersampling strategy will be employed for the aim of this research. The dataset will be balanced by randomly oversampling the majority class labelled as "target: 1". Therefore, reducing the likelihood of biases.

5.2.4 Train Test Split

A crucial phase in the creation of every machine learning model is the division of data into training and testing sets, known as the train-test split. The dataset is partitioned into two distinct subsets: a training set and a test set. This split is beneficial for training the machine learning model and subsequently assessing the model's performance.

The primary goal of this partition is to create a subset of data that is not used during model training, but is instead reserved for evaluating the model's performance. Given that this particular portion of data already has labels assigned to it, it can be verified by comparing the predicted label with the existing labels. This will allow us to determine the accuracy of the trained model. An effective strategy is to divide the dataset using the 80-20 rule, which is sometimes referred to as the Pareto principle. 80% of the dataset is allocated as the training dataset, while the remaining 20% is designated as the test dataset. It is essential to refrain from training on the test dataset to prevent overfitting.

Randomness and Reproducibility: It is crucial to ensure that the train-test split is conducted in a random manner to prevent any potential bias in the splitting of the data. The use of randomization in the process aids in achieving an impartial evaluation of the model's effectiveness. Furthermore, in order to guarantee reproducibility, it is advisable to establish a random seed value prior to executing the split.

Cross-validation is a technique that includes dividing the data into several subsets, or folds, and conducting several train-test splits. This method is used in addition to the traditional train-test split. Cross-validation offers a more resilient assessment by calculating the average performance across various divisions and can be especially valuable when the dataset is restricted.

5.2.5 Tokenization

Tokenization is the process of dividing or segmenting the original text into distinct units. The term used to refer to these chunks is "tokens". A token can refer to either a word or a letter, depending on the level of detail and the specific need. For instance, the phrase 'It is a nice product' will be divided into individual components such as 'It', 'is', 'a', 'good', and 'product'. Each of these texts is considered a token. Next, the procedure of removing stop words is carried out to eliminate tokens that have no significant impact on the phrase or any processing (Srinivas et al., 2021). Various libraries and techniques can be employed to carry out tokenization. Some examples of these libraries include NLTK, Keras, and Gensim.

Lemmatization is the procedure of reducing words to their base form. Lemmatization and tokenization are closely related, as tokens are transformed into their corresponding dictionary form, known as 'lemma'. For instance, the process of lemmatization will transform words such as 'running' and 'drinking' into 'run' and 'drink' respectively. This approach facilitates the precise capturing of the semantic significance of the words.

5.2.6 Feature extraction

As said succinctly in section 2.3, feature extraction is crucial in the process of sentiment analysis. Here is an illustration of how feature extraction operates on the chosen dataset of Amazon reviews.

After applying Word2Vec to the amazon review dataset, the model is capable of discerning the associations between groups of words.

Example 1: Similar Words

Script:
```
model.wv.most_similar('guitar')
```

Output:
```
[('guitars', 0.7190992832183838), ('ukelele', 0.6840307116508484), ('taylor',
0.6734611988067627), ('mandolin', 0.6629037261009216), ('banjo',
0.6451244950294495), ('acoustic', 0.6443262696266174), ('electric',
0.6412191390991211), ('ukulele', 0.6366580128669739), ('electricacoustic',
0.6203750967979431)]
```

In above example as it is clearly visible when we pass the word ‘guitar’, Word2Vec is able to recognize which terms in the document correspond to it at what degree.

Example 2: Similarity Score

Script:
```
model.wv.similarity('guitar', 'banjo')
```

Output:
```
0.6451245
```

In above example as it is clearly visible when terms ‘guitar’ and ‘banjo’ are .64% similar which

can also be cross checked from Example 1.

In Natural Language Processing, vectorization or feature extraction (Bengfort et al., between 1891 and 1894) in simple terms means that the textual data is converted into vectors (numerical representation). These vectors are easily understood for processing by the machine learning models.

5.3 Model setup and results

This section will specifically address the machine learning methods utilised in the sentiment analysis experiment. The analysis investigates four distinct models: Support Vector Machine (SVM), Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM), and Bidirectional Encoder Representations from Transformers (BERT). Each model possesses a distinct design that aids in capturing the many nuances of emotion. Its performance may be assessed for sentiment label classification.

The models utilised in this experiment have undergone fine-tuning using the specifically chosen dataset of Amazon customer reviews. The later section will analyse the performances of these models using various measures like accuracy, recall, precision, F1 score, confusion matrix, and ROC curve. The purpose of the model setup section is to offer a thorough

understanding of the structure and various parameters of each model. By conducting a systematic examination, the research will be able to derive relevant findings regarding the most efficient method for sentiment classification on the chosen dataset.

5.3.1 SVM

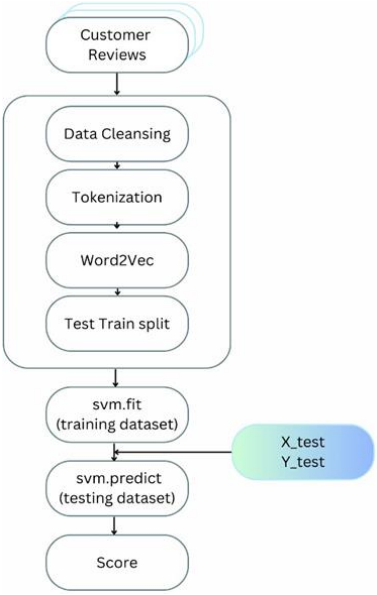

***Figure 22**. SVM Setup*

The SVM architecture with Word2Vec embeddings entails training the Word2Vec model to acquire word representations and subsequently generating feature vectors using these embeddings. In addition, the Support Vector Machine (SVM) classifier is trained using these vectors to do sentiment classification. The efficacy of this design primarily relies on the calibre of embeddings generated by Word2Vec. Similarly, the efficacy of the model utilising TF-IDF is contingent upon the quality of the TF-IDF representation and the discriminative capability that SVM can provide. The following diagram illustrates the SVM architecture employed in this experiment.

***Table 7**. SVM model parameters*

Parameter	Value
Kernel	Linear
Feature extractor	TD-IDF/Word2Vec
Vector Size (embedding)	100
C	Default (1.0)

Table 8. *SVM results using TF IDF and Word2Vec feature selection methods*

Model	Feature extraction	Accuracy	Precision	Recall	F1 Score	AUC ROC Score
SVM	TF IDF	0.853	0.857	0.848	0.852	0.925
SVM	Word2Vec	0.804	0.837	0.763	0.799	0.805

Based on the data presented in table 8, it can be concluded that the SVM model using TF-IDF achieved an accuracy of 85.3%, which is higher than the accuracy of 80.4% attained by the SVM model using Word2Vec. Therefore, this suggests that both versions of the models achieved an accuracy rate of over 80% in classifying consumer reviews as either positive or negative. According to the given data, it seems that utilising the SVM classifier with TF-IDF as the feature extraction technique resulted in the most superior overall performance for the sentiment analysis assignment. The model achieved an accuracy of 0.853, indicating that it correctly predicted the sentiment of approximately 85.3% of the test samples.

In addition, the Support Vector Machine (SVM) with Term Frequency-Inverse Document Frequency (TF-IDF) attained a precision of 0.857, indicating a high percentage of accurately predicted positive sentiment instances out of all predicted positive instances. The recall score of 0.848 indicates that the model successfully detected a significant proportion of the true positive sentiment events in the dataset. Finally, the F1-score of 0.852, which represents the harmonic mean of accuracy and recall, suggests a favourable equilibrium between precision and recall. On the other hand, using SVM with Word2Vec as the feature extraction method resulted in slightly lower performance, with an accuracy of 0.837, precision of 0.763, and recall of 0.799.. Ultimately, the Support Vector Machine (SVM) utilising the Term Frequency-Inverse Document Frequency (TF-IDF) approach shown superior performance compared to the SVM utilising Word2Vec in terms of accuracy, precision, recall, and total F1-score. Therefore, the SVM with TF-IDF is the preferred option for sentiment analysis in this particular situation.

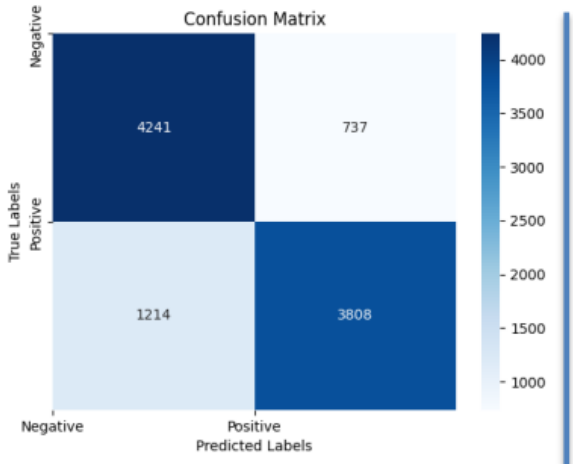

Figure 23. *SVM Wrod2Vec Confusion Matrix*

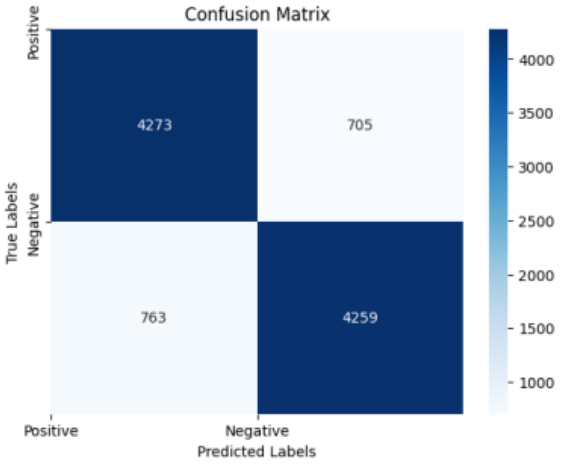

Figure 24. *SVM with TF IDF Confusion Matrix*

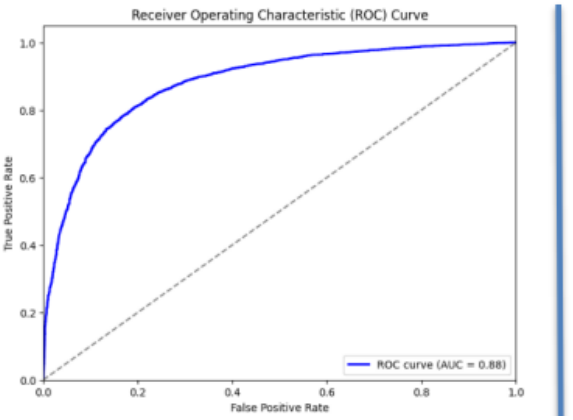

Figure 25*. Figure 25. SVM Wrod2Vec ROC Curve*

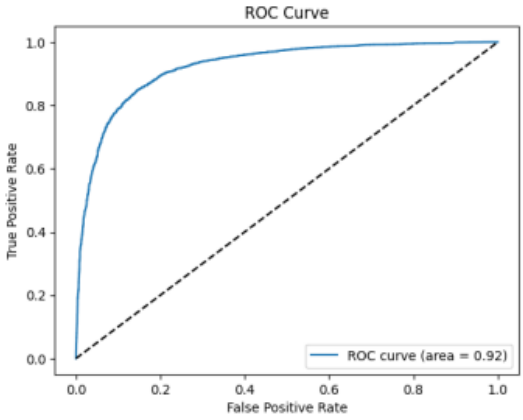

Figure 26. *SVM TF-IDF ROC Curve*

Figures 23 and 24's side-by-side confusion matrices offer a clearer picture of how both model versions were able to accurately identify the positives and negatives. The SVM model with Word2Vec achieved a high number of true positives (3808) and true negatives (4241), indicating its proficiency in correctly identifying positive and negative instances. Similarly, the SVM model with TF IDF achieved 4273 true negatives and 4259 true positives, further demonstrating its ability to accurately identify positive and negative instances. Nevertheless, the occurrence of incorrect positive identifications and incorrect negative identifications for Support Vector Machines (SVM) with Word2Vec (737 false positives, 1214 false negatives), and SVM with TF-IDF (705 false positives, 763 false negatives) respectively, suggests that there is still potential for enhancing the performance of these models.

Furthermore, based on the data presented in figures 25 and 26, it is evident that both models have the ability to differentiate between positive and negative samples. The Support Vector Machine (SVM) model using Term Frequency-Inverse Document Frequency (TF-IDF) achieved a notable Area Under the Curve (AUC) of 92%, whilst the SVM model using Word2Vec obtained an AUC of 88%. The SVM model with TF-IDF has a higher AUC value, indicating its superior ability to differentiate between positive and negative reviews compared to the SVM model using Word2Vec.

5.3.2 LSTM

LSTM, a form of recurrent neural network (RNN), is known for its ability to effectively capture sequential patterns in textual input. The LSTM model used in this experiment consists of 100 LSTM units, which correspond to 100 memory cells within the LSTM. This configuration enables the model to effectively learn and retain significant data and sequential patterns. The input length is configured to the maximum sequence length in order to ensure consistency. The LSTM model architecture included a dense layer with 64 units, triggered by the Rectified Linear Unit (ReLU) function. This brought non-linearity into the network and improved its ability to represent data. The output layer had a solitary unit operated by the sigmoid function, guaranteeing that the model generates a probability score ranging from 0 to 1, indicating the emotion polarity (positive or negative) of the input text.

The performance is enhanced by utilising the binary cross entropy loss function, which is often employed in binary classification scenarios. The Adam optimizer, renowned for its adaptable learning rates and rapid convergence, was selected to update the model's parameters during the training process. During the training phase, we employed a batch size of 32, enabling the model to change its parameters after processing each batch of 32 review samples. The training procedure was executed for 10 epochs, which indicates the number of times the model repeatedly processed the complete dataset during training. Ultimately, our LSTM model, which has been meticulously crafted and optimised with appropriate architecture and hyperparameters, effectively showcases its ability to do sentiment analysis on e-commerce review data.

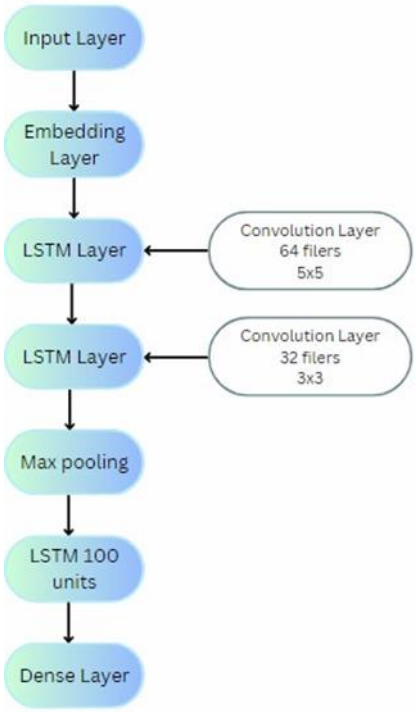

Figure 27. LSTM model setup

Table 9*. LSTM model parameters*

Parameter	**Value**
Input Length	Maximum sequence length
Embedding Output Dimension	100
LSTM Units	100
Dense Units	64
Dense Activation Function	relu
Output Layer Units	1
Output Layer Activation	sigmoid
Loss Function	binary_crossentropy
Optimizer	Adam
Batch Size	32
Epochs	10

***Table 10**. LSTM results using TF IDF and Word2Vec feature selection methods*

Model	Feature extraction	Accuracy	Precision	Recall	F1 Score	AUC ROC Score
LSTM	TF IDF	0.823	0.841	0.800	0.820	0.883
LSTM	Word2Vec	0.849	0.858	0.837	0.847	0.920

The LSTM model with TF-IDF achieved an accuracy of 82.3%. The algorithm accurately predicted the sentiment of around 82.3% of the reviews in the test set. The accuracy rating of 84.1% indicates that the model correctly predicted whether a review was positive or negative 84.1% of the time. The recall rate of 80.0% signifies that the model accurately identified 80.0% of both positive and negative feelings in the test set. The F1 score, which is 82.0%, indicates a well-balanced evaluation of both precision and recall. Additionally, the AUC ROC score, which is 88.3%, demonstrates the model's capability to differentiate between positive and negative thoughts. Conversely, the LSTM model utilising Word2Vec achieved a superior accuracy of 84.9%, suggesting enhanced overall performance in comparison to the TF-IDF model. The model's accuracy rate of 85.8% indicates that its predictions were correct 85.8% of the time. The recall rate of 83.7% signifies that the model accurately detected 83.7% of both positive and negative feelings in the test set. The F1 score, which stands at 84.7%, indicates a well-balanced evaluation of both precision and recall. Additionally, the AUC ROC score, which is 92.0%, displays the model's robust capability to differentiate between positive and negative attitudes.

An enhanced comprehension of how both models were able to accurately detect the positives and negatives can be gained from examining the side-by-side confusion matrices in figures 28 and 29. The LSTM model with Word2Vec achieved a high number of true positives (4205) and true negatives (4287), indicating its proficiency in correctly identifying positive and negative instances. Similarly, the LSTM model with TF IDF achieved 4018 true negatives and 4221 true positives, further demonstrating its proficiency in identifying positives and negatives. Nevertheless, the occurrence of incorrect positive and negative results for LSTM with Word2Vec (691, 817), and LSTM with TF-IDF (757, 1004) correspondingly suggests that there is still potential for enhancing these models.

Furthermore, based on figures 30 and 31, it is evident that both models possess the ability to differentiate between positive and negative samples. The LSTM model utilising Word2Vec acquired a notable AUC (Area Under the Curve) of 92%, whilst the LSTM model employing LSTM attained an AUC of 88.3%. The LSTM model with TF-IDF has a higher AUC value, indicating its superior ability to differentiate between positive and negative reviews compared to the SVM model with Word2Vec.

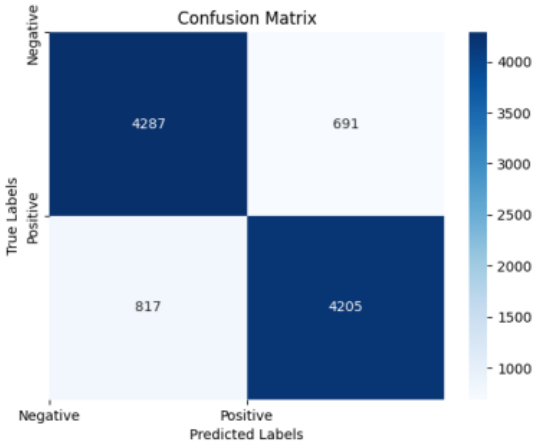

Figure 28. *LSTM with Word2Vec Confusion Matrix*

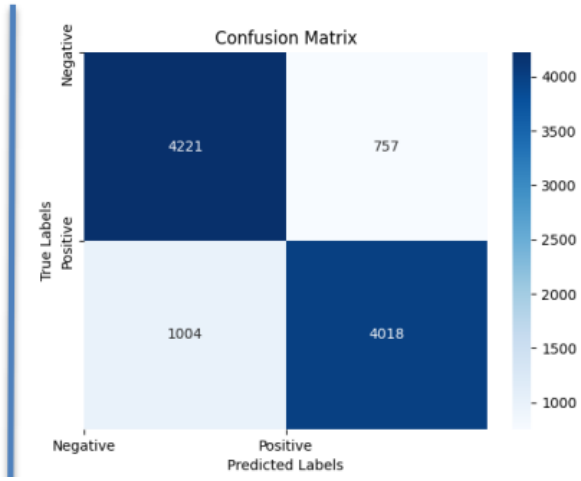

Figure 29*. LSTM with TF-IDF Confusion Matrix*

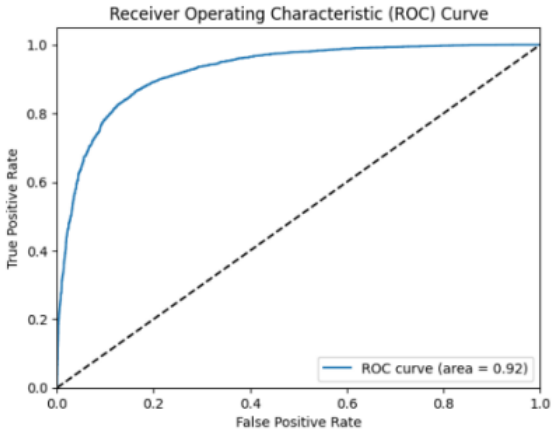

Figure 30. *Figure 30. LSTM with Word2Vec ROC Curve*

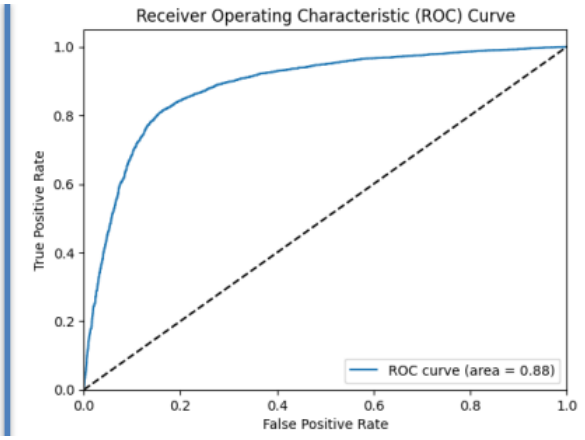

Figure 31. *LSTM with TF-IDF ROC Curve*

5.3.3 CNN

The following is a comprehensive description of the utilised configuration and structure for Convolutional Neural Network (CNN). The input shape is defined as (Number of samples, Number of features, 1), where each sample represents a feature and review text that is converted into a one-dimensional vector. The model comprises several convolutional layers, each defined by its kernel size, number of filters, and activation function. Afterwards, in order to reduce the size of the learnt features, pooling layers are used, offering the choice between MaxPooling and AveragePooling.

Dropout layers are used to address the issue of overfitting. The dropout rate is explicitly defined, for example as 0.2 or 0.5. During training, dropout selectively deactivates random neurons, promoting the network to learn properties that are not too reliant on particular neurons. The optimizers used in the experiment were Adam, SGD, and RMSprop. The learning rate, set at either 0.001 or 0.01, determines the magnitude of each step taken during the gradient descent process.

During the training process, the loss function used was binary cross entropy, which is commonly employed for binary classification tasks. The model parameters were updated using a batch size of either 32 or 64. The batch size was contingent upon the availability of computational resources. During the training, a predetermined number of epochs was established to ensure that the model was processed iteratively and uniformly. Finally, early stopping was included, which terminated the training process if there was no improvement in an epoch. This measure was taken to prevent overfitting.

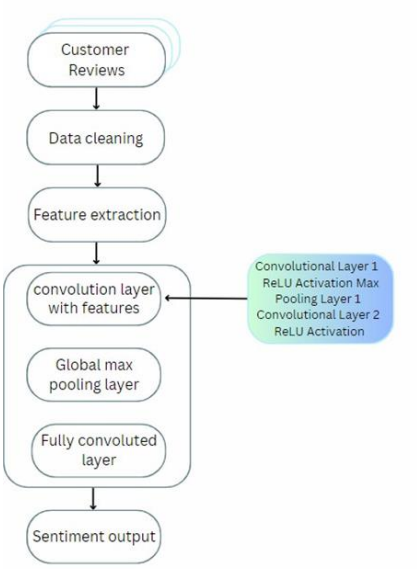

Figure 32. *CNN model setup*

Table 11*. CNN model setup*

Parameter	**Value**
Input Shape	(Number of samples, Number of features, 1)
Convolutional Layers	Number, Filters, Kernel Size, Activation
Pooling Layers	Type (MaxPooling, AveragePooling), Pool Size
Dense Layers	Number, Units, Activation
Dropout	Dropout rate (e.g., 0.2, 0.5)
Optimizer	Adam, SGD, RMSprop, etc.
Learning Rate	Value (e.g., 0.001, 0.01)
Loss Function	Binary Crossentropy, Categorical Crossentropy
Batch Size	Value (e.g., 32, 64)
Epochs	Number of training epochs (10)
Early Stopping	Patience (number of epochs with no improvement)
Model Architecture	CNN architecture, custom or predefined

Table 12. *CNN results using TF IDF and Word2Vec feature selection methods (below table)*

Model	Feature extraction	Accuracy	Precision	Recall	F1 Score	AUC ROC Score
CNN	TF IDF	0.578	0.606	0.481	0.4	0.591
CNN	Word2Vec	0.81	0.797	0.850	0.823	0.895

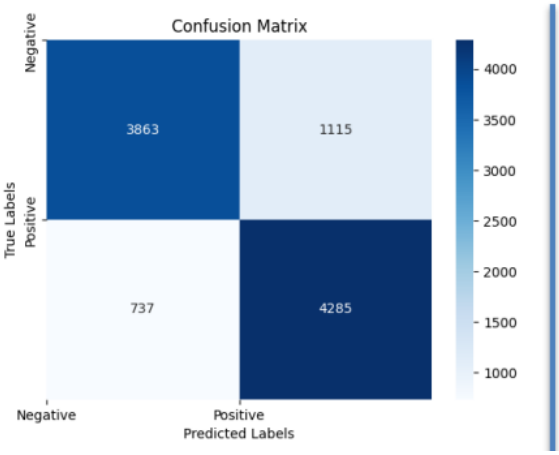

Figure 33. *CNN with Word2Vec Confusion Matrix*

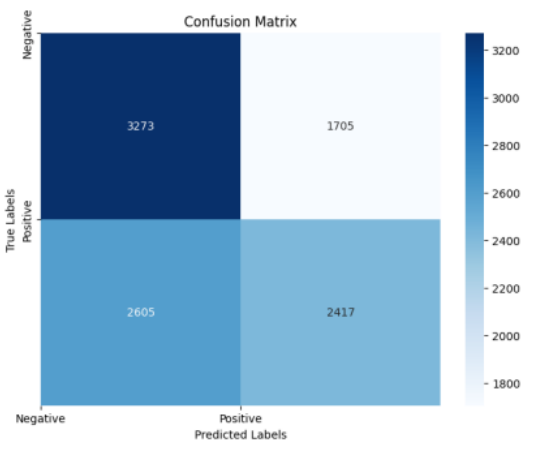

Figure 34. *CNN with TF-IDF Confusion Matrix*

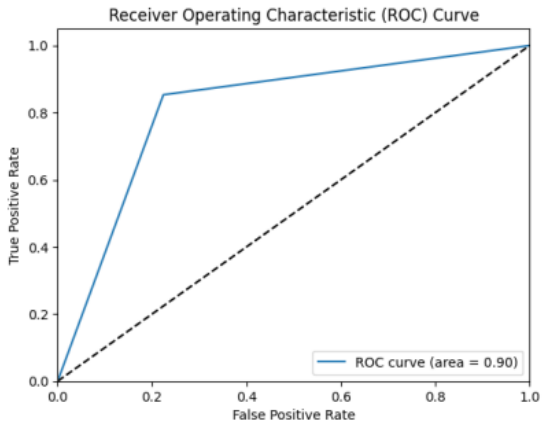

Figure 35. *CNN Word2Vec ROC Curve*

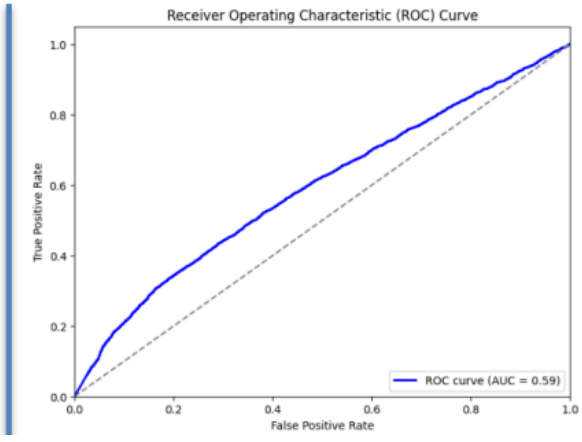

Figure 36. *CNN TF-IDF ROC Curve*

The results clearly indicate that the CNN model utilising Word2Vec as the feature extraction technique outperformed the CNN model using TF-IDF for sentiment analysis. The CNN- Word2Vec model demonstrated a commendable accuracy of 0.81, signifying its ability to accurately predict the emotion of around 81% of the test samples. Furthermore, the model exhibited a high level of precision (0.797), indicating that it made a minimal number of incorrect positive sentiment classifications. The recall score of 0.850 suggests that the model successfully detected a significant part of the actual instances of positive sentiment in the dataset. The F1-score, which is calculated as the harmonic mean of precision and recall, indicates a balanced performance between the two measures with a value of 0.823.

The CNN-Word2Vec model achieved a notable AUC-ROC score of 0.892, demonstrating its great accuracy in distinguishing between positive and negative sentiment instances. On the other hand, the CNN-TF IDF model demonstrated inferior overall performance, with an accuracy of 0.578, precision of 0.606, recall of 0.40, F1-score of 0.482, and AUC-ROC score of 0.607. Ultimately, the CNN-Word2Vec model demonstrated its superiority in sentiment analysis, surpassing the CNN-TF IDF model in all evaluation metrics.

5.3.4 BERT

BERT is a highly potent language model that has been pre-trained to effectively capture contextual information from extensive collections of texts. The BERT arrangement utilised the "bert-base-uncased" variation. This variation has undergone pre-training using 12 transformer layers and has a total of 110 million parameters. The BERT tokenizer is employed to segment the input text and produce input embeddings. The model architecture utilised the "BertForSequenceClassification", which was specifically built for jobs involving classification. This system generates a solitary output that represents the sentiment label. The tokenized input sequence is required to perform this task.

The maximum sequence length is defined as 150, ensuring that all sentences are of the same length by either padding or truncating them during tokenization. The binary cross entropy loss function was utilised, which is a commonly used option for binary classification tasks. The objective is to forecast the sentiment as either negative or positive. The Adam optimizer was employed, utilising a learning rate of 0.00002. This parameter determines the

magnitude of the update applied to the model parameter during the training process. During the training process, the data is partitioned into batches, with each batch consisting of 16 samples. Dividing the data into batches is crucial for addressing memory limitations and enabling efficient gradient changes. Ultimately, the training processes are repeated for a total of 10 epochs.

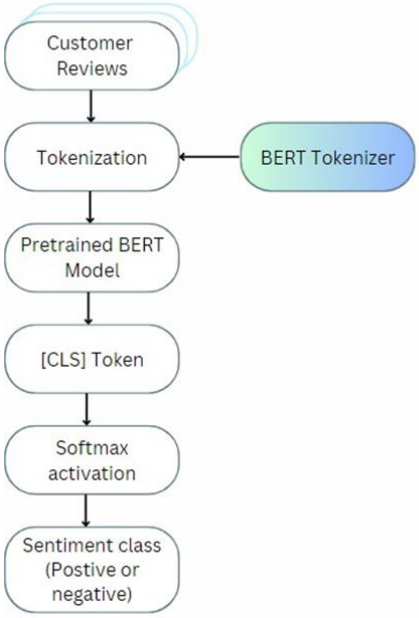

Figure 37. *BERT model setup*

Table 13. *BERT model setup*

Parameter	**Value**
Model Name	Bert-base-uncased (BERT pre-trained Model)
Tokenizer	Bert Tokenizer
Model Architecture	BertForSequenceClassification
Maximum Sequence Length	150
Loss Function	Binary Cross-Entropy (BCE) Loss
Optimizer	Adam
Learning Rate (Step size of Optimizer)	0.00002
Batch Size	16
Epochs	10

***Table 14**. BERT sentiment analysis performance result*

Model	Feature extraction	Accuracy	Precision	Recall	F1 Score	AUC ROC Score
BERT	BERT	0.863	0.838	0.903	0.869	0.928

The aforementioned findings achieved by the BERT model demonstrate its robust performance in accurately categorising sentiment for the given amazon ecommerce dataset. The model's accuracy of 86.3% indicates that it can accurately predict the sentiment label for a substantial proportion of the review samples. The model's excellent accuracy demonstrates its efficacy in comprehending and capturing the inherent sentiment patterns inside the textual data.

A recall rating of 90.3% signifies that the model is capable of correctly identifying 90% of the genuine positive instances, which is crucial in sentiment analysis to minimise the occurrence of false negatives. The F1 score of 86.9% indicates that the model achieves a favourable balance in accurately categorising positive attitudes, as it is calculated as the harmonic mean of precision and recall.

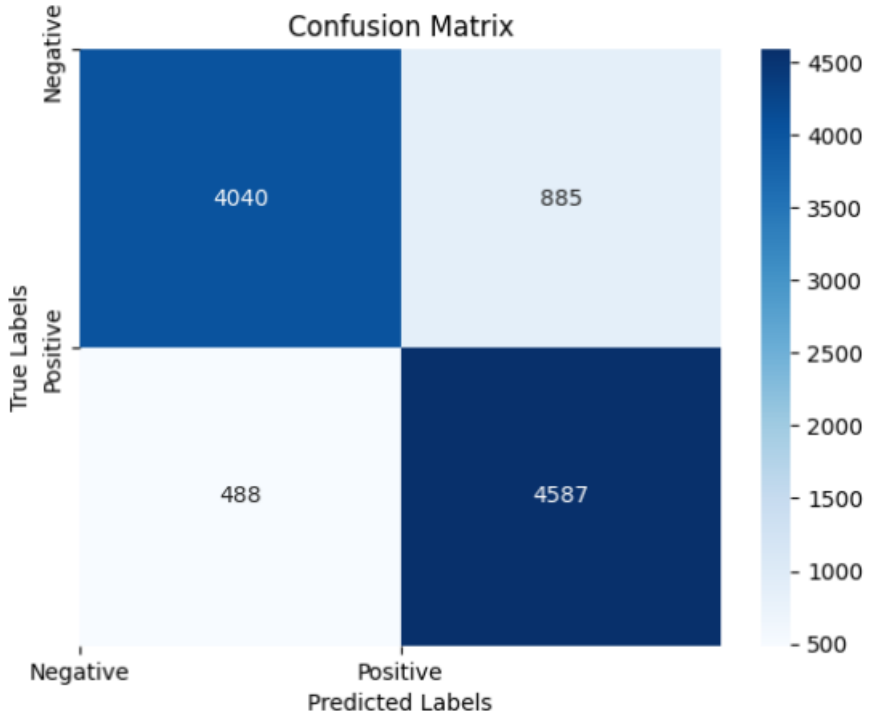

***Figure 38.** BERT confusion matrix*

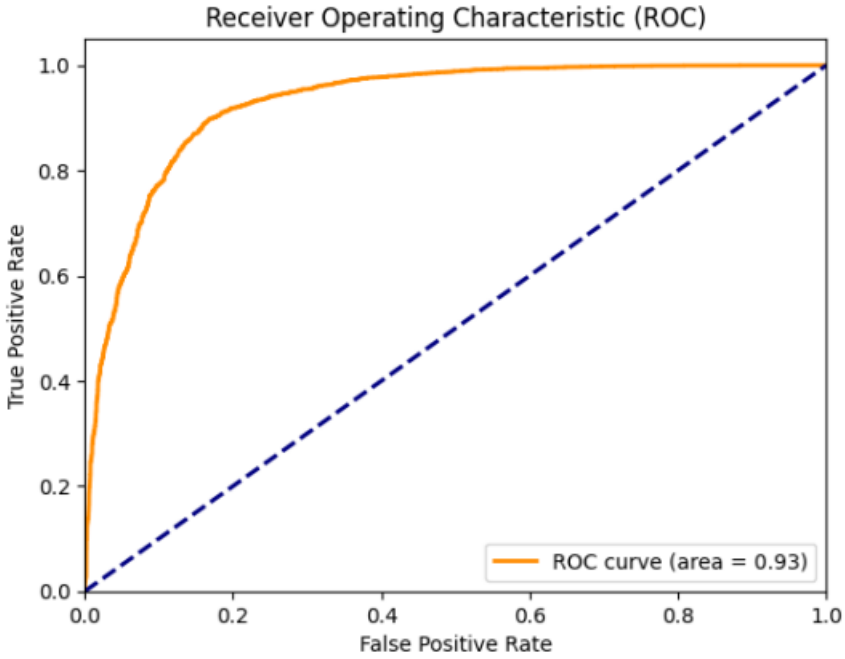

Figure 39. *BERT ROC Curve, AUC = 93%*

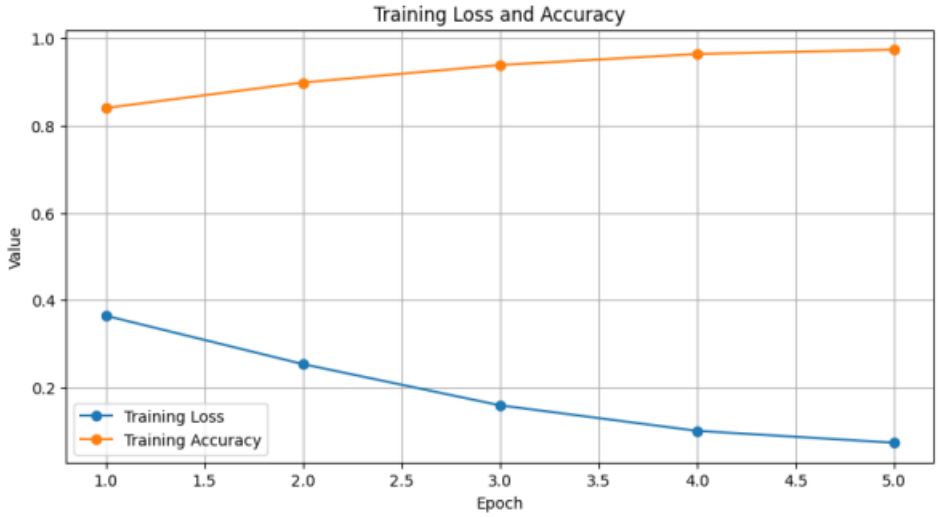

Figure 40. *BERT Training loss and training accuracy over 5 epochs*

The AUC ROC score is a crucial parameter for binary classification as it quantifies the model's ability to differentiate between positive and negative sentiment classes. A score of 92.8% indicates that this model has a good ability to differentiate between the two sentiments classes, resulting in confident predictions.

5.4. Performance comparison and addressing the research questions

The 'Performance evaluation' section will provide a comparison and analysis of the outcomes produced by the trained models. Each model will be evaluated on the test dataset using several evaluation approaches. Furthermore, the objective of this section is to acquire a more profound understanding of the models' efficacy in obtaining optimal accuracy.

The table below displays the performances of SVM, LSTM, CNN, and BERT models in order to determine which model produces the most optimal outcome for the given dataset and the goal of sentiment analysis.

Table 15. *Final evaluation matrix, comparing performances of all models*

Model	Feature extraction	Accuracy	Precision	Recall	F1 Score	AUC ROC Score
SVM	TF IDF	0.853	0.857	0.848	0.852	0.925
SVM	Word2Vec	0.804	0.837	0.763	0.799	0.805
LSTM	TF IDF	0.823	0.841	0.800	0.820	0.883
LSTM	Word2Vec	0.849	0.858	0.837	0.847	0.920
CNN	TF IDF	0.578	0.606	0.481	0.400	0.591
CNN	Word2Vec	0.81	0.797	0.850	0.823	0.895
BERT	**BERT**	**0.863**	**0.838**	**0.903**	**0.869**	**0.928**

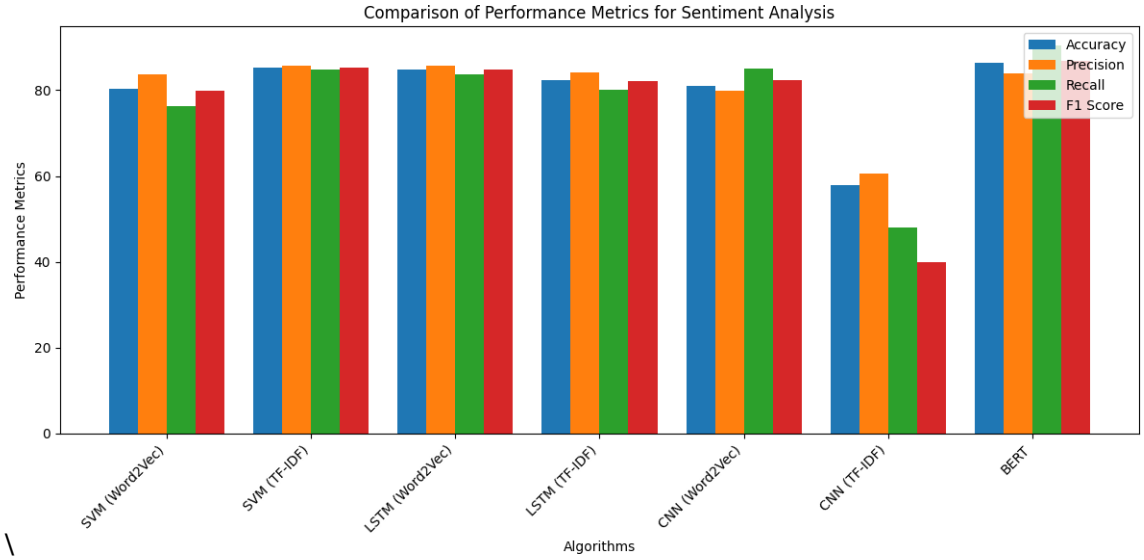

Figure 41. Comparison of performance metrics for sentiment analysis

Table 15 and figure 41, give a good overview to compare the different models employed for sentiment analysis for this research. This enables the research to answer the research questions posed in section 1.2. Accuracy, will serve as the most pivotal score to determine of these has the best performance. From a quick glance of the table, BERT has performed exceptionally well in comparison to other methods.

Furthermore, answering the research question will help dive deep into the performances of these models.

RQ1: *What is the comparative performance and effectiveness of various machine learning models for sentiment analysis on e-commerce review data?*

Based on the comparison done in section 4.4 with various machine learning models to perform sentiment analysis on the selected Amazon ecommerce review dataset, the research has pointed that BERT outperformed other included models. In the experiment BERT achieved the highest accuracy of 86.3%, this goes on to showing that it has a better classification capability of the customer reviews. Furthermore, BERT solidifies its superiority by remarkable demonstration of precision, recall and F1score.

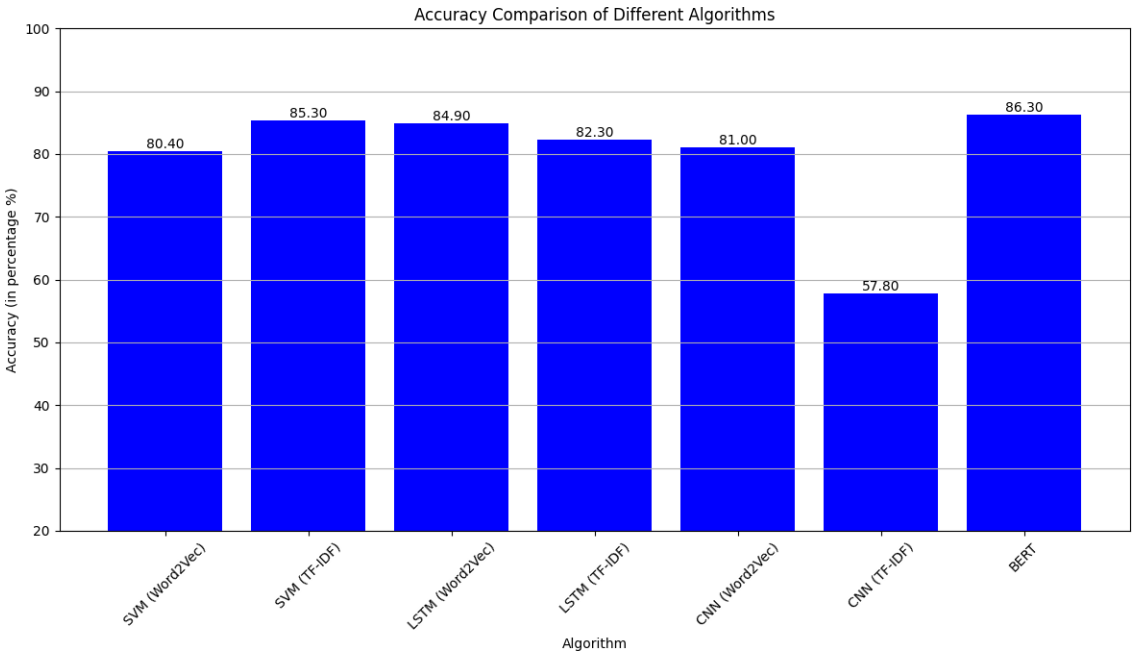

Figure 42. *BERT with highest accuracy compared to other models*

The Support Vector Machine (SVM) model, using the Term Frequency-Inverse Document Frequency (TF-IDF) technique, obtained the second highest performance, with an accuracy rate of 85.3%. Although it exhibited somewhat poorer Precision, Recall, and F1 Score in comparison to BERT, it nevertheless achieved impressive outcomes in sentiment analysis.

The LSTM model, together with Word2Vec, achieved a third-place ranking with an accuracy rate of 84.9%. The model had a good level of precision and recall, while its F1 score was slightly lower when compared to BERT and SVM with TF-IDF. The LSTM model, combined with TF-IDF, obtained a ranking of fourth place, with an accuracy rate of 82.3%. Although the F1 Score demonstrated comparable Precision and Recall, it was marginally inferior than the top-performing models.

The Convolutional Neural Network (CNN) model, trained using the Word2Vec algorithm, achieved an accuracy of 81.0% and ranked sixth in performance. The results showed high precision and AUC ROC score, but were worse in terms of recall and F1 score. The Convolutional Neural Network (CNN) implemented with the Term Frequency-Inverse Document Frequency (TF-IDF) technique had the least satisfactory performance, attaining an accuracy rate of 57.8%. This model had the worse Precision, Recall, F1 Score, and AUC ROC Score compared to all the examined models.

RQ2: *Which feature extraction method is more suitable for sentiment analysis of Amazon reviews, TF-IDF or Word2Vec?*

When it comes to analysing the sentiment of customer reviews on Amazon, BERT has outperformed standard approaches of extracting features such as Word2Vec and TF-IDF. One of the objectives of this research was to ascertain the optimal feature extraction method between Word2Vec and TF-IDF. Word2Vec has clearly outperformed other methods in most circumstances.

The LSTM model using Word2Vec achieved an accuracy of 84.9%, surpassing the accuracy of 82.3% reached by the LSTM model using TF-IDF. Similarly, the performance of CNN with Word2Vec was superior to that of CNN with TF-IDF, achieving accuracies of 81% and 57.8% respectively.

5.5 Summary

Section 4 presents the research that has put the theory covered in the previous parts into practice. The outcome of the models is determined by their performance and the configuration of the model. To ensure a fair assessment, the models employed with Word2Vec and TF-IDF were kept unchanged, with only modifications made to their feature extraction techniques. The final matrix in table 41 provides a comparison of the selected models, revealing that BERT emerges as the superior performer among the others.

CHAPTER 6

DISCUSSION

The conclusion drawn from this thesis work are found to be highly relevant and informative towards the use of deep learning models for sentiment analysis of e-commerce product reviews. This section considers the implications of finding comprised of the study, the difficulties experienced during the research process and possible areas for further research.

The comparative analysis of different machine learning models, including SVM, LSTM, CNN, and BERT, highlights several key observations.

Model Performance:

BERT: As predicted, both training and testing results with all the evaluation measures including Accuracy, Precision, Recall, and F1 spoke volumes about the efficiency of the BERT model. Hence, its bidirectional contextual capturing is helpful in enabling the model to capture the shade of emotions contained in the reviews about products.

LSTM and CNN: Hence, we can conclude that both LSTM and CNN models were accurate in their predictions, with LSTM taking a slight edge compared to CNN on account of its ability to learn sequence information within the text. Similarly, CNN, specially designed for pattern identification in local regions, performed well in the study but yielded slightly lower results compared to LSTM in modeling longer sequences.

SVM: While using SVM, the results were quite promising, but not very groundbreaking compared to the results of deep learning models. The reason for saying so is because it has a simpler architecture than LSTMs and is computationally less intensive, hence suitable for use in small data sets but not efficient in large Scale and Text data.

Feature Extraction Methods:

The study further validated the feasibility in extracting features utilizing Word2Vec and TF-IDF where Word2Vec was slightly more effective than TF-IDF in most scenarios. Word2Vec outperforms the other models due to its capability of generating vectorsizes that represent the semantic connections or radiations of words.

TF-IDF though stands as a relatively powerful baseline and is computationally less complex compared to other methods making it appropriate for specific applications.

Data Preprocessing:

Preprocessing steps like clean text, tokenization, and normalization were very important in improving the outcomes of the model. The process of eliminating stop words and the management of special symbols were critical in the process of purging the models of irrelevant, analytic data.

Several challenges were encountered during the research process:

Data Imbalance:

When analyzing the data we also identified three distinct classes and overwhelming majority of the data was positive. This was true mainly in the sense that it created an imbalance with regard to model training and making of biased predictions. One major challenge was the data imbalance whereby the number of samples belonging to some classes far exceeded others.

Computational Resources:

Performing the training of deep learning models on a large set took considerable amount of computation power. This was tackled through the use augmented cloud-based tools such as Google Colab which supports computational resources with GPUs for training, though the training duration was still sizable.

Handling Sarcasm and Irony:

Depending on the type of review, it is not easy to distinguish between sarcasm and irony. Examples included cases where models or algorithms became confused and could not predict the sentiments properly, pointing out to a specific deficiency in existing models that need to be investigated and optimized.

CHAPTER 7

CONCLUSION

7.1 Conclusion

This thesis primarily focuses on sentiment analysis, which involves the classification of consumer reviews. It also addresses the obstacles encountered in sentiment analysis. This literature is a comparative study that showcases the proficiency of several machine learning models. This thesis's research examines four distinct machine learning models to contribute to the expanding understanding of the classification accuracy of consumer reviews.

This research use a dataset consisting of consumer reviews from Amazon. The dataset is annotated with reviews that span a rating scale of 1 to 5 stars. In this study, those that had ratings of 0 to 3 stars were categorised as unfavourable, whereas those with ratings of 4 to 5 stars were considered positive. In order to mitigate biases, a balanced dataset was obtained by under-sampling the data, as there was an imbalance between positive and negative ratings in the class distribution. Four distinct machine learning algorithms were utilised, namely SVM (Support Vector Machine), LSTM (Long Term Short Memory), CNN (Convolutional Neural Network), and BERT (Bi-directional Encoder Representations from Transformers), to classify user reviews on Amazon.

The results unequivocally demonstrate that BERT surpassed the other models, attaining the greatest accuracy of 86.3% and exhibiting exceptional precision, recall, and AUC ROC Score. The discovery emphasises the effectiveness of deep learning models, specifically BERT, in comprehending intricate patterns and contextual information in textual data. BERT, being one of the most advanced models currently available, was anticipated to have achieved superior performance.

Notably, Support Vector Machines (SVM) using Term Frequency-Inverse Document Frequency (TF-IDF) and Long Short-Term Memory (LSTM) using

Word2Vec achieved the highest performance after BERT. Both achieving an accuracy of 85.3% and 84.9% respectively. Furthermore, it should be noted that a significant number of these models can require a substantial amount of resources. In particular, BERT and LSTM required greater computational resources in comparison to SVM and CNN. Furthermore, the comparison between TF-IDF and Word2Vec aided in determining the most suitable feature extraction method. Word2Vec emerged as the superior choice, suggesting that Word2Vec embeddings are more suitable for sentiment analysis of the amazon customer review dataset.

Ultimately, the study sought to assess the efficacy and proficiency of several machine learning models in doing sentiment analysis on e-commerce consumer review data. The study has gained useful insights into the strengths and limits of the selected models through tests and evaluation. This study adds to the growing field of natural language processing in sentiment analysis through the application of machine learning. The study's insights can assist firms in transitioning towards data-driven decision making. In order to make more informed decisions, stakeholders should familiarise themselves with the relative advantages of various models and utilise one of these strategies.

7.2 Scope for Further research

As with any research, it is essential to consider the limitations of this study. For instance, the performance of the models may change on different datasets, domain specificity, and the generalizability of the findings should be verified across multiple domains and data sources. Additionally, further optimization of hyperparameters and tuning of model architectures may enhance the performance of some models. This study can be further extended and the further research can be carried out in the following direction.

- Fine tuning BERT: While BERT demonstrated exceptional performance in sentiment analysis, further exploration can be done to fine-tune the BERT model specifically for e- commerce review data. Fine-tuning involves adapting the pre-trained BERT model to better suit the characteristics and language

patterns of the target domain, which may lead to even higher accuracy and efficiency.

- Ensemble methods: Investigating the possibility of combining multiple models via ensemble techniques. Ensemble learning can help improve the overall performance and robustness for the task of sentiment analysis by leveraging the strength of individual models.
- Multi-lingual Sentiment Analysis: Extend the study to accommodate multi-lingual sentiment analysis if your e-commerce platform has reviews in multiple languages. Investigate pre-trained models and techniques that can handle sentiment analysis in different languages effectively.
- Real-time Sentiment Analysis: Explore the implementation of real-time sentiment analysis on live data streams. Investigate how the model can be deployed and updated in a production environment to provide up-to-date sentiment insights for businesses.
- Deployment and scalability: Consider the practical implications of deploying the sentiment analysis models in a real-world setting. Address challenges related to model deployment, scalability, and integration into existing business processes.
- Handling of textual data: Addition to above points, there is also a room to handle modern day reviews that include more than textual data. Not restricted to ecommerce data but also different sources that generate huge amount of data daily for example Twitter, and Facebook. Different reviews contain emoticons or in simple terms symbols like (☺, ☹) assist the user to express their emotions in a more impactful manner. Moreover, the models also need to take into consideration the stress that user tend to give on certain words that is directly proportional to the strength of their emotion. For example, words like ‘greatttt!’ and ‘niiice!!!’. Usually, they might not have a proper meaning but should be further processed to help identify the sentiment that is associated with the sentence.
- Evaluation methods: One of the matrices used to evaluate the models is the confusion matrix. Going ahead, this performance can also be evaluated using statistical tests like, ANOVA, Wilkinson test and t-test can be included to evaluate the performances of the employed systems.

REFERENCES

[1] P. Aline Bessa. (2022, March 17). Lexicon-Based Sentiment Analysis: A Tutorial. Retrieved from https://www.knime.com/blog/lexicon-based-sentiment-analysis

[2] Andrienko, N., Andrienko, G., Miksch, S., Schumann, H., & Wrobel, S. (2021). A theoretical model for pattern discovery in visual analytics. Visual Informatics, 5(1), 23–42.

[3] Bengfort, B., Bilbro, R., & Ojeda, T. (1891–1894). Enabling language-aware data products with machine learning. Sebastopol, CA: O'Reilly Media Inc.

[4] Chen, N. (2022). E-Commerce Brand Ranking Algorithm Based on User Evaluation and Sentiment Analysis. Frontiers in Psychology, 13, 907818.

[5] Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. Retrieved from https://arxiv.org/pdf/1409.1259

[6] Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Retrieved from https://arxiv.org/pdf/1412.3555

[7] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[8] Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery. In Proceedings of the twelfth international conference on World Wide Web - WWW '03 (p. 519). New York, NY: ACM Press.

[9] Britz, D. (2015). Understanding Convolutional Neural Networks for NLP. Retrieved from https://dennybritz.com/posts/wildml/understanding-convolutional-neural-networks-for-nlp/

[10] 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 (pp. 4171–4186). Stroudsburg, PA: Association for Computational Linguistics.

[11] Eke, C. I., Norman, A. A., Shuib, L., & Nweke, H. F. (2020). Sarcasm identification in textual data: systematic review, research challenges and open directions. Artificial Intelligence Review, 53(6), 4215–4258.

[12] Elzeheiry, S., Gab-Allah, W. A., Mekky, N., & Elmogy, M. (2023). Sentiment Analysis for E-commerce Product Reviews: Current Trends and Future Directions.

[13] Cramer-Flood, E. (2020). Ecommerce Decelerates amid Global Retail Contraction but Remains a Bright Spot. Retrieved from https://www.insiderintelligence.com/content/global-ecommerce-2020

[14] Fausett, L. (1994). Fundamentals of neural networks: Architectures, algorithms, and applications. Englewood Cliffs, NJ: Prentice-Hall.

[15] Ghose, A., & Ipeirotis, P. G. (2011). Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. IEEE Transactions on Knowledge and Data Engineering, 23(10), 1498–1512.

[16] Haque, T. U., Saber, N. N., & Shah, F. M. (2018). Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE International Conference on Innovative Research and Development (ICIRD) (pp. 1–6). IEEE.

[17] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

[18] Hota, H. S., Sharma, D. K., & Verma, N. (2021). Lexicon-based sentiment analysis using Twitter data. In Data Science for COVID-19 (pp. 275–295). Elsevier.

[19] Iqbal, A., Amin, R., Iqbal, J., Alroobaea, R., Binmahfoudh, A., & Hussain, M. (2022). Sentiment Analysis of Consumer Reviews Using Deep Learning. Sustainability, 14(17), 10844.

[20] Nabi, J. (2019). Recurrent Neural Networks (RNNs): Implementing an RNN from scratch in Python. Retrieved from https://towardsdatascience.com/recurrent-neural-networks-rnns-3f06d7653a85

[21] Abah, J. O. (2021). Sentiment Analysis of Amazon Electronic Product Reviews using Deep Learning. Master's dissertation, Dublin Business School. Retrieved from https://esource.dbs.ie/handle/10788/4291

[22] Read, J. (2005). Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. 43-48. Retrieved from https://aclanthology.org/P05-2008

[23] Kaur, J., & Sidhu, B. K. (2018). Sentiment Analysis Based on Deep Learning Approaches. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1496–1500). IEEE.

[24] Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746–1751). Stroudsburg, PA: Association for Computational Linguistics.

[25] Li, H., Ma, Y., Ma, Z., & Zhu, H. (2021). Weibo Text Sentiment Analysis Based on BERT and Deep Learning. Applied Sciences, 11(22), 10774.

[26] Liu, B. (n.d.). Sentiment Analysis and Subjectivity.

[27] Liu, Q., Wang, J., Zhang, D., Yang, Y., & Wang, N. (2018). Text Features Extraction based on TF-IDF Associating Semantic. In 2018 IEEE 4th International Conference on Computer and Communications (ICCC) (pp. 2338–2343). IEEE.

[28] Aithal, C. T. M. (2021). On Positivity Bias in Negative Reviews.

[29] Phi, M. (2018). Illustrated Guide to LSTM’s and GRU’s: A step by step explanation. Retrieved from https://towardsdatascience.com/illustrated-guide-to-lstms-and-grus-a-step-by-step-explanation

[30] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Retrieved from http://arxiv.org/pdf/1301.3781v3

[31] Mohammad, S. M., & Turney, P. D. (2013). CROWDSOURCING A WORD-EMOTION ASSOCIATION LEXICON. Computational Intelligence, 29(3), 436–465.

[32] Ni, J., Li, J., & McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 188–197). Stroudsburg, PA: Association for Computational Linguistics.

[33] Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1–135.

[34] Rambocas, M., & Pacheco, B. G. (2018). Online sentiment analysis in marketing research: a review. Journal of Research in Interactive Marketing, 12(2), 146–163.

[35] Kurban, R. (2019). CNN Sentiment Analysis: Use Convolutional Neural Networks to Analyze Sentiments in the IMDb Dataset. Retrieved from https://towardsdatascience.com/cnn-sentiment-analysis-9b1771e7cdd6

[36] Paknejad, S. (2018). Sentiment classification on Amazon reviews using machine learning approaches. Master's dissertation, KTH Royal Institute of Technology. Retrieved from https://web.archive.org/web/20200610153943/http://kth.diva-portal.org/smash/get/diva2:1241547/FULLTEXT01.pdf

[37] Srinivas, A. C. M. V., Satyanarayana, C., Divakar, C., & Sirisha, K. P. (2021). Sentiment Analysis using Neural Network and LSTM. IOP Conference Series: Materials Science and Engineering, 1074(1), 012007.

[38] Chevalier, S. (2022). Global retail e-commerce sales 2014-2026. Retrieved from https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/

[39] Tan, J. Y., Chow, A. S. K., & Tan, C. W. (2022). A Comparative Study of Machine Learning Algorithms for Sentiment Analysis of Game Reviews. The Journal of The Institution of Engineers, Malaysia, 82(3).

[40] Thompson, M., Duda, R. O., & Hart, P. E. (1974). Pattern Classification and Scene Analysis. Leonardo, 7(4), 370.

[41] Tripathy, A., & Rath, S. K. (2017). Classification of Sentiment of Reviews using Supervised Machine Learning Techniques. International Journal of Rough Sets and Data Analysis, 4(1), 56–74.

[42] turbolab. (2021). Feature Extraction in Natural Language Processing. Retrieved from https://turbolab.in/feature-extraction-in-natural-language-processing-nlp/

[43] Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.

[44] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. Retrieved from http://arxiv.org/pdf/1706.03762v5

[45] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731–5780.

[46] Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10), 1550–1560.

[47] Wikipedia. (2023). Recurrent neural network. Retrieved from https://en.wikipedia.org/w/index.php?title=Recurrent_neural_network&oldid=1162017246 (Accessed 17 July 2023).