Sentiment Analysis for Nykaa website using SVM, Twin SVM, Naïve Bayes, and LSTM

A DISSERTATION

THESIS SUBMITTED IN PARTIAL FULFILMENT OF REQUIREMENTFOR THE AWARD OF THE DEGREE
OF

Master of Technology
In
Information Systems

Under the guidance of

Dr. RITU AGARWAL
(Professor)

Submitted By
PARUL GUPTA
(Roll No. 2K21/ISY/18)

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahabad Daulatpur, Main Bawana Road, Delhi-110042

June 2023

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I/We, Parul Gupta, 2K21/ISY/18 student of MTech (IT-ISY), hereby declare that the project dissertation titled "**Sentiment Analysis for Nykaa website using SVM, Twin SVM, Naïve Bayes, and LSTM**" which is submitted by me to the Department of Information Technology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship, or other similar title or recognition.

Place: Delhi   Name- Parul Gupta

Date: Roll No. 2K21/ISY/18

INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled “Sentiment Analysis for Nykaa website using SVM, Twin SVM, Naïve Bayes, and LSTM” which is submitted byParul Gupta, 2K21/ISY/18 Information Technology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this university or elsewhere.

Place: Delhi **SUPERVISOR**

Date: **Dr. RITU AGARWAL**

ACKNOWLEDGEMENT

I would like to express my gratitude to my guide Dr. Ritu Agarwal without her valuable input and guidance this report wouldn’t come to fruition. I would also like tothank my parents for their constant support throughout the process. Also, I’d like to thank myclassmates who kept my spirits up when I was down and helped me whenever I was stuck somewhere.

PARUL GUPTA

ABSTRACT

Many websites such as Twitter, blogs, and e-commerce sites are popular nowadays, which display a tremendous amount of information about various topics, such as reviews and discussions on events. To manually try to understand the essential opinion regarding something is very time-consuming. Opinion mining or sentiment analysis is used, which automatically analyzes text using machine learning approaches and tries to give the idea of people's sentiments regarding a topic or product. Nykaa is one of the leading online shopping websites, where a large amount of information is available. The paper uses sentiment analysis on the Nykaa dataset, where we train the machine to generate the ability to define the overall opinion about a particular context, such as negative or positive. The input data is first pre-processed, then this minimalized data is converted into vector space as the machine understands numbers, not text, using sentiment score. Then machine learning algorithms such as Naïve Bayes, SVM, Twin SVM, and LSTM are applied, and results are evaluated and compared where LSTM performs better.

.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLTK :	Natural Language Toolkit
NLP :	Natural Language Processing
SA :	Sentiment Analysis
NB :	Naïve Bayes
SVM :	Support Vector Machine
LSTM :	Long Short-Term Memory
KNN :	K - Nearest Neighbor
BRM :	Brand Reputation Management

CHAPTER 1

INTRODUCTION

Sentiment Analysis is the natural language processing for understanding an attitude, thought, or judgment of the public about a particular product or topic. With the increased usage of e-commerce, the number of sellers selling their products has increased, thereby asking to review their products. A massive number of reviews exist for a product that is mainly in demand, and because of this, it is difficult for an individual to read them and decide whether to purchase the product. This problem can be solved using sentiment analysis, where one can easily understand the emotion of the content and make a purchase decision accordingly.

It can be difficult to accurately represent the precise intent and intricacies of written text due to language's complexity and nuance. The meaning of a sentence might vary depending on context, irony, ambiguity, and cultural connotations. Additionally, it is challenging to develop models that can consistently and accurately anticipate ideas due to the variety and heterogeneity of human expression.

NLP is being used to address these issues in projects in fields like sentiment analysis, language translation, text summarization, and question-answering systems. We seek to enhance the computer's comprehension of language and capacity to properly predict human thinking, making it a valuable tool in a variety of applications and fields.

1.1 Introduction of python

In my work, I have used Python. Python is a high-level programming language easy to read, understand and write. It is a robust and versatile programming language. It is one of the most used languages as it uses an open-source library. In Python, testing and debugging are fast as an interpreter is used instead of a compiler.

Python is used in many places such as web development, data analysis, software development, artificial intelligence, and System scripting. It can work on various platforms such as Windows, Raspberry, R, etc. Python is very easy to write as compared to other languages as it is similar to English. Some key features of Python are simple and expressive syntax, interpreted language, cross-platform compatibility, and a large standard library.

1.2 Introduction of Anaconda and Jupyter

Anaconda is an open-source, package manager and, distribution platform for the Python programming language. With Anaconda comes the package manager Conda which allows users to install, update and manage dependencies and packages.

Jupyter is a web-based open-source environment that allows users to share notebooks. These notebooks allow you to take notes, draw charts and, execute commands. Various key components of Jupyter are Notebook interface which consists of cells where one can create, write or update the code. Jupyter Notebook supports various languages such as R, Python, and others. It allows interactive visualization as it has various libraries for that such as matplotlib, seaborn, etc. It's very easy to share Jupyter's notebook. Jupyter assures interactive and reproducibility properties in others' work. One can easily identify what the other person was trying to write or produce with that code.

1.3 What is Sentiment Analysis?

Sentiment Analysis is part of Natural Language Processing (NLP) where we try to train the machine in such a way that it generates the ability to define the overall opinion about certain contexts such as negative, neutral, or positive. By analyzing sentiments using sentiment analysis one can analyze the mass opinion such as feedback, customer reviews, and opinion on the available data

Sentiment analysis is a computational process that involves identifying and categorizing opinions expressed in text, enabling the determination of the writer's positive, negative, or neutral attitude toward a specific topic or product. For instance, imagine you're interested in purchasing a product. Before making a purchase, you typically seek feedback from other customers to gauge whether the product is good or bad. This manual analysis involves examining the feedback provided by customers. Now, consider this process at a company level, where the number of customers is not just one or a few but rather millions. In such cases, companies employ sentiment analysis techniques to understand what their customers truly think about their products. This allows them to assess whether their product is performing well in the market or not.

1.4 Sentiment Analysis Classification

Classification is done based on polarity categorization nonpolar | negative | positive, which isfrequently studied in sentiment analysis. An analysis of sentiment is often conducted at one of the following levels: the document level | the attribute level | sentence level. You can conduct sentiment analysis using two approaches, namely machine learning and lexicon-based analysis.

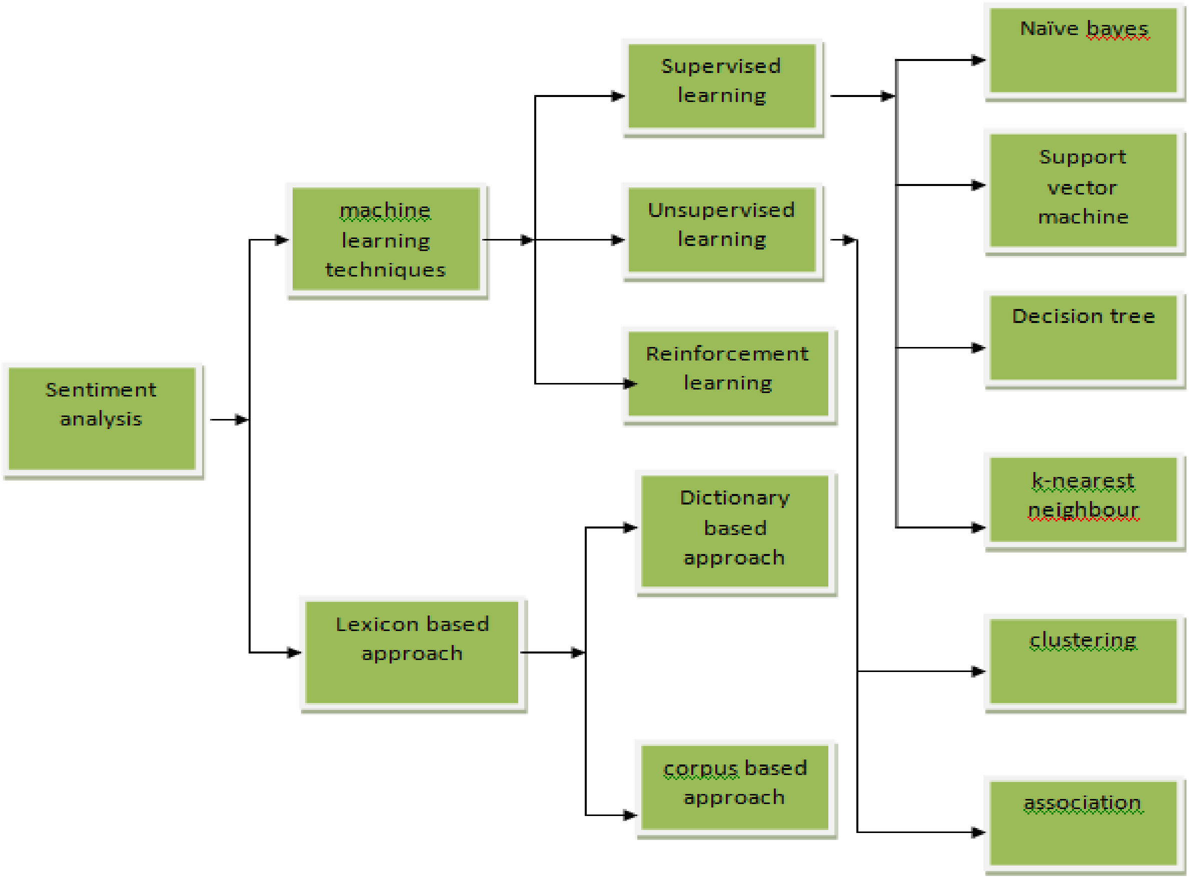

Figure 1: Sentiment classification techniques

Sentiment analysis is a highly researched field that encompasses various classification techniques. It is widely recognized as one of the most popular areas of study. The classification of sentiment analysis techniques can be broadly divided into two types: machine learning-based approaches and lexicon-based approaches. Lexicon-based approaches primarily concentrate on identifying negative and positive terms and can be further categorized into two types: dictionary-based and corpus-based approaches. On the other hand, machine learning approaches focus on two main techniques: supervised and unsupervised methods. Additionally, supervised techniques within machine learning can be further classified, which we will explore in the subsequent section.

1.4.1 Machine Learning techniques

Text classification in sentiment analysis is commonly accomplished using machine learning techniques. To train a model, a set of labeled training data records is utilized, which are then employed to create a predictive model. Each record is assigned a label corresponding to a specific class. When a new, unlabeled record is provided to the model, it is then assigned a label representing one of the different classes. In sentiment analysis, these classes typically include positive, negative, and neutral. The neutral class often represents mixed opinions and is not as frequently considered. Machine learning techniques in sentiment analysis can be broadly categorized into three types: supervised learning techniques, unsupervised learning techniques, and reinforcement techniques.

Supervised learning

Different classification techniques are employed in the field of machine learning to categorize unlabeled data. These methods trained the dataset using several classifiers. Naive Bayes, Support Vector Machine, KNN, and Decision Tree are examples of machine learning classifiers. These fall under the category of supervised machine learning classifiers, which need training data as a prerequisite. We do have multiple data points that explain features/variables and target variables in supervised machine learning. To predict the outcome variable from the predictor variable is the goal of supervised learning. Automating time-consuming or expensive manual tasks is the aim of supervised learning. Take "doctor's diagnosis" as an example. And we can forecast things like "Will a customer click on an ad or not" in the future. supervised education.

Unsupervised learning

It is a machine-learning strategy. Information that is not labeled or categorized is not used. It enables the algorithm to operate on that data without human supervision. Without any prior knowledge of the data, unsupervised learning aims to group unsorted data according to patterns, differences, and similarities. Since there is no human supervision in unsupervised learning, the machine is not trained with any data. As a result, a machine cannot independently discover the hidden structure in unlabeled data. Clustering and association are the next two divisions of unsupervised learning.

There is no data training given to the machine because instruction is offered. Therefore, a machine cannot identify the hidden information in unlabeled data.

Reinforcement learning

It is a machine learning technique that is based on feedback where agents take an action in the current environment and then learn from the outcome of that action. For every correct action reward is given and for every wrong move penalty is applied. As such no human interaction is required in reinforcement learning for training the model to how to behave in a particular environment to maximize the rewards. Reinforcement Learning is different from supervised and unsupervised machine learning where the dataset is divided into training and testing datasets and the model is trained using the training dataset and later accuracy is checked using the testing dataset but in reinforcement learning no dataset is used instead it learns from the experience by taking action in the given environment.

1.4.2 Lexicon-Based Approach

The lexicon-based approach is a technique for sentiment analysis in unsupervised learning. It makes use of a lexicon that lists the sentiment polarity—positive, negative, or neutral—as well as the related opinion guidelines—of words and phrases. This method makes use of the dictionary to find the opposites and equivalents of sentimental terms, facilitating the examination of sentiment. The lexicon-based approach can be further divided into two main approaches: the dictionary-based approach and the corpus-based approach.

Dictionary-based approach

The main objective of the dictionary-based sentiment analysis method is to create a thorough list of opinionated words. The dictionaries used in this method often include both words with positive and negative connotations. The dictionary-based method has a reasonably simple procedure.

The dictionary is built around these terms as its foundation. The sentiment dictionary is expanded if any further synonyms or antonyms are discovered by adding them to the wordlist.

The dictionary-based approach uses manually gathered sentiment words and iteratively adds synonyms and antonyms to the lexicon to provide a simple methodology for sentiment analysis. This method lays the groundwork for additional sentiment analysis and classification in text data, allowing researchers to gain insightful knowledge from massive amounts of textual data.

Corpus-based approach

The corpus-based approach to sentiment analysis has two basic goals: it finds new words with strong opinions from the corpus and builds a sentiment dictionary from existing words in the corpus.

By examining a corpus of text data, the corpus-based approach can be used to identify new opinionated words, which is one of its main applications. This method entails reading the text to find words that express emotion or opinion. By comparing these words to a list of known sentiment words, researchers can identify new words that exhibit sentiment but may not be present in the initial sentiment dictionary. This contributes to boosting sentiment analysis's coverage and broadening the vocabulary of sentiment words.

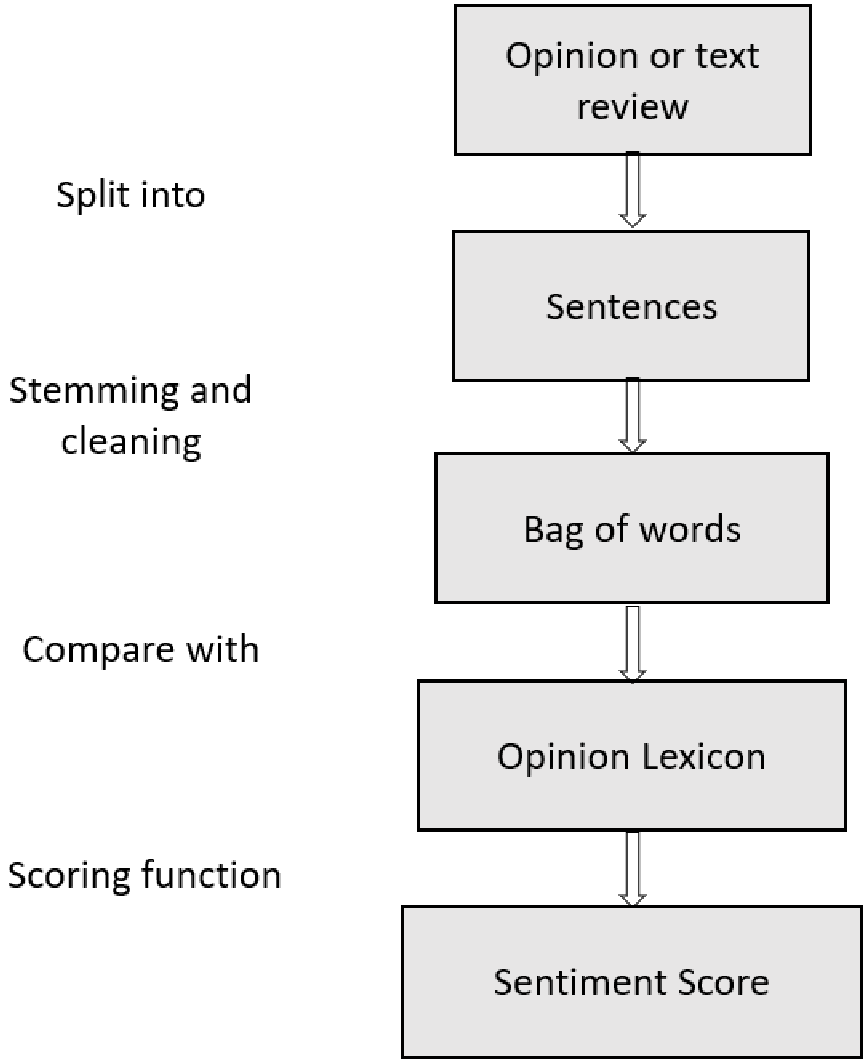

Figure 2: Flow diagram of lexicon based approach

1.5 Need for sentiment analysis

1.5.1 Industry development

From a business standpoint, it is obvious that unstructured data must be processed in part rather than in full to be useful. Sentiment analysis is essential in this context because it enables businesses to get knowledge from the data they already have. Industries can learn a great deal about the opinions, preferences, and levels of customer satisfaction by examining the feelings expressed in customer feedback, reviews, and other textual data.

Sentiment analysis can be useful for a variety of industries. Sentiment analysis, for instance, can be used to determine how customers feel about certain dishes, the quality of the service they receive, or their overall dining experiences. The menu selections, customer service, and general consumer happiness may all be enhanced with the help of this data.

1.5.2 Demand for research

Due to its study emphasis on evaluation, classification, and opinion analysis, sentiment analysis has seen substantial growth. Text extraction, artificial intelligence, machine learning, natural language processing (NLP), voting applications, and automated content analysis are a few of the computer-related areas that have benefited from improvements. These innovations make it possible for sentiment analysis systems to quickly process and decipher massive amounts of text input, eliminating the need for manual labeling. Utilizing these techniques, sentiment analysis offers companies insightful information from consumer feedback, assisting in decision-making and improving customer experiences.

1.5.3 Decision Making

People share information and thoughts in the current digital era via a variety of web applications, blogs, and social media platforms. A certain approach to data analysis is needed to get usable results to pull pertinent information from this enormous social web. For businesses, it would be difficult to obtain user feedback on the finest goods and conduct daily polls. However, consumer reviews and opinions are vital for assisting others in making educated selections and for advancing the fields of business and study.

1.5.4 Recognizing contextual cues

Due to its intrinsic complexity, understanding human language presents major difficulties for machines. Comprehending a language can be challenging due to its intricacies, slang, spelling mistakes, and cultural variations. To close the communication gap between humans and machines, innovative methods must be developed. These solutions should make it easier for people to understand language produced by computers. We can work to develop better systems that facilitate a greater human understanding of machine language by utilizing technologies like natural language processing (NLP) and machine learning.

1.5.5 Online Marketing

The extensive use of online marketing by big businesses and sectors is another important factor influencing the quick development of sentiment analysis. These organizations frequently collect online customer feedback about their brands, goods, social media posts, blogs, and events. Sentiment analysis is a useful marketing tool that enables businesses to learn more about the attitudes and perceptions of their customers. Businesses can evaluate the effectiveness of their marketing activities, comprehend client preferences, and make data-driven decisions to improve their marketing strategy by analyzing the sentiment expressed by users. Sentiment research is therefore essential for maximizing marketing efforts and elevating general client pleasure.

1.6 Applications of sentiment analysis

In the area of natural language processing (NLP), sentiment analysis has several uses. There is a rising need for social media data in sentiment analysis due to the importance of understanding emotions and sentiments. To streamline their procedures, many large corporations have used sentiment analysis tools. Here are a few significant uses for sentiment analysis.

1.6.1 Monitoring social media

Sentiment analysis aids in keeping track of and examining public opinion, consumer reviews, and brand mentions on social media platforms. It lets businesses comprehend how consumers feel about their brands or goods and base decisions on that understanding.

1.6.2 Analysis of customer feedback

The examination of consumer evaluations, ratings, and comments regarding goods or services is frequently done using sentiment analysis. It enables companies to better understand customer happiness, pinpoint areas for development, and make data-driven choices that improve customer experiences.

1.6.3 Market Analysis

Sentiment analysis is a critical component of market research since it examines consumer attitudes and opinions about particular goods, companies, or market trends. It aids companies in comprehending market dynamics, identifying consumer preferences, and gaining insightful information for product creation and marketing plans.

1.6.4 Financial Analysis

To understand market sentiment and identify trends, financial markets utilize sentiment analysis to examine news stories, social media messages, and other textual data. It helps traders and investors make wise choices and comprehend how market mood affects stock prices and investment opportunities.

1.6.5 Brand reputation management(BRM)

Sentiment research helps manage and uphold a company's brand reputation by looking at sentiment across numerous channels. It assists in spotting possible PR disasters, monitoring brand sentiment over time, and taking preventative action to deal with any unfavorable sentiment.

1.6.6 Political Analysis

Sentiment analysis is used in political campaigns to analyze voter attitudes, monitor public opinion, and assess voter preferences. Understanding public opinion enables politicians and campaign managers to adjust their plans.

The use of textual data for better decision-making, consumer engagement, and brand management is made possible by the widespread use of sentiment analysis in a variety of industries.

CHAPTER 2

LITERATURE SURVEY

In this research, a summarized review of multiple works related to sentiment analysis is considered. Various approaches have been seen using different data sets, algorithms, and pre-processing techniques to enhance the efficiency of sentiment analysis. The most basic step of sentiment analysis is discrimination between subjective text and objective text. The subjective text contains some opinion whereas objective text is just a fact it carries no emotion.

Example: Objective – Apple launched a new iPhone on the 7th of September (it's a fact)
Subjective – iPhone 13 is a very beautiful phone (opinion).

Sentiment analysis is performed on the second example which is subjective by nature as it represents opinion regarding phones. Opinions can be categorized into various classes most analyzed in our survey are neutral, positive, and negative.

Sentiment Analysis is part of Natural Language Processing (NLP) where we try to train the machine in such a way that it generates the ability to define the overall opinion about certain contexts such as negative, neutral, or positive. Data is taken and various pre-processingsteps are applied and tagging of data is done to define its orientation then this minimalized data is converted into vector space as the machine understands numbers not text using sentiment score or weightage to each word or frequency count method. Then various machine learningalgorithms are applied, and results are evaluated.

Understanding the emotion behind the text is a complex process many emotions such as sarcasm, and irony are hard to determine by machine for example word jaguar whether it's a car or an animal is a complex process to identify and a lot of new methods such as deep learning, convolutional neural network, genetic algorithms and combination of old one are coming into scene to deal with implicit sentences such as 'Samsung phones are any given day better than oppo phone' this statement is negative but implicitly hence hard to determine its polarity.

Xing Fang and Justin Zhan have done sentiment classification on the sentence level and review level using various steps such as sentiment phrase identification that is negative of phrases, score computation, feature vector generation, and polarity categorization [4]. POS tagger isthe part of speech tagging of the word in the speech category such as nouns, pronouns, verbs, adverbs, adjectives, conjunction, preposition, and many others [1]. It helps in having a better understanding of the relation between words in a sentence. POS tagging of words is done to extract subjective content from the text. The subjectivity of content is words that signify the sentiment or represent some opinion. Sentiment score computation is done using the formula:

Where t is the number of tokens and the occurrence of t is the number of times that token appears in theith review. Here balancing is done using the formula:

Here no tagging is required as we already have star tagging on a scale of 1 to 5 where the value of score below 3 is negative, the value above 3 is positive and the value of score 3 is neutral. Sentence level categorization for tagging or labeling we use the bag-of-words model which contains a list of positive words and negative words and the score is computed based on the number of positive or negative word counts in a statement and whose count is greater than the sentiment of the statement.

The feature vector is developed on four features such as two hash tables, a tag, and averaged sentiment score. Where the hash table represents several bits to represent the number of words token and phrase tokens. Then F1-score is computed using:

The F1-score for sentence level is better as compared to the review level. In general, both showed apromising result, but as neutral class is taken into consideration in the review level its score is low as compared to the sentence level.

Performance is increased in sentiment analysis by pre-processing the data using various techniques [5]. Text preprocessing and then evaluating its performance has shown good results in work published by **Haddi.** Sentiment Analysis is done in three steps: features areextracted using transformation and filtering using chi-square statistics, SVM classifier on feature matrix is applied and computed using feature weighing methods feature frequency, feature presence, Term Frequency Inverse Document Frequency, and then performance is evaluated. These weights determine the importance of the feature by attaching weights.

Feature-based sentiment analysis can also be done such that it helps us to understand the opinion of the features of the product such as size, weight, and build quality extra [6]. This is sentiment analysis on the aspect level. Three-step processes for mining customer review: Finding out which product features customers have expressed an opinion about, finding out the number of positive or negative reviews for each feature, and result evaluation.Using POS tagger words in sentences are tagged. Using **association mining** frequent features are extracted. Features that are of no use are removed using compactness pruning and redundancy pruning.

Compactness pruning is the removal of unlikely features by calculating the distance between words inthe feature phrase such as “I had searched for a digital pen for 3 months.” “This is the best digital pen in the market” and “The pen does not have a sharp digital nib” Here digital pen appeared together in two sentences hence it is a compact feature used frequently

Redundancy pruning uses p-support (pure support) such that for a feature that is either a noun or noun phrase no superset must exist for example pen or digital is redundant if a digital pen exists hence, they should be pruned.

While predicting orientation when the feature review is straight negative or positive then no issue is there but if multiple feature review is given in a single statement and orientation comes out to be neutral that it will result in the loss of information for that we use distance such that opinion words close to that feature represents the opinion for that feature. For example, ‘though the touch screen of the phone is amazing, the camera is not up to the mark’ here 1positive word and 1 negative word results in the neutral hence if we look for a particular feature then we can classify it as for touch screen it's positive and for a camera it's negative.

Accuracy improves as we use compactness pruning and p-support as compared to only using association mining. The average accuracy for predicting sentence orientation is 84%. This shows a method of analyzing reviews using a seed list of adjectives is effective. The limitation of the paper is itdidn’t consider the pronoun such as ‘it’ for example ‘it has a very good touchpad’ here this approach will fail to identify it. Apart from this, it gives promising results on adjectives, verb and all is not taken into consideration

Feature-based sentiment analysis helps us to understand the opinion of the features of the product, such as size, weight, and build quality. In the paper[5], association mining extracts frequent features, and compactness and redundancy pruning remove irrelevant features.

While performing Sentiment Analysis the impact of **feature extraction using TF-IDF** and n-gram on the SS Tweet dataset shows that TF-IDF performs better[6]. TF-IDF stands for Term Frequency – Inverse Document Frequency; itdefines the significance of the word. Calculated using the formula:

$$W_{a,b} = t'^{f_{a,b}} * \log(Nx/d'f_a) \qquad (1)$$

$W_{a,b}$ = word a within document b.
$t'f_{a,b}$ = Frequency of a in b.
$d'f_a$ = the number of documents containing a.
Nx = the total number of documents.

TF-IDF transforms the textual information into a vector form. Let us consider a document having 100 words. The word keyboard appears five times in the document; the term frequency will be 5/100=0.05. Suppose there are 25000 documents and 250 documents containing the term keyboard; the inverse document frequency will be 25000/250=100, and TF-IDF is equal to 0.05*100=5. The n-gram feature extraction value of n can be 1,2,3.. . Where n=1 is unigram, n=2 is bigram, and so on. In a statement "laptop build quality is very good," for unigram single word is considered in a sentence {'laptop', 'build',' quality',' is',' very',' good} whereas in bigram group of words is considered such as {'laptop build', 'build quality, 'quality is', 'is very',' very good'}.

Chaining classifiers such that two or classifiers are chained one after the other for better performance is used in paper [8] where the first classifier is used to classify text into three categoriesopinion oriented such as polar, unnecessary, and neutral. Then in the second stage data under polar categorization is used and further classified as positive and negative.

Removing skewness from the dataset can also enhance the performance. That is balancing the unbalanced training dataset using sampling processes such as under-sampling and over-sampling.
Under sampling – eliminating the objects from the majority of
Over sampling - increasing the objects in the minority class.

The dataset used for the first classification irrelevant|polar|neutral discrimination has shown thataccuracy has been improved to an extent as we have reduced the degree of skewness [8].

The average accuracy of SMO has increased by 4 to 5% using under-sampling and over-sampling process. In the second classifier, SVM gives a consistent result, but the overall accuracy is less as compared to the first classifier.

Deep neural networks are better for text classification as they have more hidden layers than neural networks. Automatic feature extraction makes them faster and more efficient[7]. The paper[8] uses LSTM for sentiment analysis on IMBD and Amazon datasets. It performs well due to its ability to handle large amounts of information

CHAPTER 3

PROBLEM STATEMENT

Nykaa.com is one of the emerging shopping sites entitled the first Indian unicorn startup headed by a female in 2020. Nykaa sells various products, from cosmetics to clothes to general wellness, and hence, many opinions or reviews are available, which we have analyzed in this research. Here we are trying to understand the mass view of the website, which in turn will help the customers, users, and buyers decide on the buy. E-commerce websites display the products using images, so it is hard to determine the quality of their products or services. User ratings of the products help us understand people's sentiments as to whether they like the website based on the quality of the products offered. Therefore, a review gives us an idea of the product's or website's positives and negatives.

The main intention of this research is to perform a sentiment analysis on the Nykaa website using various algorithms known to perform well in the case of natural language processing, such as LSTM, SVM, Twin SVM, and Naïve Bayes. Nonpolar, positive, or negative sentiments are associated with textual content, which helps measure an article's feelings.

Sentiment analysis can be utilized to understand public opinion about a product, movie, or sentiment regarding some topic of conversation between two persons. One can capitalize on this analysis by understanding the public demand, opinion, or attitude and taking measures to meet the needs.

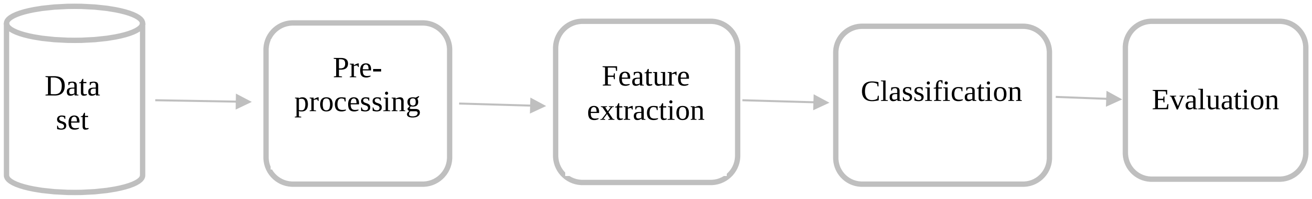

Figure 3: Flow chart

The basic sentiment analysis is done by collecting the data, applying pre-processing steps and feature extraction to make data machine compatible then applying classification algorithms and evaluating the performance.

3.1 MOTIVATION

Nykaa is one of the fastest-growing e-commerce sites these days which is used by a large number of people in India. Nykaa is the first cosmetic line that launched its IPO in such a short period hence capturing the huge market. As one of the frequent users of the website myself, I decided to do a sentiment analysis on Nykaa to understand the overall feedback of the website. Due to being one of the most trending websites nowadays a lot of feedback, and customer reviews are available about the delivery services of the website, product quality, and authenticity. A lot of reviews about the website are available over social media, blogs, or the App Store and to manually go through each review by a user is very time-consuming therefore sentiment analysis is done of the available data to understand the sentiment of public or mass opinion.

3.1.1 Perspectives of consumers

Sentiment analysis is essential for enabling consumers to make wise decisions. In the past, people relied on their friends and family's feedback to learn about items. But now that social media and the internet have been developed, people may access a huge variety of reviews and opinions from many sources. This wealth of knowledge is really helpful for customers who are thinking about buying a particular product.[18]

Utilizing sentiment analysis, people can gauge the general attitude towards a product and use this knowledge to inform their choice. It assists in addressing the essential query of whether or not they ought to purchase specific goods. Usually, this decision-making process has a binary stance: either you buy the thing or you don't. However, users may find it impractical to manually sort through and analyze each viewpoint due to the vast number of information provided.

Automation is necessary to address this problem. Users can quickly analyze and evaluate massive amounts of data by using technology-driven sentiment analysis tools. These technologies can provide aggregated opinions and categorize them as either good or negative, supporting users in making well-informed judgments.

Users can more successfully sift through the deluge of information by automating the sentiment analysis process with the help of technology. It enables them to gain useful insights, which supports confident decision-making.

3.1.2 Perspectives of producer

In a dynamic system, decisions made by producers and consumers are interrelated. Producers closely observe consumer behavior while consumers express their thoughts online to inform their business decisions. Through this interaction, a multi-user ecosystem is created where consumer spending habits affect other consumers' decisions and motivate manufacturers to provide and sell goods.

The opinions and input of customers are very important in determining market trends. Customers have an impact on others' purchasing decisions by sharing their insights and advice. Positive comments and recommendations can generate talk about a product, increasing demand. Negative reviews, on the other hand, can put off potential customers. Collectively, consumer decisions impact a product's success or failure.

To gather knowledge of market trends and consumer preferences, producers actively analyze and learn from consumer feedback. Producers can modify their future business strategy by tracking the development of sales and consumer feedback. Through this feedback loop, they may improve the quality of their products, streamline their offers, and match their business strategies to market expectations. Consumer choices provide manufacturers with useful feedback that they may use to guide future product development, marketing plans, and business strategy.

In this multi-user system, choices made by both customers and producers help products and services communicate their impression and utility. The market environment is shaped by the ongoing feedback and decision-making between consumers and producers, which also has an impact on how products and services develop. Because of this dynamic interaction, customer preferences influence producer choices, which in turn influence consumer choices.

CHAPTER 4

IMPLEMENTATION AND METHODOLOGY

The main intention of this research is to perform a sentiment analysis on the Nykaa website using various algorithms known to perform well in the case of natural language processing, such as LSTM, SVM, Twin SVM, and Naïve Bayes. Nonpolar, positive, or negative sentiments are associated with textual content, which helps measure an article's feelings.

Sentiment analysis can be utilized to understand public opinion about a product, movie, or sentiment regarding some topic of conversation between two persons. One can capitalize on this analysis by understanding the public demand, opinion, or attitude and taking measures to meet the needs.

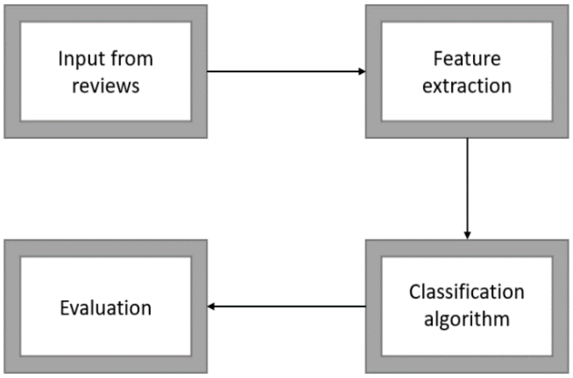

Figure 4: Block Diagram

4.1 Proposed Methodology

One can perform a sentimental analysis at various levels, including aspect, sentence, and document levels[9].

Document level: The sentiment of a complete document is evaluated into negative, positive, or neutral views based on which one can determine the reading for a product or service of a website.

Sentence level: When sentence-level sentiment analysis is done, the sentiment classifier classifies the statement as nonpolar, positive, or negative. It is used in review sentiment analysis or comment analysis.

Entity Level: Entity level sentiment analysis works on the aspect or feature of the product, which helps the customer to understand the liking or disliking of the different components of the product, like size, build quality, the weight of the camera, etc.

We can divide Aspect level sentiment Analysis as:

Aspect- description of quality, quantity, size, color, and all other product features.

Sentiment- positive or negative opinions about the features of a product.

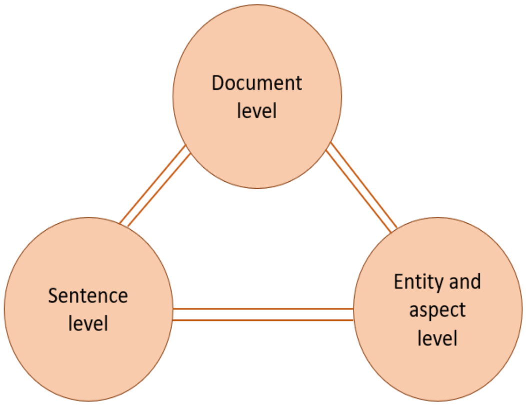

Figure 5: Different Levels

4.2 Data Analysis

The models used in this paper have been applied to the review dataset, downloaded from Kaggle (https://doi.org/10.34740/KAGGLE/DSV/2607279), where reviews are scrapped from the google play store, consisting of 194430 rows and two columns. The first column is the app review text representing the content, and the second column is the score binned sentiment consisting of the sentiment labels concerning to content column. For sentiment labels, reviews with a five-star rating are considered positive, three stars and four-star rating is considered neutral, and the rest are taken as negative. We have trained models using negative and positive comments. The skewness present in the dataset is removed to enhance the performance using the sampling process. A balanced dataset consists of 25000 positive and 17134 negative reviews, as shown in Fig. 6:

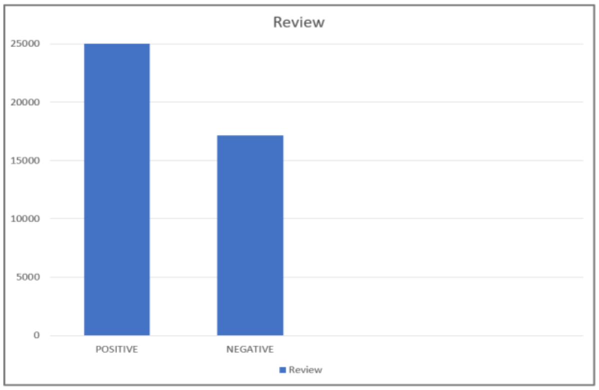

Figure 6: Histogram of the dataset

4.3 Data Pre-Processing

Stop-word removal: In this technique, we remove *stop-words* that are commonly used in the formation of English statements, such as *a, an, the, and from*, which are of no use as they contain no value concerning sentiment identification.

Removing Punctuations and Whitespaces: helps in reducing the size of training datasets like *commas, and full stops*, which have no value addition in opinion mining.

Removing Punctuations and Whitespaces: helps in reducing the size of training datasets like *commas, and full stops*, which have no value addition in opinion mining n=1 means unigram, i.e., a single word as a token, e.g., 'nice,' 'good,' etc.

n=2 means bigram, which is a combination of two terms as a token 'not appropriate,' 'poor build,' etc.[2]

This research uses the unigram tokenization model.

Lemmatization: Removal of inflection from a token and converting it into its original form, such as *loving* into love, *failing* into fail, and *passing* into a pass.

4.4 Classification Algorithm

Classification is done based on polarity categorization negative | positive frequently studied in sentiment analysis. Sentiment analysis is conducted on the sentence level using three popular algorithms: Naïve Bayes, SVM, Twin SVM, and LSTM.

Various pre-processing steps, as shown in Fig. 7, have been used to enhance the performance of models. We have used one hot encoding feature extraction technique on pre-processed data to convert text data into numerical data. The model is trained using classification algorithms, and the performance of the algorithms is evaluated and analyzed.

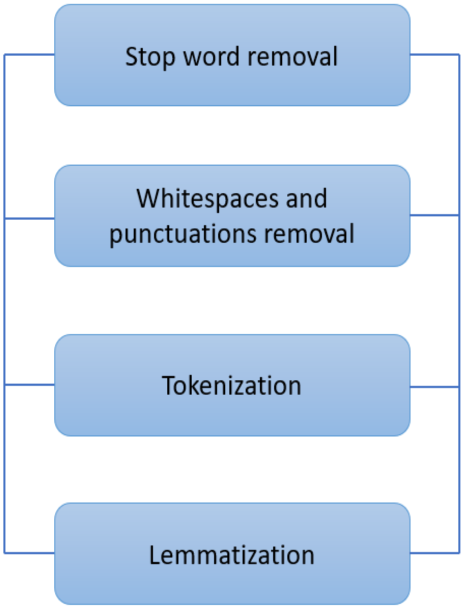

Figure 7: Pre-processing steps

4.4.1 Naive Bayes

It is a probabilistic algorithm that uses Bayes' Theorem to predict the label mentioned in the text and outputs the label with the maximum probability. It determines the likelihood of the label in the sentence[13] and then calculates the probability of each word using the Bayes formula.

$$P(X'/Y') = (P(X'/Y') * P(X'))/P(Y') \qquad (2)$$

Here X' is 'The movie was very bad,' and N can be positive or negative. We calculate the probability for both labels and consider the one whose value is greater.

P (the movie was very bad | positive) and

P (the movie was very bad | negative).

Here, we assume every word is independent of each other, and the class or label with higher probability is the correct output

4.4.2 Support Vector Machine

SVM is a non-probabilistic supervised machine learning algorithm that takes data points as input, plots them on a multidimensional space, and outputs the hyperplane, separating them into groups. This hyperplane is the decision boundary that helps decide where a tag belongs. The best hyperplane is one whose distance from the tags closest to the hyperplane is the maximum. When data is linearly separable, then the hyperplane is a straight line in two-dimensional space, but when data is non-linearly separable, we need to add more dimensions and to make this process less expensive, we use the kernel function. In comparison to a neural network, SVM shows better performance and speed for limited sample data

4.4.3 LSTM

LSTM (Long Short-Term Memory) is the advancement of RNN (Recurrent Neural Network) RNN suffers from the vanishing gradient problem and exploding problem where information is stored only for a short period. LSTM has overcome this problem as it can store data for longer. It uses the concept of forgetting information that is not required further and adds new details using three gates: forget gate, input gate, and output gate[12].

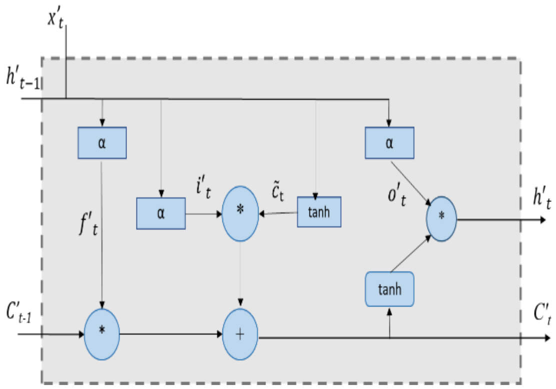

Figure 8: LSTM

$$i'_t = \alpha(w'_i[h'_{t-1}, x'_t] + b_i) \quad (3)$$

$$f'_t = \alpha(w'_t[h'_{t-1}, x'_t] + b_i) \quad (4)$$

$$o'_t = \alpha(w'_o[h'_{t-1}, x'_t] + b_o) \quad (5)$$

$$\tilde{c}_\mathrm{t} = tanh\,(w'_t[h'_{t-1}, x'_t] + b_i \quad (6)$$

i'_t→ input gate

f'_t→ forget gate

o'_t→ output gate

α→ sigmoid function.

w'_t→ weight for the given gates.

h'_{t-1}→ hidden state.

x'_t→ current input.

b_i→ biases for the given gate.

C'_{t-1}→ previous cell state.

On multiplying *f't*C'*$_{t-1}$ decision is taken whether to keep the information or discard it. Similarly, using (i'_t* $\tilde{c}$)+ (*f't*C'*$_t$), we can decide whether to update the previous value or add a new value and based on these calculations, we find h'_t using o'_t*tanh (C'_t).

4.4.4 Twin SVM

A machine learning algorithm called Twin SVM is made to address binary classification problems. Since two distinct support vector machines (SVMs) are trained, it is referred regarded as a "twin" method. An identical set of data is used to train each SVM, but they all have different goals.

The objective of Twin SVM is to identify two hyperplanes that can successfully discriminate between the two classes of data. By minimizing the classification errors and increasing the margin between the support vectors and the hyperplanes, these hyperplanes are optimized.

The initial SVM, also known as the positive hyperplane SVM, is concerned with accurately categorizing the positive instances in the data. It learns to construct a hyperplane with the greatest possible margin separating the positive examples from the negative one

The negative hyperplane SVM, the second SVM, is trained to accurately classify the negative cases. The goal is to choose a hyperplane that has the greatest margin of separation between the negative and positive examples.

Twin SVM corrects for any imbalances or asymmetry between the two classes in the data by training two SVMs independently. Different margins and decision boundaries are possible, which is advantageous in situations when the importance of the classes is unevenly distributed or when the data is unbalanced.

The final classification is established during prediction based on the decision values acquired from both SVMs. The anticipated class is designated as the one with the highest decision value.

CHAPTER 5

RESULT AND ANALYSIS

This analysis is entirely experimental, with the practical implementation described in detail. We have used Naïve Bayes, SVM, one of the most used machine learning algorithms in sentiment analysis, Twin SVM which uses two support vector machines, and LSTM, commonly used in text classification as it has memory. To comprehend the sentence, we need to consider the whole sentence instead of words. The accuracy of Naïve Bayes is less than SVM due to overlapping positive and negative comments and the accuracy of SVM is less than Twin SVM.

First, we pre-processed our data set. Then we used Naïve Bayes, SVM, Twin SVM, and LSTM to train our model. We used multinomial Naive Bayes, SVC linear kernel for SVM, and Twin SVM The Twin SVM gives a more accurate result than the Naïve Bayes and SVM algorithm in our analysis. For LSTM, we have used the embedding Layer - Which generates Embedding Vector for each input sequence, the LSTM layer, the dense layer and activation function, dropout layer to drop some neurons from previous layers to avoid overfitting problems. LSTM gives much better results on the dataset than SVM, Twin SVM, and Naïve Bayes.

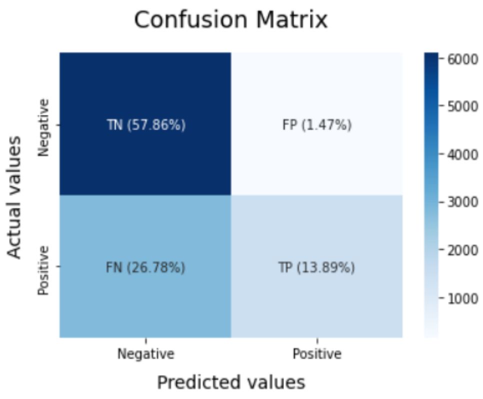

Figure 9: Confusion matrix for Naïve

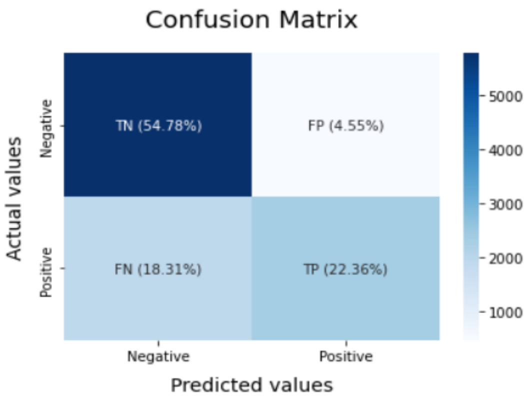

Figure 10: Confusion matrix for SVM

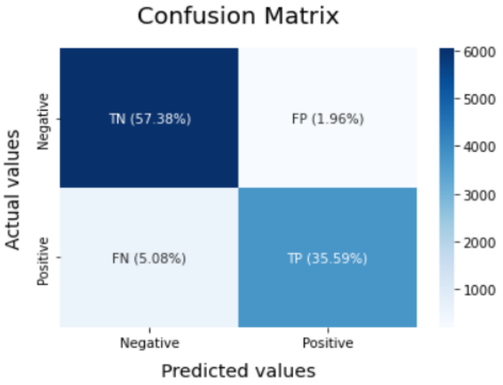

Figure 11: Confusion matrix for LSTM

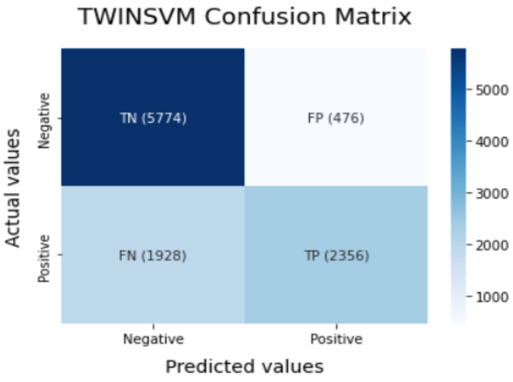

Figure 12: Confusion matrix for Twin SVM

Based on the confusion matrix, we calculate performance evaluating parameters such as recall, precision, f1-score, and accuracy. The comparative evaluation table for all three models is shown in Table.1.

	Naïve Bayes	**SVM**	**Twin SVM**	**LSTM**
Accuracy	0.72	0.77	0.88	0.93

Table 1: Comparative table

CHAPTER 6

CONCLUSION AND FUTURE WORK

A huge amount of data is available or generated these days on social media, e-commerce sites, by organizations, etc and this day may be structured or unstructured. To find out the polarity of data or emotion behind it sentiment analysis is performed.

In this thesis, we analyze user sentiment regarding online shopping sites. To check the accuracy, we evaluated the LSTM, linear SVM, Twin SVM, and Naïve Bayes algorithms on the online shopping site dataset. Many times Naïve Bayes and SVM outperform the neural network. However, Naïve Bayes' accuracy is 72% because of the overlapping negative and positive words. SVM's accuracy is 77% which can be improved by using feature extraction techniques such as TDF-IDF, bag-of-words, etc. Basic Twin SVM accuracy was 87% which is further improved by tuning the hyperparameters to 88%. LSTM is mainly used for text classification, and its accuracy is better than both naïve Bayes, Twin SVM, and SVM, which is 93%.

In the future, LSTM accuracy can be further improved by exploring the hybrid model, and its weights can be optimized using a genetic algorithm. SVM and TwinSVM accuracy can also be enhanced using feature extraction techniques, and we can analyze the performance of SVM, TwinSVM, and LSTM.

CHAPTER 7

REFERENCES

[1] A. Mridula and C. R. Kavitha, “Opinion Mining and Sentiment Study of Tweets Polarity Using Machine Learning,” 2018. doi: 10.1109/ICICCT.2018.8473079.

[2] D. Ramachandran and R. Parvathi, “Analysis of Twitter Specific Preprocessing Technique for Tweets,” in *Procedia Computer Science*, 2019, vol. 165. doi: 10.1016/j.procs.2020.01.083.

[3] P. Patil and P. Yalagi, “Sentiment Analysis Levels and Techniques : A Survey,” *International Journal of Innovations in Engineering and Technology*, vol. 6, no. 4, 2016.

[4] S. Dhar, S. Pednekar, K. Borad, and A. Save, “Sentiment Analysis Using Neural Networks: A New Approach,” 2018. doi: 10.1109/ICICCT.2018.8473049.

[5] B. Duncan and Y. Zhang, “Neural networks for sentiment analysis on Twitter,” 2015. doi: 10.1109/ICCI-CC.2015.7259397.

[6] J. Akilandeswari and G. Jothi, “Sentiment classification of tweets with non-language features,” in *Procedia Computer Science*, 2018, vol. 143. doi: 10.1016/j.procs.2018.10.414.

[7] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera, “Opinion mining and sentiment analysis on a Twitter data stream,” 2012. doi: 10.1109/ICTer.2012.6423033.

[8] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, “The impact of features extraction on the sentiment analysis,” in *Procedia Computer Science*, 2019, vol. 152. doi: 10.1016/j.procs.2019.05.008.

[9] M. Hu and B. Liu, “Mining and summarizing customer reviews,” 2004. doi: 10.1145/1014052.1014073.

[10] E. Haddi, X. Liu, and Y. Shi, “The role of text pre-processing in sentiment analysis,” in *Procedia Computer Science*, 2013, vol. 17. doi: 10.1016/j.procs.2013.05.005.

[11] X. Fang and J. Zhan, “Sentiment analysis using product review data,” *Journal of Big Data*, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0015-2.

[12] P. F. Muhammad, R. Kusumaningrum, and A. Wibowo, “Sentiment Analysis Using Word2vec and Long Short-Term Memory (LSTM) for Indonesian Hotel Reviews,” in *Procedia Computer Science*, 2021, vol. 179, pp. 728–735. doi: 10.1016/j.procs.2021.01.061.

[13] Institute of Electrical and Electronics Engineers, *Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies (NGCT) : October 14th-16th, 2016, Center for Information Technology, University of Petroleum and Energy Studies, Dehradun.*

[14] G. S. N Murthy, S. Rao Allu, B. Andhavarapu, M. Bagadi, and M. Belusonti, "Text based Sentiment Analysis using LSTM; Text based Sentiment Analysis using LSTM." [Online]. Available: www.ijert.org.

[15] Q. Tul *et al.*, "Sentiment Analysis Using Deep Learning Techniques: A Review," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, 2017, doi: 10.14569/ijacsa.2017.080657.

[16] University of Colombo School of Computing., IEEE Computer Society. Sri Lanka Chapter, IEEE Sri Lanka Section, and Institute of Electrical and Electronics Engineers, *International Conference on Advances in ICT for Emerging Regions : ICTer2012 : conference proceedings : 13th-14th December 2012, Colombo, Sri Lanka*. IEEE, 2012.

[17] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. ”Thumbs up?: sen- timent classification using machine learning techniques.” Proceedings of the ACL-02 conference on Empirical methods in natural language processing- Volume 10. Association for Computational Linguistics, 2002.

[18] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. ”Sentiment analysis algorithms and applications: A survey.” Ain Shams Engineering Journal (2014)