Unmasking the Best ML Algorithm for Sentiment Analysis

by

Md. Rifat Hossain
ID: CSE2101022083

Md. Bipul Hossain
ID: CSE2101022101

Jewel Biswas
ID: CSE2101022140

Supervised by
Arifur Rahaman

Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science and Engineering



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SONARGAON UNIVERSITY (SU)

January 2025

APPROVAL

The thesis titled "**Unmasking the Best ML Algorithm for Sentiment Analysis**" submitted by Md. Rifat Hossain (CSE2101022083), Md. Bipul Hossain (CSE2101022101) and Jewel Biswas (CSE2101022140) to the Department of Computer Science and Engineering, Sonargaon University (SU), has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering and approved as to its style and contents.

Board of Examiners

Arifur Rahaman
Assistant Professor & Coordinator,
Department of Computer Science and Engineering
Sonargaon University (SU)

Supervisor

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 1

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 2

(Examiner Name and Signature)
Department of Computer Science and Engineering
Sonargaon University (SU)

Examiner 3

DECLARATION

We, hereby, declare that the work presented in this report is the outcome of the investigation performed by us under the supervision of **Arifur Rahaman, Assistant Professor & Coordinator,** Department of Computer Science and Engineering, Sonargaon University, Dhaka, Bangladesh. We reaffirm that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

(Arifur Rahaman)
Supervisor

Signature

Md. Rifat Hossain
ID: CSE2101022083

Md. Bipul Hossain
ID: CSE2101022101

Jewel Biswas
ID: CSE2101022140

ABSTRACT

The enormous quantity of information that we encounter every day on the Internet originates from a variety of sources: from online review platforms to social media. Such an avalanche of data conceals a valuable opportunity for text analysis-the estimation of subjective opinions, and most importantly, their sentiment or polarity. Sentiment analysis and text classification are useful in drawing valuable information out of data by classifying text into certain classes based on their content. The paper presents a comparative study of five algorithms for text classification, namely Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), using a Twitter Comments dataset. We observed that a support vector machine (SVM) and Logistic Regression algorithms are doing better on all the metric scores in comparison to its peers, namely accuracy, precession, recall, and F1 score.

ACKNOWLEDGMENT

At the very beginning, we would like to express my deepest gratitude to the Almighty Allah for giving us the ability and the strength to finish the task successfully within the schedule time.

We are auspicious that we had the kind association as well as supervision of **Arifur Rahaman**, Assistant Professor & Coordinator, Department of Computer Science and Engineering, Sonargaon University whose hearted and valuable support with best concern and direction acted as necessary recourse to carry out our project.

We would like to convey our special gratitude to **Brig. Gen. (Retd) Prof. Habibur Rahman Kamal, ndc, psc,** Dean, Faculty of Science and Engineering for his kind concern and precious suggestions.

We are also thankful to all our teachers during our whole education, for exposing us to the beauty of learning.

Finally, our deepest gratitude and love to my parents for their support, encouragement, and endless love.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
SVM	Support Vector Machine algorithm
KNN	K-Nearest Neighbors algorithm

TABLE OF CONTENTS

Title Page No.

DECLARATION ……... iii

ABSTRACT …… iv

ACKNOWLEDGEMENT ……. v

LIST OF ABBREVIATION …….. vi

CHAPTER 1 **1 – 5**

INTRODUCTION TO MACHINE LEARNING

1.1 Introduction ……..……..... 1

1.2 Objectives ……...……... 1

1.3 Importance of Machine Learning (ML) ……..……... 1

1.4 Different types of Machine Learning (ML) …….…….. 2 – 3

1.4.1 Supervised learning ……...…….…….. 2

1.4.2 Unsupervised learning …….……. 2

1.4.3 Semi-supervised learning …...…….…. 2

1.4.4 Reinforcement learning …...…….…… 2

1.5 Features of Machine Learning (ML) …….……. 3

1.6 Application of Machine Learning (ML) …… 3 – 5

1.6.1 The role of Machine Learning (ML) in Healthcare ……... 3

1.6.2 Applying Machine Learning (ML) to Financial Markets ……. 4

1.6.3 Marketing & Advertising with Machine Learning (ML) …….. 4

1.6.4 Autonomous Vehicles using Machine Learning (ML) …… 4

1.6.5 Retail with Machine Learning (ML) ……. 4

1.6.6 Machine Learning (ML) for Energy ……. 5

1.7 Organization of the Thesis Book ……... 5

CHAPTER 2 **6 – 7**

LITERATURE REVIEW

2.1 Introduction …….…..... 6

2.2 Case Study and Overview of Related Works ……. 6

2.3 The Novelty of Our Approach ……... 7

CHAPTER 3 **8 – 9**

SUPERVISED AND UNSUPERVISED LEARNING

3.1 Introduction ……... 8

3.2 Difference between Supervised and Unsupervised Learning …… 8

3.3 Graphical Difference ……. 9

CHAPTER 4 **10 – 16**

4.1 Introduction ……... 10

4.2 How does Supervised Learning works? ……. 10

4.3 Steps involved in Supervised Learning …….. 11

4.4 Types of Supervised Learning 12 – 13

4.4.1 Regression …… 12

4.4.2 Classification …… 13

4.5 Supervised Machine Learning algorithms …… 13 – 15

4.5.1 Linear Regression …… 13

4.5.2 Logistic Regression …….. 14

4.5.3 Decision Trees ……. 14

4.5.4 Random Forests ……... 14

4.5.5 Support Vector Machine (SVM) …….. 14

4.5.6 K-Nearest Neighbors (KNN) ……... 14

4.5.7 Gradient Boosting …… 14

4.5.8 Naive Bayes Algorithm …… 14

4.6 Areas of Application …….. 16

4.7 Advantages of Supervised Learning …….. 16

4.8 Disadvantages of Supervised Learning ……. 16

CHAPTER 5 **17 – 20**

NATURAL LANGUAGE PROCESSING

5.1 Introduction ……..…….. 17

5.2 Machine learning (ML) for NLP …… 17

5.3 Supervised Learning for NLP and Text Analytics …… 18

5.4 Phases of Natural Language Processing (NLP) …… 19

5.5 Technologies related to Natural Language Processing (NLP) ……... 19

5.6 Contexts of Practical Application of NLP …… 20

5.7 Future Scope …….. 20

CHAPTER 6 **21 – 22**

SENTIMENT ANALYSIS

6.1 Introduction ……..…….. 21

6.2 Why is Sentiment Analysis Important?…….. 21

6.3 What are the types of sentiment analysis? …….……. 21

6.3.1 Fine-Grained Sentiment Analysis …… 21

6.3.2 Emotion detection…….…… 22

6.3.3 Aspect-Based Sentiment Analysis ……... 22

6.3.4 Multilingual Sentiment Analysis ……. 22

6.4 How does Sentiment Analysis work? …… 22

6.4.1 Pre-processing ……. 22

6.4.2 Analysis ……... 22

CHAPTER 7 **23 – 28**

CLASSIFICATION ALGORITHMS FOR SENTIMENT ANALYSIS

7.1 Introduction……..…….. 23

7.2 Different Types of Classification ……... 24

7.2.1 Logistic Regression …….. 24

7.2.2 Decision Tree ……... 25

7.2.3 Random Forest ……. 25

7.2.4 Support Vector Machine (SVM) …….. 26

7.2.5 K-Nearest Neighbor (KNN) …….……... 27

CHAPTER 8 **29 – 32**

EXPERIMENTAL DATASETS AND METHODOLOGY

8.1 Introduction ……..……. 29

8.2 Experimental Dataset …… 29

8.3 Methodology ……. 29

CHAPTER 9 **33 – 38**

9.1 Introduction 33

9.2 Analysis of Result 33

9.2.1 Evaluation Report Based on Precision, Recall, and F1 Score ... 33

9.2.2 Evaluation Report Based on Accuracy ……. 34

9.2.3 Overall Evaluation Report …… 35

CHAPTER 10 **39 – 38**

10.1 Conclusion …… 39

10.2 Limitations …… 39

10.3 Future Works ……. 40

REFERENCES ……..…….. **41 – 42**

LIST OF TABLES

Table No.	Title	Page No.
Table 3.1	Difference between Supervised and Unsupervised Learning	8
Table 4.1	Comparison of supervised machine learning algorithms	15
Table.9.1	Evaluation report of recall, precision and f1-score	33
Table.9.2	Evaluation report of recall, precision and f1-score (Database-II)	33
Table.9.3	Evaluation report of accuracy (Database-I)	34
Table.9.4	Evaluation report of accuracy (Database-II)	34

LIST OF FIGURES

Figure No.	Title	Page No.
Fig.1.1	Types of Machine Learning (ML)	3
Fig.3.1	Supervised learning	9
Fig.3.2	Unsupervised learning	9
Fig.4.1	Supervised learning system	10
Fig.4.2	Mechanism of Supervised Learning	11
Fig.4.3	Types of Supervised Learning	12
Fig.4.4	Regression	12
Fig.4.5	Classification	13
Fig.5.1	Machine learning (ML)	17
Fig.5.2	NLP helps data analysts turn unstructured text	18
Fig.5.3	Natural Language Processing (NLP)	18
Fig.5.4	Phases of NLP	19
Fig.6.1	Sentiment	21
Fig.7.1	Generalized example of Classification Algorithms	23
Fig.7.2	Logistic Regression	24
Fig.7.3	Sigmoid function	24
Fig.7.4	Decision Tree	25
Fig.7.5	Random Forest Classification	25
Fig.7.6	Support Vector Machine (SVM)	26
Fig.7.7	Example of SVM	26
Fig.7.8	Steps of KNN	27
Fig.7.9	Example of KNN	28
Fig.8.1	Flowchart of the research	29
Fig.9.1	Graphical Representation of Recall (Dataset-I)	35
Fig.9.2	Graphical Representation of Recall (Dataset-II)	35
Fig.9.3	Graphical Representation of Precision (Dataset-I)	36
Fig.9.4	Graphical Representation of Precision (Dataset-II)	36
Fig.9.5	Graphical Representation of F1 Score (Dataset-I)	37
Fig.9.6	Graphical Representation of F1 Score (Dataset-II)	37
Fig.9.7	Graphical Representation of Accuracy (Dataset-I)	38
Fig.9.8	Graphical Representation of Accuracy (Dataset-II)	38

CHAPTER 1

INTRODUCTION TO MACHINE LEARNING

1.1 Introduction

Artificial intelligence (AI) is technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy [1]. AI offers numerous benefits across various industries and applications. A simple way to think about AI is as a series of nested or derivative concepts that have emerged over more than 70 years. Machine learning (ML) is a branch of artificial intelligence (AI) focused on enabling computers and machines to imitate the way that humans learn, to perform tasks autonomously, and to improve their performance and accuracy through experience and exposure to more data [2]. There are many types of machine learning techniques or algorithms, including linear regression, logistic regression, decision trees, random forest, support vector machines (SVMs), k-nearest neighbor (KNN), clustering and more. Each of these approaches is suited to different kinds of problems and data.

1.2 Objective

Machine learning (ML) algorithms are efficient in automating the repetitive and time-consuming tasks of a business. It will certainly free up the human resources for more strategic activities in near future. If we look at the historical data analysis, the Machine learning (ML) models may predict the future outcomes-such as sales forecasts, customer churn, or outbreaks of any disease. These algorithms will highlight the hidden patterns and relationship in data that might not strike to the human brain. Also, ML provides valuable insights to support decision-making in various domains: finance, healthcare, and marketing. This will improve the personalized experience for each user according to their preferences and behavior, which increases customer satisfaction.

1.3 Importance of Machine Learning (ML)

Since its inception in the mid-20th century by AI founders Walter Pitts, Warren McCulloch, Alan Turing, and John von Neumann, who laid the computational foundations of the field, machine learning has grown in importance in human society. The ability of machines to learn from data themselves and thereby improve over time enables organizations to automate routine tasks -- which in theory would free humans to pursue more creative and strategic work. Machine learning has enormous and wide-ranging practical applications. In finance, the algorithms of machine learning help banks spot fraudulent transactions by processing massive amounts of data in real time at speeds and accuracies no human can achieve. In healthcare, ML is helping doctors diagnose diseases from medical images and inform treatment plans with predictive models of patient outcomes. In retail, many businesses are using ML to personalize shopping, predict inventory needs, and optimize supply chains. Another role of ML is performance of manual tasks that are out of human capability in terms of scale-for example, processing of huge volumes of data generated each day by digital devices. The ability to extract patterns and insights from large datasets has become a competitive differentiator in industries ranging from banking to scientific discovery. Today, many of the world's leading companies, including Meta, Google, and Uber,

integrate machine learning into their operations to inform decisions making and improve efficiency. The philosophical aspect, where machines process massive data, questions human understanding of our intelligence and our role in interpretation and action on complex information. More practically, this opens important ethical considerations with regards to the decisions made by the advanced machine learning (ML) models. The machine learning (ML) training and decision-making require transparency and explain ability; likewise, the effects of such models on employment and societal structures remain a continuous subject of oversight and discussion.

1.4 Types of Machine Learning (ML)

Classical machine learning (ML) is often categorized by how an algorithm learns to become more accurate in its predictions. The four basic types of machine learning (ML) are [3]:

1.4.1 Supervised Learning

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output [4].

1.4.2 Unsupervised Learning

Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data[5].

1.4.3 Semi-supervised Learning

Semi-supervised learning is a type of machine learning that falls in between supervised and unsupervised learning. It is a method that uses a small amount of labelled data and a large amount of unlabeled data to train a model [6].

1.4.4 Reinforcement Learning

Reinforcement Learning (RL) is a branch of machine learning focused on making decisions to maximize cumulative rewards in a given situation. Unlike supervised learning, which relies on a training dataset with predefined answers, RL involves learning through experience [7].

Those four types of machine learning (ML) encompass a wide range of algorithms and techniques, each suited to different types of data, learning objectives, and application domains. Depending on the specify problem at hand and the available data, practitioners may choose employee one or a combination of these types build effective machine learning solutions.

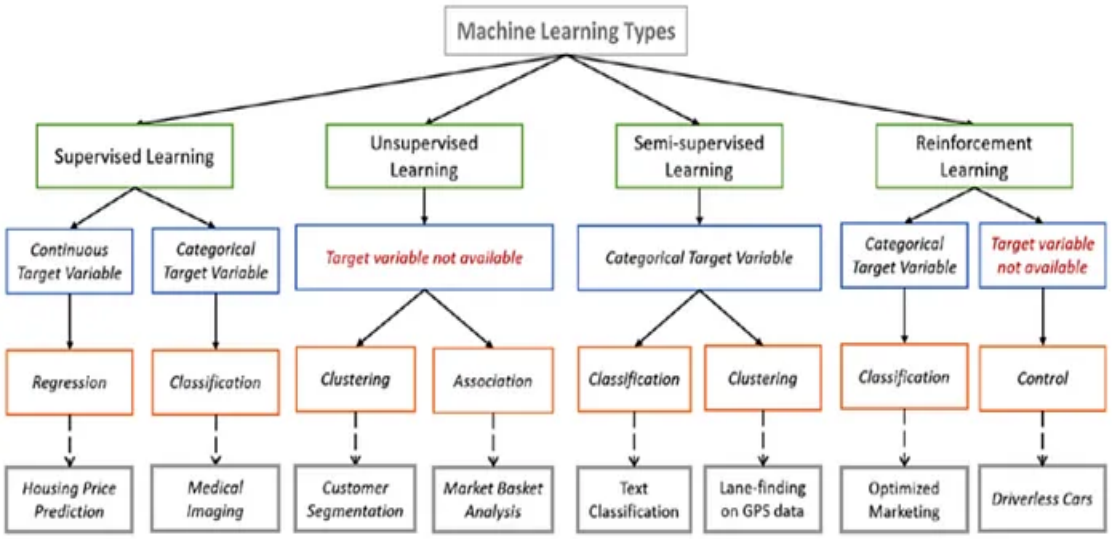


Fig.1.1: Types of Machine Learning (ML)

1.5 Features of Machine Learning (ML)

Machine learning (ML) is a subdomain of artificial intelligence (AI) that focuses on developing systems that learn or improve performance, based on the data.

a. Machine learning (ML) is the field of study that gives computers the capability to learn without being explicitly programmed.
b. It is similar to data mining, as both deal with substantial amounts of data.
c. For large organizations, branding is crucial, and targeting a relatable customer base becomes easier.
d. Given a dataset, machine learning (ML) can detect various patterns in the data.
e. Machines can learn from past data and automatically improve their performance.
f. Machine learning (ML) is a data-driven technology. A large amount of data is generated by organizations daily, enabling them to identify notable relationships and make better decisions [8].

1.6 Application of Machine Learning (ML)

The importance of machine learning is to demonstrate, effectively, the knowledge that one may have. It aids in decision-making using alternative information like statistics, data, and patterns. With this, we may make better decisions, solutions to problems, and helping in improving society.

Machine learning (ML) can be used to boost productivity, enhance security, refresh marketing strategies, and solve social issues. Its importance will continue to increase as it is helping us move closer to a better and safer future.

Let's look at the Real-Life Application of Machine Learning:

1.6.1 The role of Machine Learning (ML) in Healthcare

The tool of machine learning in the healthcare industry helps medical professionals to care for patients and manage clinical data. The machine learning (ML) applied its application to artificial intelligence (AI) that contains computer programmers to mimic human thinking. With the rise of artificial intelligence (AI) technology, we can apply healthcare to collect information on patient data. The ability of machine learning to improve decision-making and reduce the risks in the medical field. According to IDC, the future value of the AI market will be expected to triple between 2025 and 2030. At this time, we are seeing an increase in the number of jobs in start-ups for ML skills and career opportunities in this field [9]. The development of the machine learning domain is more accessible than users think. Building on the basics of machine learning will allow for innovative solutions and a wide range of career opportunities in healthcare.

1.6.2 Applying Machine Learning (ML) to Financial Markets

Machine learning is playing an important role in finance, transforming data into information and revolutionizing decision-making processes. It is a type of Artificial Intelligence (AI) that learns from data and is used in various financial operations.

1.6.3 Marketing & Advertising with Machine Learning (ML)

Machine learning (ML) in marketing and advertising is set to revolutionize actions in digital marketing, simplifying tasks such as digital advertising campaigns, content creation, and personalized recommendations. This article provides a guide to machine learning in marketing, examining its definition and business applications. Machine learning (ML) software can also help automate many different processes, thereby abstracting away some marketing tasks and saving marketers valuable time and energy. Many of the repetitive tasks that marketers typically have to do themselves, such as data analysis and reporting, optimizing content, and segmenting audiences, can be done faster and more effectively with machine learning (ML).

1.6.4 Autonomous Vehicles using Machine Learning (ML)

The autonomous vehicle's machine learning forms a revolutionary technology that builds in them the ability for navigation with self-determined instructions. This is the technology that provides the capability for vehicles to update the data that will be required for operating and updating themselves from time to time. Machine learning enables vehicles to get adapted to new locations, adverse travel conditions, and different road transportation problems. For that, data from sensitive sensors of the vehicles like cameras, radar, lasers, etc., are directed to support the autonomous driving of the vehicle. Moreover, machine learning (ML) vehicles can adapt themselves to such a degree that they can be driven according to the needs and circumstances of the passenger. This technology contributes to the development of safer, more energy-efficient vehicles, making travel easier on the road and enabling future generations to travel safely and sustainably.

1.6.5 Retail with Machine Learning (ML)

Machine learning (ML) in marketing is one of the major techniques used in the marketing field to support the processing and selling of products and communicate with customers. It can be applied in several areas in marketing, including marketing strategy, market analysis, customer service, and expertise in product context. The machine learning (ML) brings the understanding of the

forecasted demand of their products and services. They base their data analysis on studying customers' behavior and analyzing market trends. This helps them manage the inventories when the products are needed, adjusting marketing strategies, and suggesting specific products to customers. Machine learning (ML) in marketing also finds its application in areas related to customers, such as understanding customer preferences and applications. This helps them have a greater understanding of the business and, therefore, plan to deliver their services. This leads to the highest level of customer satisfaction and productivity in the marketing field. Let’s discuss on the topic.

1.6.6 Machine Learning (ML) for Energy

Machine learning (ML) in the energy sector helps predict energy demand, manages energy production and distribution, and also aims at increasing energy efficiency. Therefore, with machine learning, energy companies can better handle resources and help make the production of energy safer and more efficient. Energy organizations make use of machine learning (ML) in analyzing data related to sun, electricity, water, etc. This data helps in characterizing energy production, energy use patterns, and discovers energy savings opportunities. Furthermore, it helps ensure greater security of energy resources, regularity, and greater concentration of production. Machine learning (ML) can help energy organizations adjust operating plans to meet expected production demand. It can help energy organizations improve the efficiency of operations and determine the science and technology that can be used to make energy production safest and more effective. Applications of machine learning can, in the near future, be really significant in designing more secure, sustainable, and useful energy systems. It will also contribute to helping energy organizations improve the quality, safety, and performance of their production.

Machine learning (ML) is one of the most powerful technologies that has ever transformed many industries, from healthcare and finance to marketing and retail. Machine learning makes processes much more efficient and effective. As machine learning continues to develop, we can expect even more innovative applications to come out in the years ahead. "This course is very well structured and easy to learn. Anyone with zero experience of data science, python or ML can learn from this. This course makes things so easy that anybody can learn on their own. It's helping me a lot. Thanks for creating such a great course."- Ayushi Jain | Placed at Microsoft.

1.7 Organization of the Thesis Book

This thesis consists of 10 chapters ranging from Chapter 1 to Chapter 10. Chapter 1 gives an overview of the topic. It additionally comprises of Problem Statements, Objectives. In the meantime, Chapter 2 reviews the past research works that was led by others explores. All the significant specialized papers, diaries and books taken from those inquires about will be talked about in detail. Chapter 3 deals with several classification difficulties. In chapter 5 Convolutional Neural Network has been discussed along with its three architectures. The methodology of our thesis has been depicted in chapter 8. In chapter 9 Findings and Results have been uncovered respectfully. Finally in chapter 10 the thesis has been concluded and future scope has been portrayed.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, we embark on a journey through the extensive research. The ability to accurately classify data is a fundamental task in machine learning. From predicting customer churn to detecting fraudulent transactions, classification algorithms play a pivotal role in various industries. However, the choice of algorithm can significantly impact the performance of a machine learning model. We have collected research from renowned sites.

2.2 Case Study and Overview of Related Works

Millions of users share opinions on various topics using micro-blogging every day. Several studies on machine learning-based sentiment analysis have been carried out in the past. This work aims at building a classifier able of predicting the polarity of a comment while using Machine Learning (ML) algorithms. A study was conducted by Jie Ying Tan (2022) through comparison between Support Vector Classifier (SVC), Multi-layer Perceptron Classifier (MLP), Extreme Gradient Boosting Classifier (XGB), Logistic Regression (LR) and Multinomial Naïve Bayes (MNB), it was evident that SVC had the most superior performance [10]. Yassine Al Amrani (2018) and others proposed hybrid approach to identify product reviews offered by Amazon. The results show that the proposed system approach outperforms these individual classifiers in this amazon dataset [11]. Fatma Jemai, Mohamed Hayouni and Sahbi Baccar had done another research. They have divided into three tasks: data extraction, processing and modelling. Based on a supervised probabilistic machine learning algorithm, they tended to create a classifier to classify our tweets into positive and negative sentiments then we opt for two experiments to evaluate the performance of our model. Compared to previous reported works, they achieved greater precision [12]. Another study was done by Yaoyong Li, Kalina and Hamish on 2009. Support Vector Machines (SVM) have been used successfully in many Natural Language Processing (NLP) tasks. The novel contribution of this paper is in investigating two techniques for making SVM more suitable for language learning tasks. Firstly, we propose an SVM with uneven margins (SVMUM) model to deal with the problem of imbalanced training data [13]. Rodrigo Rocha Silva (2019) shows that Sentiment Analysis and Text Classification helps to extract precious information about data and assigning a text into one or more target categories according to its content. This paper proposes a comparison between four of the most popular Text Classification Algorithms - Naive Bayes, Support Vector Machine, Decision Trees and Random Forest - based on the Amazon Unlocked mobile phone reviews dataset. Moreover, we also study the impact of some attributes (Brand and Price) on the polarity of the review. Our results demonstrate that the Support Vector Machine is the most complete algorithm of this study and achieve the highest values in all the metrics such as accuracy, precision, recall, and F1 score [14]. Bangladeshi author, Mohammad Rezwanul Huq, Ahmad Ali and Anika Rahman had done another research on this sector on 2017. They had introduced two methods: one of the methods is known as sentiment classification algorithm (SCA)

based on k-nearest neighbor (KNN) and the other one is based on support vector machine (SVM). They also evaluated their performance based on real tweets [15].

Based on the previous studies, there exists several research gaps for sentiment analysis. It is evident from the literature that a diverse range of algorithms, including Support Vector Machines (SVM), Logistic Regression, Naive Bayes, and Decision Trees, have been used. While the utilization of numerous algorithms is time-consuming, it also casts doubt on the accuracy of the results, thereby complicating the decision-making process for researchers. To address this challenge, we aimed to identify the most effective algorithm among the top five commonly used ones when applied to identical features using substantial datasets. Our findings are presented below. This approach enables us to expedite tasks by selecting a specific algorithm based on our analysis of extensive data, rather than iteratively testing multiple algorithms and introducing unnecessary complexities.

2.3 The Novelty of Our Approach

Our research offers a unique perspective that sets it apart from other studies in the field. The main contributions of this work are the following:

a. Compare Logistic Regression, Decision Trees, Random Forest, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) algorithms based on Accuracy, and F1 score.
b. This comparison is based on a massive dataset of over a million customer responses.
c. We have undertaken a multifaceted study across various social media platforms, with a particular focus on Twitter, Facebook, Instagram, and Amazon and all other social medias.
d. Our decision to employ pandas' built-in functions for data cleaning and manipulation has significantly reduced the complexity of our analysis, making it easier for others to replicate our findings.

CHAPTER 3

SUPERVISED AND UNSUPERVISED LEARNING

3.1 Introduction

Machine learning (ML) is a subfield of artificial intelligence (AI) computer science that gives computers the ability to learn without being explicitly programmed. Supervised learning and unsupervised learning are two main types of machine learning (ML). In supervised learning, the machine is trained on a set of labeled data, which means that the input data is paired with the desired output. The machine then learns to predict the output for new input data. Supervised learning is often used for tasks such as classification, regression, and object detection.

On the other hand, in unsupervised learning, the machine is trained on a set of unlabeled data, which means that the input data is not paired with the desired output. The machine then learns to find patterns and relationships in the data. Unsupervised learning is often used for tasks such as clustering, dimensionality reduction, and anomaly detection.

3.2 Difference between Supervised and Unsupervised Learning

The difference between supervised and unsupervised learning is mentioned below:

Supervised Learning	Unsupervised Learning
1. Supervised learning is a machine learning approach that's defined by its use of labeled data sets.	1. Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled data sets.
2. Simpler method.	2. Computationally complex method.
3. Highly accurate.	3. Less accurate.
4. Also called classification.	4. Also called clustering.
5. Example: Optical character recognition.	5. Example: Find a face in an image.

Table 3.1: Difference between Supervised and Unsupervised Learning

3.3 Graphical Difference

Difference between supervised and unsupervised learning in graphical mode is mentioned below:

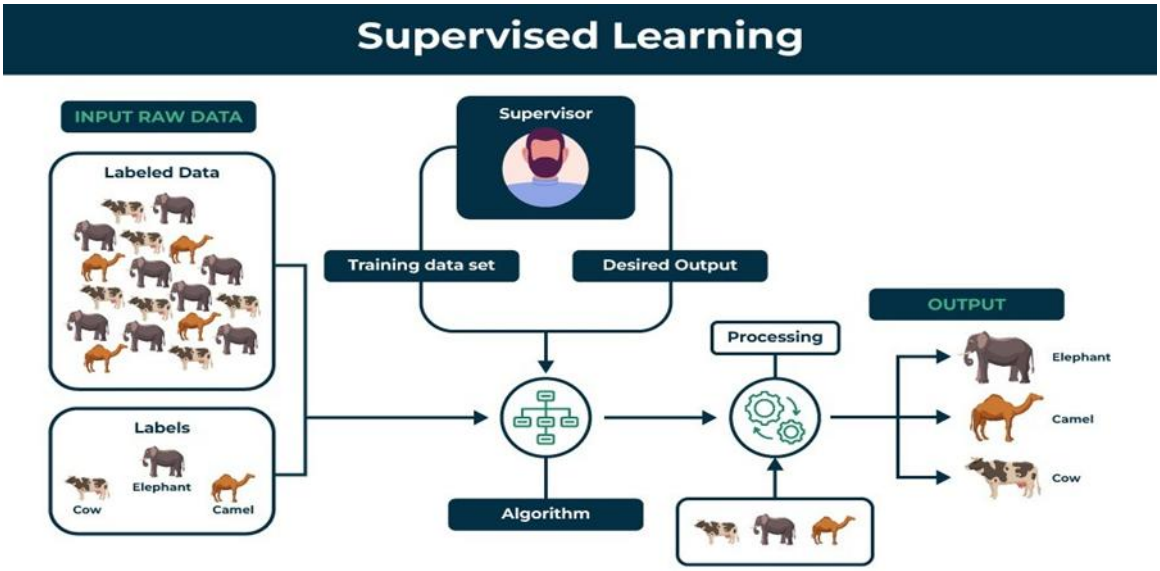


Fig.3.1: Supervised learning

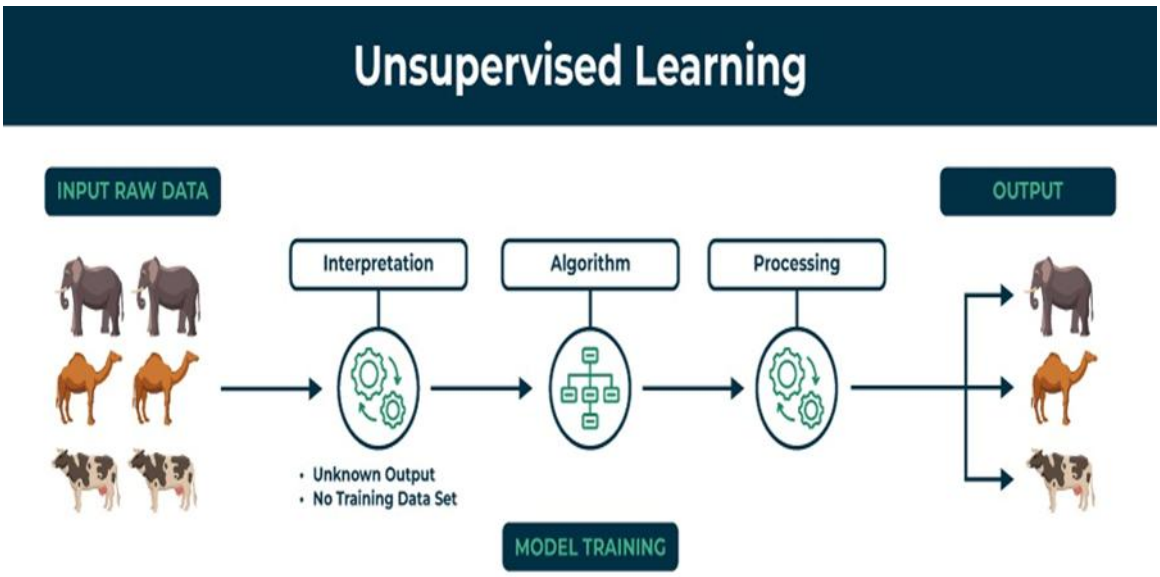


Fig.3.2: Unsupervised learning

CHAPTER 4

SUPERVISED LEARNING

4.1 Introduction

Machine learning (ML) is a field of computer science that gives computers the ability to learn without being explicitly programmed. Supervised learning one of the main types of machine learning (ML). Supervised learning is the types of machine learning (ML) in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. It is a process of providing input data as well as correct output data to the machine learning (ML) model. In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.

4.2 How does Supervised Learning works?

In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

The working of Supervised learning can be easily understood by the below example and diagram:

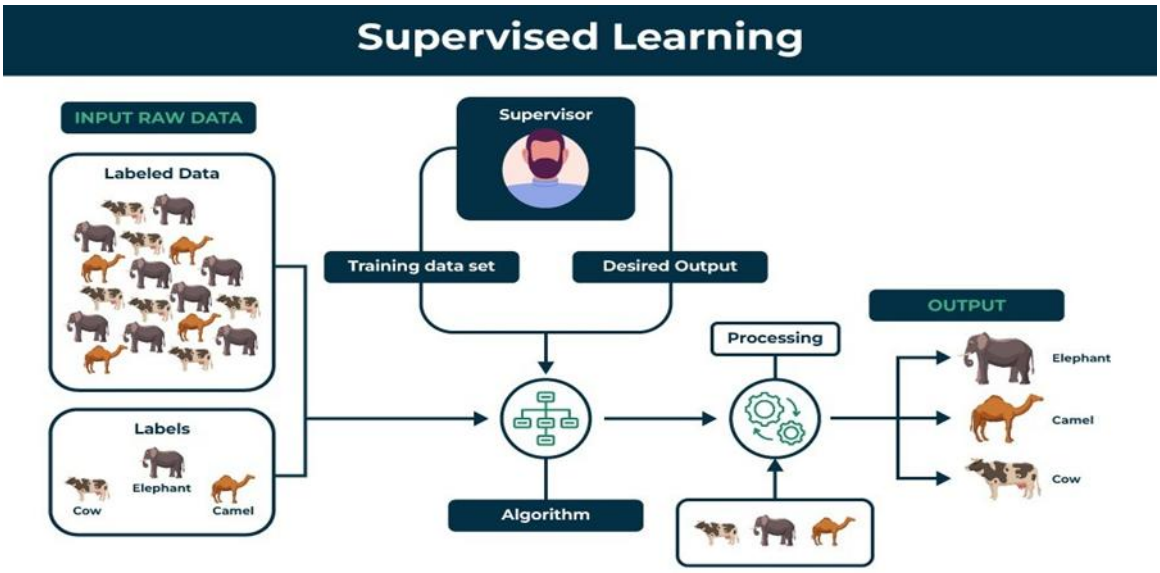


Fig.4.1: Supervised learning system

Suppose we have a dataset of different animals which includes elephant, camel, and cow. Now the first step is that we need to train the model for each label.

a. If the given shape has four legs and trunk, then it will be labelled as an elephant.
b. If the given shape has four legs and long neck, then it will be labelled as a camel.
c. If the given shape has four legs and long tail then it will be labelled as a cow.

Now, after training, we test our model using the test set, and the task of the model is to identify the shape. The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape, and predicts the output.

Let's learn how a supervised machine learning model is trained on a dataset to learn a mapping function between input and output, and then with learned function is used to make predictions on new data:

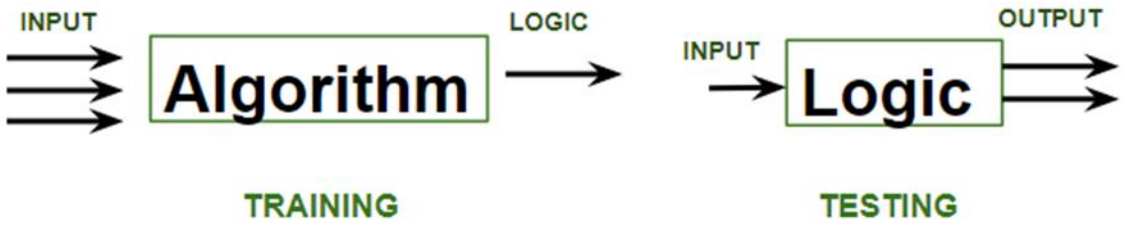


Fig.4.2: Mechanism of Supervised Learning

In the image above, Training phase involves feeding the algorithm labeled data, where each data point is paired with its correct output. The algorithm learns to identify patterns and relationships between the input and output data. Testing phase involves feeding the algorithm new, unseen data and evaluating its ability to predict the correct output based on the learned patterns.

4.3 Steps involved in Supervised Learning

a. First Determine the type of training dataset
b. Collect/Gather the labelled training data.
c. Split the training dataset into training dataset, test dataset, and validation dataset.
d. Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
e. Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
f. Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
g. Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

4.4 Types of Supervised Learning

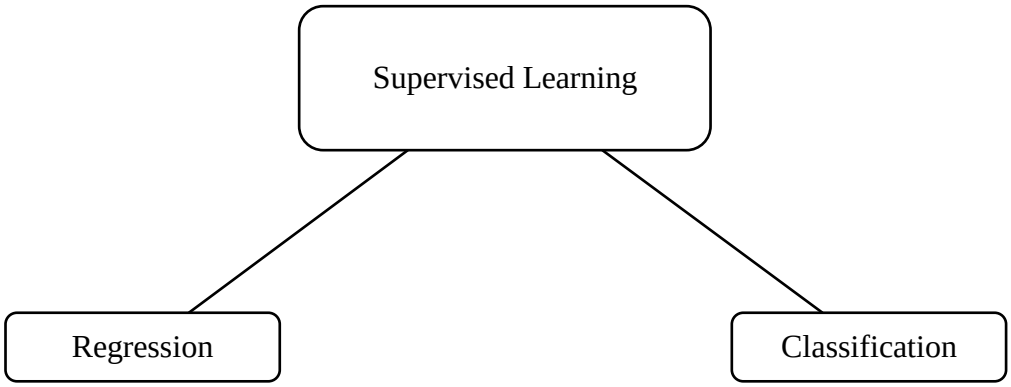


Fig.4.3: Types of Supervised Learning

4.4.1 Regression

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, and etc. Below are some popular Regression algorithms which come under supervised learning [10].

a. Linear Regression
b. Regression Trees
c. Non-Linear Regression
d. Bayesian Linear Regression
e. Polynomial Regression

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Fig.4.4: Regression

Fig.3.4: It is a Meteorological dataset that serves the purpose of predicting wind speed based on different parameters.

Input: Dew Point, Temperature, Pressure, Relative Humidity, Wind Direction Output: Wind Speed

4.4.2 Classification

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

a. Random Forest
b. Decision Trees
c. Logistic Regression
d. Support vector Machines (SVM)

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Fig.3.5: Classification

Fig.3.5: It is a dataset of a shopping store that is useful in predicting whether a customer will purchase a particular product under consideration or not based on his/ her gender, age, and salary.

Input: Gender, Age, Salary

Output: Purchased i.e., 0 or 1; 1 means yes, the customer will purchase and 0 means that the customer won't purchase it.

4.5 Supervised Machine Learning Algorithms

Supervised learning can be further divided into several different types, each with its own unique characteristics and applications. Here are some of the most common types of supervised learning algorithms:

4.5.1 Linear Regression

Linear regression is a type of supervised learning regression algorithm that is used to predict a continuous output value. It is one of the simplest and most widely used algorithms in supervised learning.

4.5.2 Logistic Regression

Logistic regression is a type of supervised learning classification algorithm that is used to predict a binary output variable.

4.5.3 Decision Trees

Decision tree is a tree-like structure that is used to model decisions and their possible consequences. Each internal node in the tree represents a decision, while each leaf node represents a possible outcome.

4.5.4 Random Forests

Random forests again are made up of multiple decision trees that work together to make predictions. Each tree in the forest is trained on a different subset of the input features and data. The final prediction is made by aggregating the predictions of all the trees in the forest.

4.5.5 Support Vector Machine (SVM)

The SVM algorithm creates a hyperplane to segregate n-dimensional space into classes and identify the correct category of new data points. The extreme cases that help create the hyperplane are called support vectors, hence the name Support Vector Machine.

4.5.6 K-Nearest Neighbors (KNN)

KNN works by finding k training examples closest to a given input and then predicts the class or value based on the majority class or average value of these neighbors. The performance of KNN can be influenced by the choice of k and the distance metric used to measure proximity.

4.5.7 Gradient Boosting

Gradient Boosting combines weak learners, like decision trees, to create a strong model. It iteratively builds new models that correct errors made by previous ones.

4.5.8 Naive Bayes Algorithm

The Naive Bayes algorithm is a supervised machine learning algorithm based on applying Bayes' Theorem with the "naive" assumption that features are independent of each other given the class label. Let's summarize the supervised machine learning algorithms in table:

Algorithm	Regression, Classification	Purpose	Method	Use Cases
Linear Regression	Regression	Predict continuous output values	Linear equation minimizing sum of squares of residuals	Predicting continuous values
Logistic Regression	Classification	Predict binary output variable	Logistic function transforming linear relationship	Binary classification tasks
Decision Trees	Both	Model decisions and outcomes	Tree-like structure with decisions and outcomes	Classification and Regression tasks
Random Forests	Both	Improve classification and regression accuracy	Combining multiple decision trees	Reducing overfitting, improving prediction accuracy
SVM	Both	Create hyperplane for classification or predict continuous values	Maximizing margin between classes or predicting continuous values	Classification and Regression tasks
KNN	Both	Predict class or value based on k closest neighbors	Finding k closest neighbors and predicting based on majority or average	Classification and Regression tasks, sensitive to noisy data
Gradient Boosting	Both	Combine weak learners to create strong model	Iteratively correcting errors with new models	Classification and Regression tasks to improve prediction accuracy

Table 4.1: Comparison of supervised machine learning algorithms

These types of supervised learning in machine learning vary based on the problem you're trying to solve and the dataset you're working with. In classification problems, the task is to assign inputs to predefined classes, while regression problems involve predicting numerical outcomes.

4.6 Areas of Application

a. Image classification
b. Weather forecasting
c. Heart Disease Prediction
d. Cryptocurrency Prediction
e. Stock Price Prediction
f. Analyzing selling price of for cars
g. Comment analysis using NLP

4.7 Advantages of Supervised Learning

The power of supervised learning lies in its ability to accurately predict patterns and make data-driven decisions across a variety of applications. Here are some advantages of supervised learning listed below:

a. Supervised learning excels in accurately predicting patterns and making data-driven decisions.
b. Labelled training data is crucial for enabling supervised learning models to learn input-output relationships effectively.
c. Supervised machine learning encompasses tasks such as supervised learning classification and supervised learning regression.
d. Applications include complex problems like image recognition and natural language processing.
e. Established evaluation metrics (accuracy, precision, recall, F1-score) are essential for assessing supervised learning model performance.
f. Advantages of supervised learning include creating complex models for accurate predictions on new data.
g. Supervised learning requires substantial labelled training data, and its effectiveness hinges on data quality and representativeness.

4.8 Disadvantages of Supervised Learning

a. Overfitting: Models can overfit training data, leading to poor performance on new data due to capturing noise in supervised machine learning.
b. Feature Engineering: Extracting relevant features is crucial but can be time-consuming and requires domain expertise in supervised learning applications.
c. Bias in Models: Bias in the training data may result in unfair predictions in supervised learning algorithms.
d. Dependence on Labelled Data: Supervised learning relies heavily on labelled training data, which can be costly and time-consuming to obtain, posing a challenge for supervised learning techniques.

CHAPTER 5

NATURAL LANGUAGE PROCESSING

5.1 Introduction

Machine learning (ML) for natural language processing (NLP) and text analytics involves using machine learning algorithms and "narrow" artificial intelligence (AI) to understand the meaning of text documents. These documents can be just about anything that contains text: social media comments, online reviews, and survey responses, even financial, medical, legal and regulatory documents.

5.2 Machine Learning (ML) for NLP

Before diving deep into how to apply machine learning (ML) and AI for NLP and text analytics, let's clarify some basic ideas.

Most importantly, "machine learning (ML)" really means "machine teaching." We know what the machine needs to learn, so our task is to create a learning framework and provide properly-formatted, relevant, clean data for the machine to learn from.



Fig.5.1: Machine learning (ML)

When we talk about a "model," we're talking about a mathematical representation. Input is key. A machine learning model is the sum of the learning that has been acquired from its training data. The model changes as more learnings are acquired.

Unlike algorithmic programming, a machine learning model is able to generalize and deal with novel cases. If a case resembles something the model has seen before, the model can use this prior "learning" to evaluate the case. The goal is to create a system where the model continuously improves at the task you've set it.

Machine learning (ML) for NLP and text analytics involves a set of statistical techniques for identifying parts of speech, entities, sentiment, and other aspects of text. The techniques can be expressed as a model that is then applied to other text, also known as supervised machine learning.

It also could be a set of algorithms that work across large sets of data to extract meaning, which is known as unsupervised machine learning. It's important to understand the difference between supervised and unsupervised learning, and how you can get the best of both in one system.



Fig.5.2: NLP helps data analysts turn unstructured text

Text data requires a special approach to machine learning (ML). This is because text data can have hundreds of thousands of dimensions (words and phrases) but tends to be very sparse. For example, the English language has around 100,000 words in common use. But any given tweet only contains a few dozen of them. This differs from something like video content where you have very high dimensionality, but you have oodles and oodles of data to work with, so, it's not quite as sparse. [11]

5.3 Supervised Learning for NLP and Text Analytics

In supervised learning (ML), a batch of text documents are tagged or annotated with examples of what the machine should look for and how it should interpret that aspect. These documents are used to "train" a statistical model, which is then given un-tagged text to analyze.



Fig.5.3: Natural Language Processing (NLP)

Later, we can use larger or better datasets to retrain the model as it learns more about the documents it analyzes. For example, we can use supervised learning to train a model to analyze comment analysis, and then later train it to factor to analysis the comments.

The most popular supervised NLP machine learning algorithms are:

a. Support Vector Machines
b. Bayesian Networks
c. Maximum Entropy
d. Conditional Random Field
e. Neural Networks/Deep Learning

All you really need to know if come across these terms is that they represent a set of data scientist guided machine learning algorithms.

5.4 Phases of NLP

There are two components of Natural Language Processing:

a. Natural Language Understanding
b. Natural Language Generation

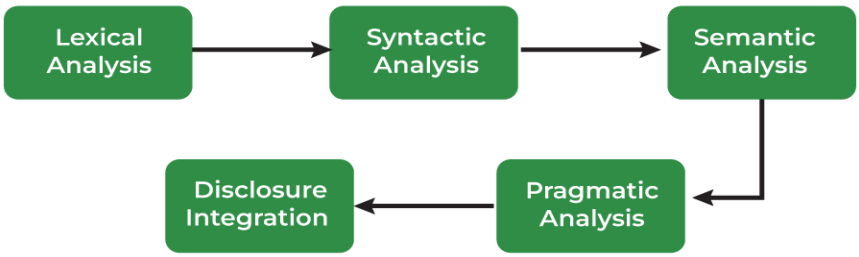


Fig.5.4: Phases of NLP

5.5 Technologies related to Natural Language Processing (NLP)

There are a variety of technologies related to natural language processing (NLP) that are used to analyze and understand human language. Some of the most common include:

a. Machine learning: NLP relies heavily on machine learning techniques such as supervised and unsupervised learning, deep learning, and reinforcement learning to train models to understand and generate human language.
b. Natural Language Toolkits (NLTK) and other libraries: NLTK is a popular open-source library in Python that provides tools for NLP tasks such as tokenization, stemming, and part-of-speech tagging. Other popular libraries include spaCy, OpenNLP, and CoreNLP.
c. Parsers: Parsers are used to analyse the syntactic structure of sentences, such as dependency parsing and constituency parsing.
d. Text-to-Speech (TTS) and Speech-to-Text (STT) systems: TTS systems convert written text into spoken words, while STT systems convert spoken words into written text.

e. Named Entity Recognition (NER) systems: NER systems identify and extract named entities such as people, places, and organizations from the text.
f. Sentiment Analysis: A technique to understand the emotions or opinions expressed in a piece of text, by using various techniques like Lexicon-Based, Machine Learning-Based, and Deep Learning-based methods
g. Machine Translation: NLP is used for language translation from one language to another through a computer.
h. Chatbots: NLP is used for chatbots that communicate with other chatbots or humans through auditory or textual methods.
i. AI Software: NLP is used in question-answering software for knowledge representation, analytical reasoning as well as information retrieval.

5.6 Contexts of Practical Application of NLP

a. Spam Filters: One of the most irritating things about email is spam. Gmail uses natural language processing (NLP) to discern which emails are legitimate and which are spam. These spam filters look at the text in all the emails you receive and try to figure out what it means to see if it's spam or not.
b. Algorithmic Trading: Algorithmic trading is used for predicting stock market conditions. Using NLP, this technology examines news headlines about companies and stocks and attempts to comprehend their meaning in order to determine if you should buy, sell, or hold certain stocks.
c. Questions Answering: NLP can be seen in action by using Google Search or Siri Services. A major use of NLP is to make search engines understand the meaning of what we are asking and generate natural language in return to give us the answers.

5.7 Future Scope

Chatbots assist clients to get to the point quickly by answering inquiries and referring them to relevant resources and products at any time of day or night. To be effective, chatbots must be fast, smart, and easy to use, to accomplish this, chatbots employ NLP to understand language, usually over text or voice-recognition interactions. Almost every connection we have with machines involves human communication, both spoken and written. Amazon's Echo is only one illustration of the trend toward putting humans in closer contact with technology in the future. The concept of an invisible or zero user interface will rely on direct communication between the user and the machine, whether by voice, text, or a combination of the two. NLP helps to make this concept a real-world thing. NLP's future also includes improved search, something we've been discussing at Expert System for a long time. Smarter search allows a chatbot to understand a customer's request can enable "search like you talk" functionality (much like you could query Siri) rather than focusing on keywords or topics. Google recently announced that NLP capabilities have been added to Google Drive, allowing users to search for documents and content using natural language.

CHAPTER 6

SENTIMENT ANALYSIS

6.1 Introduction

Sentiment analysis is a popular task in natural language processing. The goal of sentiment analysis is to classify the text based on the mood or mentality expressed in the text, which can be positive negative, or neutral.

Sentiment analysis is the process of classifying whether a block of text is positive, negative, or neutral. The goal that Sentiment mining tries to gain is to be analyzed people's opinions in a way that can help businesses expand. It focuses not only on polarity (positive, negative & neutral) but also on emotions (happy, sad, angry, etc.). It uses various Natural Language Processing algorithms such as Rule-based, Automatic, and Hybrid.

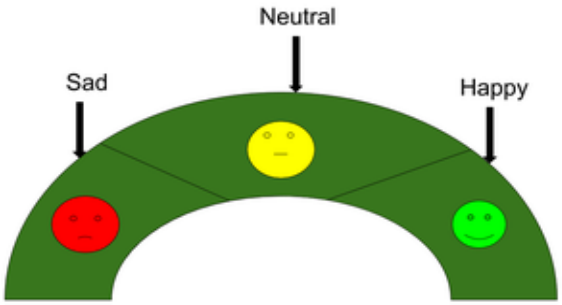


Fig.6.1: Sentiment

6.2 Why is Sentiment Analysis Important?

Sentiment analysis is the contextual meaning of words that indicates the social sentiment of a brand and also helps the business to determine whether the product they are manufacturing is going to make a demand in the market or not. According to the survey, 80% of the world's data is unstructured. The data needs to be analyzed and be in a structured manner whether it is in the form of emails, texts, documents, articles, and many more. Sentiment Analysis is required as it stores data in an efficient, cost friendly. Sentiment analysis solves real-time issues and can help you solve all real-time scenarios.

6.3 What are the Types of Sentiment analysis?

6.3.1 Fine-Grained Sentiment Analysis

This depends on the polarity base. This category can be designed as very positive, positive, neutral, negative, or very negative. The rating is done on a scale of 1 to 5. If the rating is 5 then it is very positive, 2 then negative, and 3 then neutral.

6.3.2 Emotion detection

The sentiments happy, sad, angry, upset, jolly, pleasant, and so on come under emotion detection. It is also known as a lexicon method of sentiment analysis.

6.3.3 Aspect-Based Sentiment Analysis

It focuses on a particular aspect for instance if a person wants to check the feature of the cell phone, then it checks the aspect such as the battery, screen, and camera quality then aspect based is used.

6.3.4 Multilingual Sentiment Analysis

Multilingual consists of different languages where the classification needs to be done as positive, negative, and neutral. This is highly challenging and comparatively difficult.

6.4 How does Sentiment Analysis work?

Sentiment Analysis in NLP, is used to determine the sentiment expressed in a piece of text, such as a review, comment, or social media post.

The goal is to identify whether the expressed sentiment is positive, negative, or neutral. Let us understand the overview in general two steps:

6.4.1 Pre-processing

Starting with collecting the text data that needs to be analyzed for sentiment like customer reviews, social media posts, news articles, or any other form of textual content. The collected text is pre-processed to clean and standardize the data with various tasks:

a. Removing irrelevant information (e.g., HTML tags, special characters).
b. Tokenization: Breaking the text into individual words or tokens.
c. Removing stop words (common words like “and,” “the,” etc. that don’t contribute much to sentiment).
d. Stemming or Lemmatization: Reducing words to their root form.

6.4.2 Analysis

Text is converted for analysis using techniques like bag-of-words or word embeddings (e.g., Word2Vec, GloVe). Models are then trained with labelled datasets, associating text with sentiments (positive, negative, or neutral). After training and validation, the model predicts sentiment on new data, assigning labels based on learned patterns.

CHAPTER 7

CLASSIFICATION ALGORITHMS FOR SENTIMENT ANALYSIS

7.1 Introduction

Every person has some kind of attitude towards things he experiences. We can like this handwritten note feature in the smartphone but can't stand the whole noise meter shebang. And there is also this face-lock thing that really puzzles us. It is a natural thing. All this says something about an object in question. And since this thing can be used by many people – there are dozens of such opinions from many people. When combined all these opinions paint a distinct picture of how the particular product is perceived. That's sentiment analysis in a nutshell.

Classification is a form of "pattern recognition," with classification algorithms applied to the training data to find the same pattern (similar words or sentiments, number sequences, etc.) in future sets of data.

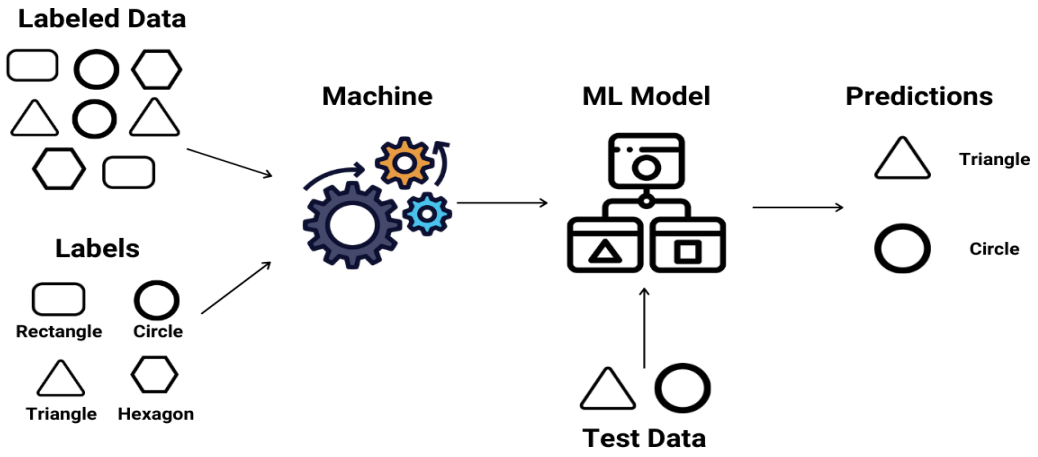


Fig.7.1: Generalized example of Classification Algorithms

7.2 Different Types of Classification

The study of classification in statistics is vast, and there are several types of classification algorithms we can use depending on the dataset we're working with. In this paper we have explored the following algorithms.

7.2.1 Logistic Regression

Logistic regression is a statistical method for developing machine learning models with binary dependent variables, i.e., binary. Logistic regression is a statistical technique used to describe data and the relationship between one dependent variable and one or more independent variables. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:

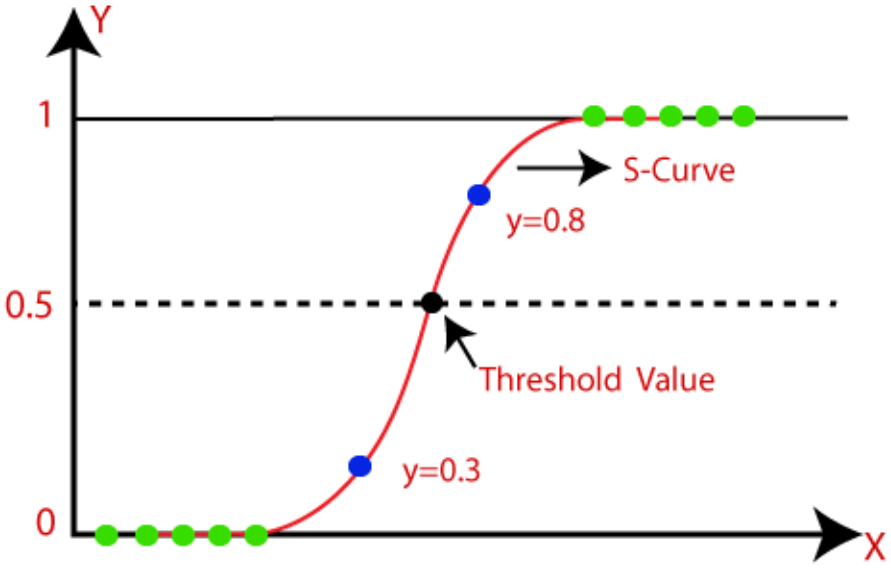


Fig.7.2: Logistic Regression

We use the sigmoid function where the input will be z and we find the probability between 0 and 1. i.e., predicted y.

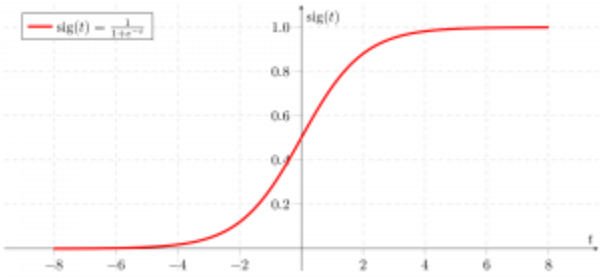


Fig.7.3: Sigmoid function

7.2.2 Decision Tree

A decision tree is a flowchart-like structure used to make decisions or predictions. It consists of nodes representing decisions or tests on attributes, branches representing the outcome of these decisions, and leaf nodes representing final outcomes or predictions. Each internal node corresponds to a test on an attribute, each branch corresponds to the result of the test, and each leaf node corresponds to a class label or a continuous value.

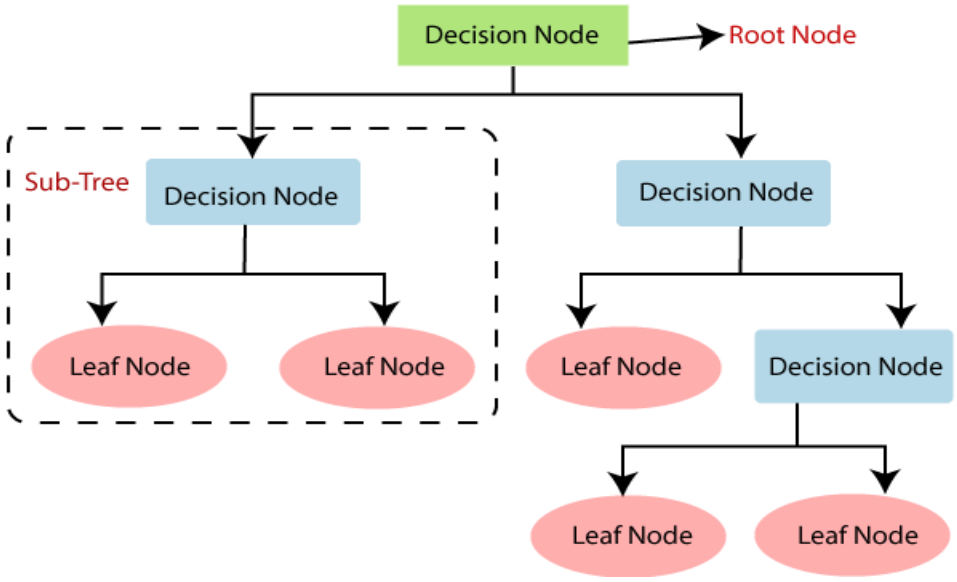


Fig.7.4: Decision Tree

7.2.3 Random Forest

Random Forest algorithm is a powerful tree learning technique in Machine Learning. It works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance.

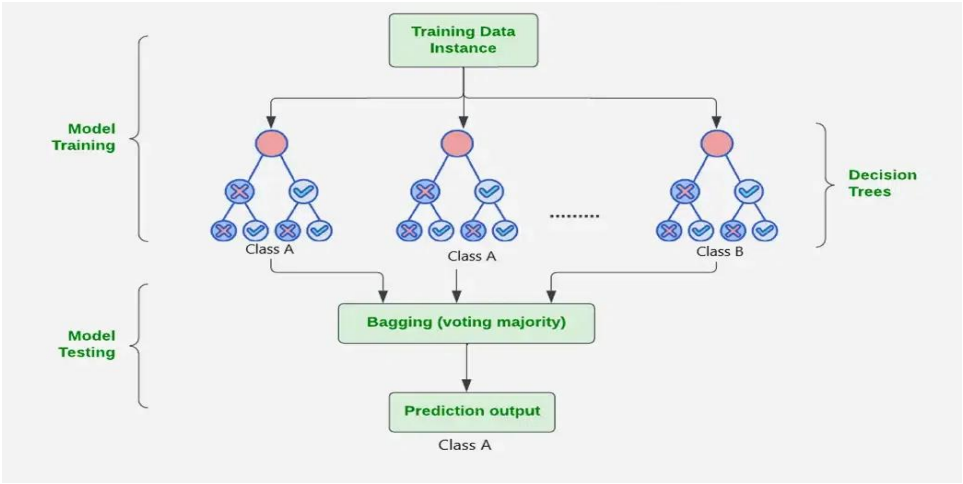


Fig.7.5: Random Forest Classification

7.2.4 Support Vector Machine (SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

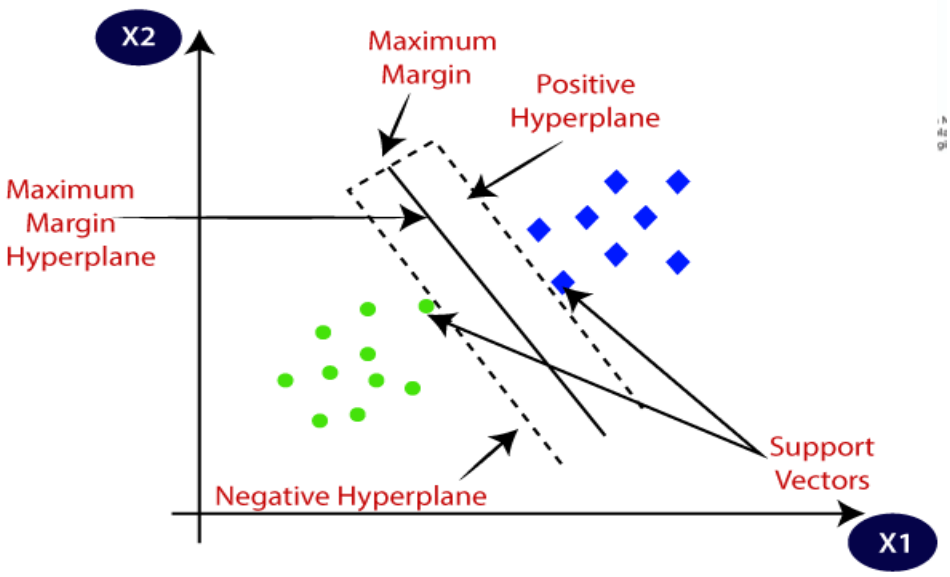


Fig.7.6: Support Vector Machine (SVM)

Example: Example: SVM can be understood with the example that we have used in the KNN classifier. Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm. We will first train our model with lots of images of cats and dogs so that it can learn about different features of cats and dogs, and then we test it with this strange creature. So as support vector creates a decision boundary between these two data (cat and dog) and choose extreme cases (support vectors), it will see the extreme case of cat and dog. On the basis of the support vectors, it will classify it as a cat. Consider the below diagram:

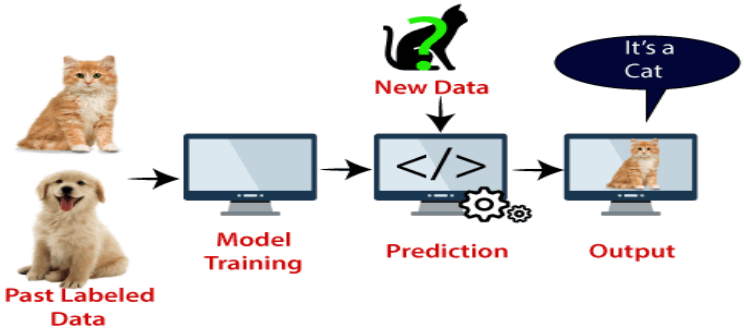
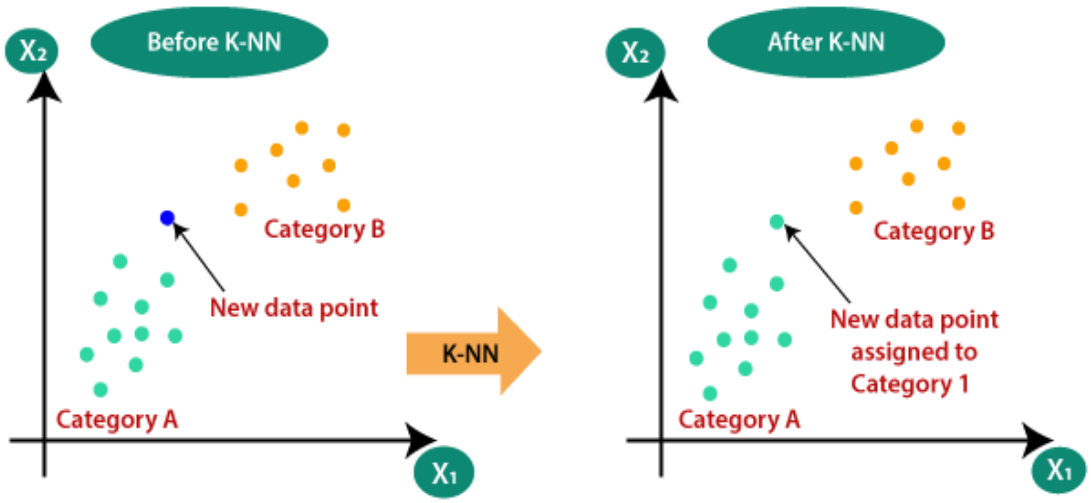
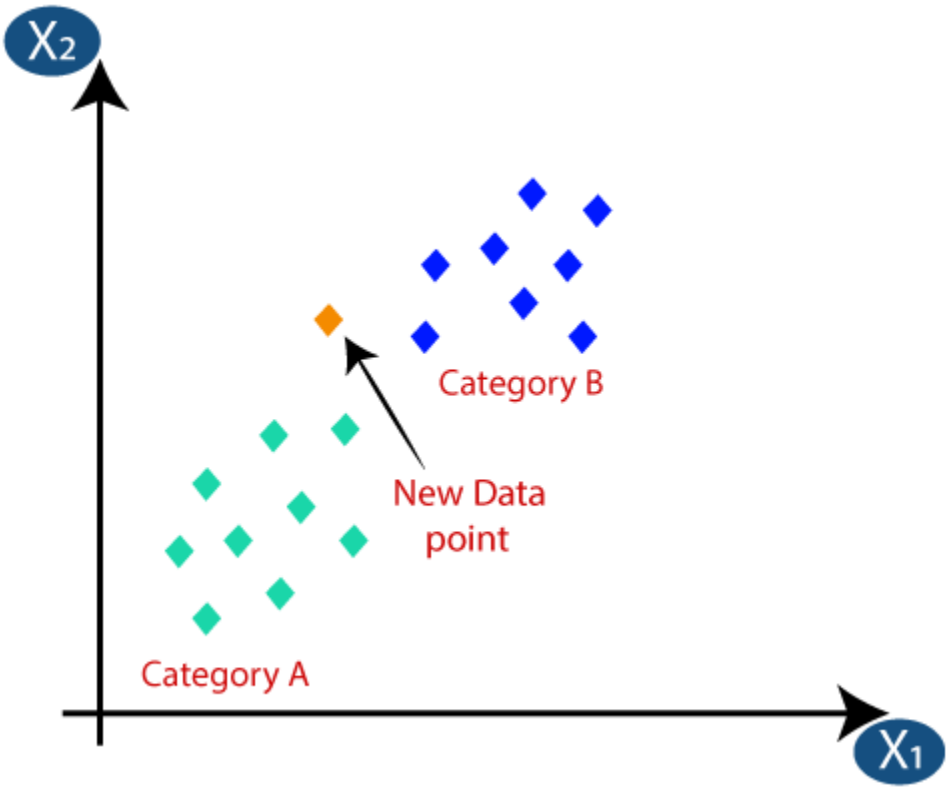


Fig.7.7: Example of SVM

7.2.5 K-Nearest Neighbor (KNN)

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. KNN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using KNN algorithm.





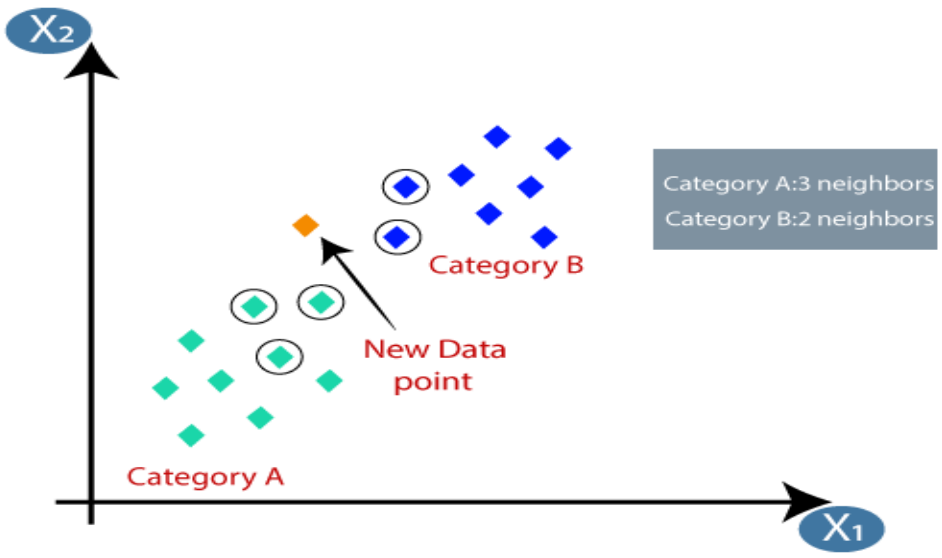


Fig.7.8: Steps of KNN

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So, for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs' images and based on the most similar features it will put it in either cat or dog category.



Fig.7.9: Example of KNN

CHAPTER 8

EXPERIMENTAL DATASETS AND METHODOLOGY

8.1 Introduction

Based on Machine Learning we have come up with three approaches to feature extraction which are CountVectorizer, HashingVectorizer, TfidfVectorizer. And also, we have used the same case on several datasets. But we have shown only two dataset results.

8.2 Experimental Dataset

Sentiment analysis, a subfield of natural language processing, has gained significant traction in recent years due to the proliferation of online reviews, social media, and other textual data. This study aims to provide a comprehensive comparison of various machine learning algorithms specifically designed for sentiment analysis tasks. While numerous studies have explored sentiment analysis, there remains a need for a systematic comparison of different algorithms to identify the most effective approach for various datasets and domains.

a. Dataset selection: The dataset used in this study consists of 133 reviews collected from https://www.kaggle.com and is annotated with sentiment labels (positive, negative, or neutral). Another dataset used in this study consists of 83 reviews collected from https://www.kaggle.com and is annotated with sentiment labels (positive, negative, or neutral).
b. Algorithm selection: A variety of machine learning algorithms, including Logistic Regression, Decision Tree, Rainforest, Support Vector Machines (SVM), and K-nearest Neighbor (KNN), were employed to classify the sentiment of the reviews.
c. Evaluation metrics: The performance of each algorithm was evaluated using standard metrics such as accuracy, recall, precision, and F1-score.

8.3 Methodology

The dataset for this study was obtained from a publicly available repository of customer comments. The data was cleaned by removing duplicates, correcting spelling errors, and handling missing values. Sentiment analysis was performed using a combination of rule-based and machine learning techniques. Machine learning models were trained on a labeled dataset to classify reviews as positive, negative, or neutral. The performance of these models was evaluated using metrics such as accuracy, recall, precision, and F1-score. Let's have look on the below flowchart:

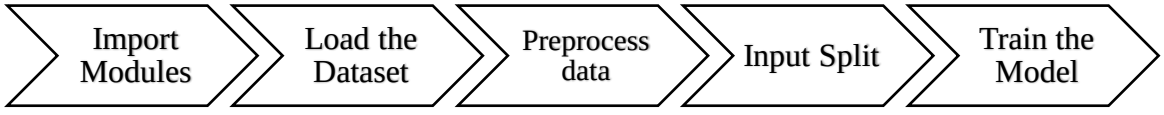


Fig.8.1: Flowchart of the research

- **Step 1: Import Modules**

First of all, we need to import the following modules.

```
import pandas as pd
import numpy as np
import seaborn as sns
import re
import string
import nltk
import warnings
```

- **Step 2: Load the Dataset**

Then we need to load the dataset.

```
df = pd.read_csv('dataset_name.csv',encoding='ISO-8859-1')
```

- **Step 3: Preprocess data**

After that we have to preprocess all the data. Follow the given steps:

a. Removes pattern in the input text

```
def remove_pattern(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for word in r:
        input_txt = re.sub(word, "", input_txt)
    return input_txt
```

b. Remove comment handles

```
df['clean_comments'] = df['comment'].apply(lambda x: remove_pattern(x, "@user"))
```

c. Remove special characters, numbers and punctuations

```
df['clean_comments'] = df['clean_comments'].str.replace("[^a-zA-Z#]", " ")
```

d. Remove short words

```
df['clean_comments'] = df['clean_comments'].apply(lambda x: " ".join([w for w in x.split() if len(w)>3]))
```

e. Individual words considered as token

```
tokenized_comments = df['clean_comments'].apply(lambda x: x.split())
```

f. Stem the words

```
from nltk.stem.porter import PorterStemmer
stemmer = PorterStemmer()
```

```
tokenized_tweet = tokenized_tweet.apply(lambda sentence: [stemmer.stem(word) for word in sentence])
```

g. Combine words into single sentence

```
for i in range(len(tokenized_tweet)):
    tokenized_comments[i] = " ".join(tokenized_comments[i])

    df['clean_comments'] = tokenized_comments
```

- **Step 4: Input Split (feature_extraction)**

a. TfidfVectorizer

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split

tfidf_vectorizer = TfidfVectorizer(max_df=0.90, min_df=2, max_features=1000, stop_words='english')
tfidf = tfidf_vectorizer.fit_transform(df['clean_tweet'])

x_train, x_test, y_train, y_test = train_test_split(tfidf, df['label'], random_state=42, test_size=0.25)
```

b. HashingVectorizer

```
from sklearn.feature_extraction.text import HashingVectorizer
from sklearn.model_selection import train_test_split

hashing_vectorizer = HashingVectorizer(n_features=1000, stop_words='english')
hashed_features = hashing_vectorizer.fit_transform(df['clean_tweet'])

x_train, x_test, y_train, y_test = train_test_split(hashed_features, df['label'], random_state=42, test_size=0.25)
```

c. CountVectorizer

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split

bow_vectorizer = CountVectorizer(max_df=0.90, min_df=2, max_features=1000, stop_words='english')
bow = bow_vectorizer.fit_transform(df['clean_tweet'])

x_train, x_test, y_train, y_test = train_test_split(bow, df['label'], random_state=42, test_size=0.25)
```

- **Step 5: Train the Model**

Finally, we need to train the model. And after that the model will be learnt using following algorithms.

a. Logistic Regression

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score, accuracy_score, precision_score, recall_score
model = LogisticRegression()
model.fit(x_train, y_train)
pred = model.predict(x_test)
```

b. Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import f1_score, accuracy_score, precision_score, recall_score
model = DecisionTreeClassifier()
model.fit(x_train, y_train)
pred = model.predict(x_test)
```

c. Random Forest

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import f1_score, accuracy_score, precision_score, recall_score
model = RandomForestClassifier()
model.fit(x_train, y_train)
pred = model.predict(x_test)
```

d. Support Vector Machine (SVM)

```
from sklearn.svm import SVC
from sklearn.metrics import f1_score, accuracy_score, precision_score, recall_score
model = SVC(kernel="rbf", gamma=0.5, C=1.0)
model.fit(x_train, y_train)
pred = model.predict(x_test)
```

e. K-Nearest Neighbors (KNN)

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import f1_score, accuracy_score, precision_score, recall_score
model = KNeighborsClassifier()
model.fit(x_train, y_train)
pred = model.predict(x_test)
```

CHAPTER 9

FINDINGS AND RESULTS

9.1 Introduction

The results indicate that Random Forest and Logistic Regression have achieved the highest accuracy of special methods on the test set, followed by Support Vector Machines with about 88%. However, in terms of training time, Decision Tree was significantly faster.

9.2 Analysis of Result

9.2.1 Evaluation Report Based on Precision, Recall, and F1 Score

The performance evaluation of the five algorithms using three feature extraction models is provided below:

a. Recall, Precision and F1-Score result base on first dataset.

Algorithm	CountVectorizer			HashingVectorizer			TfidfVectorizer		
	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score
Logistic Regression	78.93%	77.92%	78.42%	75.79%	80.55%	78.10%	75.79%	80.71%	78.17%
Decision Tree	66.36%	69.84%	68.06%	72.83%	69.73%	71.25%	64.70%	73.07%	68.63%
Random Forest	77.63%	78.07%	77.85%	79.30%	78.72%	79.01%	78.37%	79.25%	78.81%
SVM	31.05%	72.41%	43.47%	80.22%	77.78%	78.98%	79.67%	79.37%	79.52%
KNN	69.13%	72.76%	70.90%	42.70%	75.24%	54.48%	42.70%	75.24%	54.48%

Table.9.1: Evaluation report of recall, precision and f1-score (Database-I)

b. Recall, Precision and F1-Score result base on second dataset.

Algorithm	CountVectorizer			HashingVectorizer			TfidfVectorizer		
	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score
Logistic Regression	38.28%	73.79%	50.41%	23.61%	82.50%	36.72%	34.53%	80.08%	48.25%
Decision Tree	46.33%	40.03%	42.95%	35.96%	51.54%	42.36%	46.33%	57.94%	51.49%
Random Forest	45.97%	55.75%	50.39%	29.34%	88.65%	44.09%	44.72%	75.53%	56.18%
SVM	25.76%	89.44%	40.00%	22.90%	89.51%	36.47%	34.35%	81.36%	48.30%
KNN	26.30%	71.01%	38.38%	16.28%	85.85%	27.37%	33.63%	61.84%	43.57%

Table.9.2: Evaluation report of recall, precision and f1-score (Database-II)

9.2.2 Evaluation Report Based on Accuracy

a. Accuracy rate base on first dataset.

Algorithm	CountVectorizer	HashingVectorizer	TfidfVectorizer
	Accuracy	Accuracy	Accuracy
Logistic Regression	88.13%	88.38%	88.43%
Decision Tree	82.98%	83.94%	83.84%
Random Forest	87.93%	88.48%	88.48%
SVM	77.93%	88.33%	88.79%
KNN	84.49%	80.51%	80.51%

Table.9.3: Evaluation report of accuracy (Database-I)

b. Accuracy rate base on second dataset.

Algorithm	CountVectorizer	HashingVectorizer	TfidfVectorizer
	Accuracy	Accuracy	Accuracy
Logistic Regression	94.73%	94.31%	94.82%
Decision Tree	91.39%	93.15%	93.89%
Random Forest	93.67%	94.79%	95.12%
SVM	94.59%	94.42%	94.86%
KNN	94.09%	93.96%	93.91%

Table.9.4: Evaluation report of accuracy (Database-II)

9.2.3 Overall Evaluation Report

Let's follow up the graph that will help to take decision.

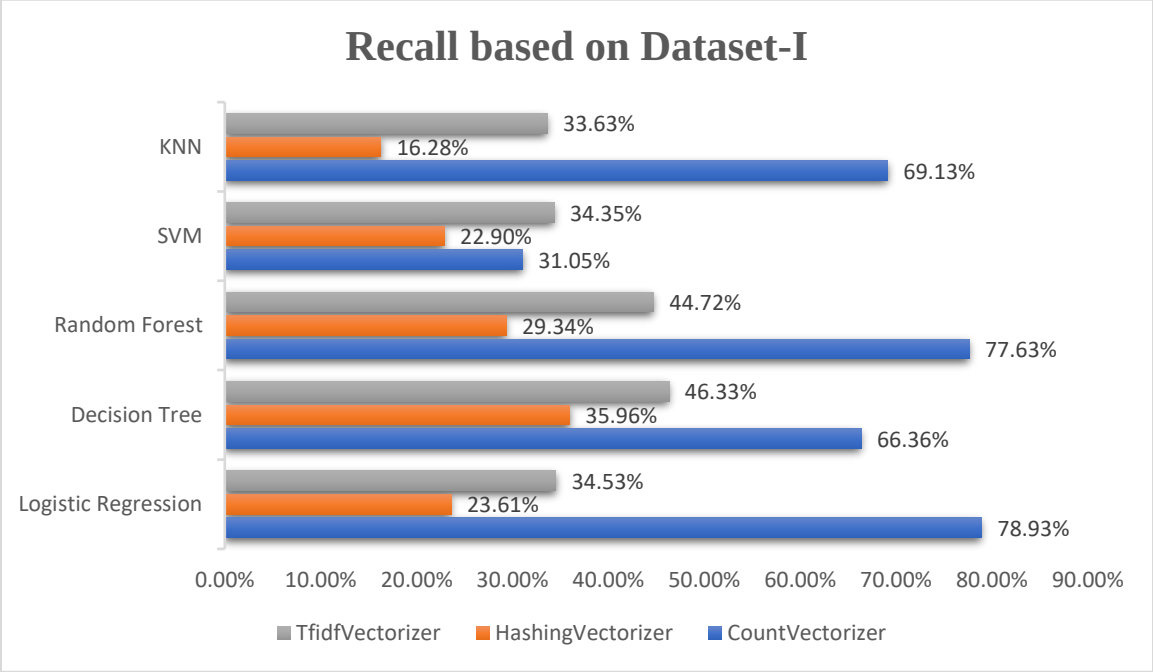


Fig.9.1: Graphical Representation of Recall (Dataset-I)

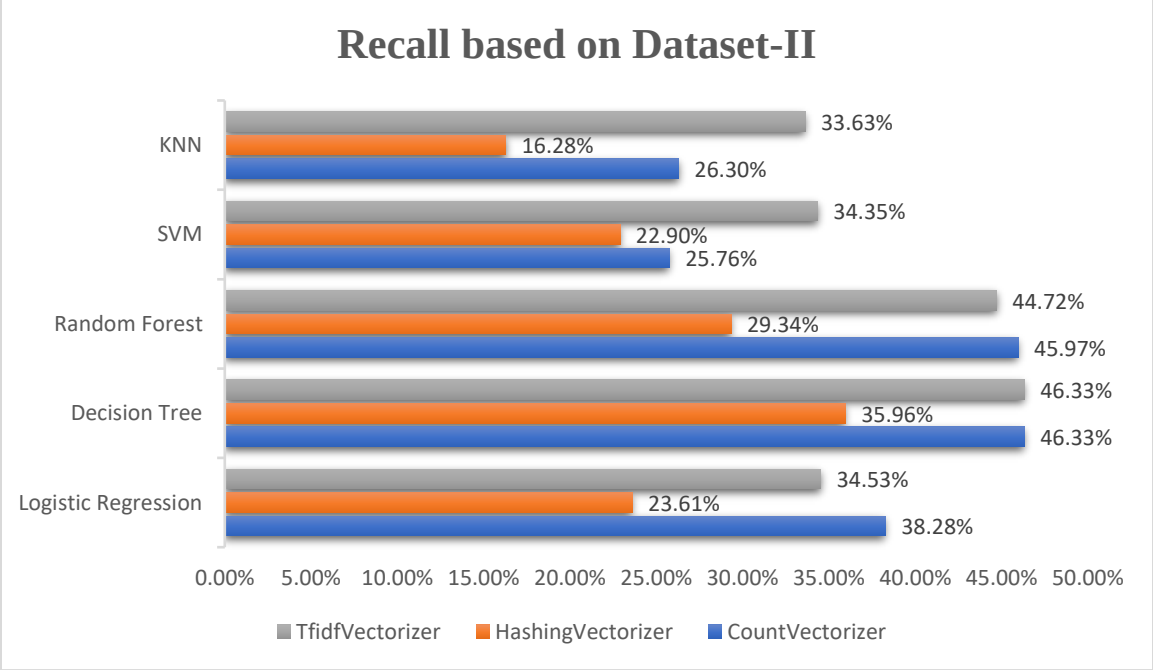


Fig.9.2: Graphical Representation of Recall (Dataset-II)

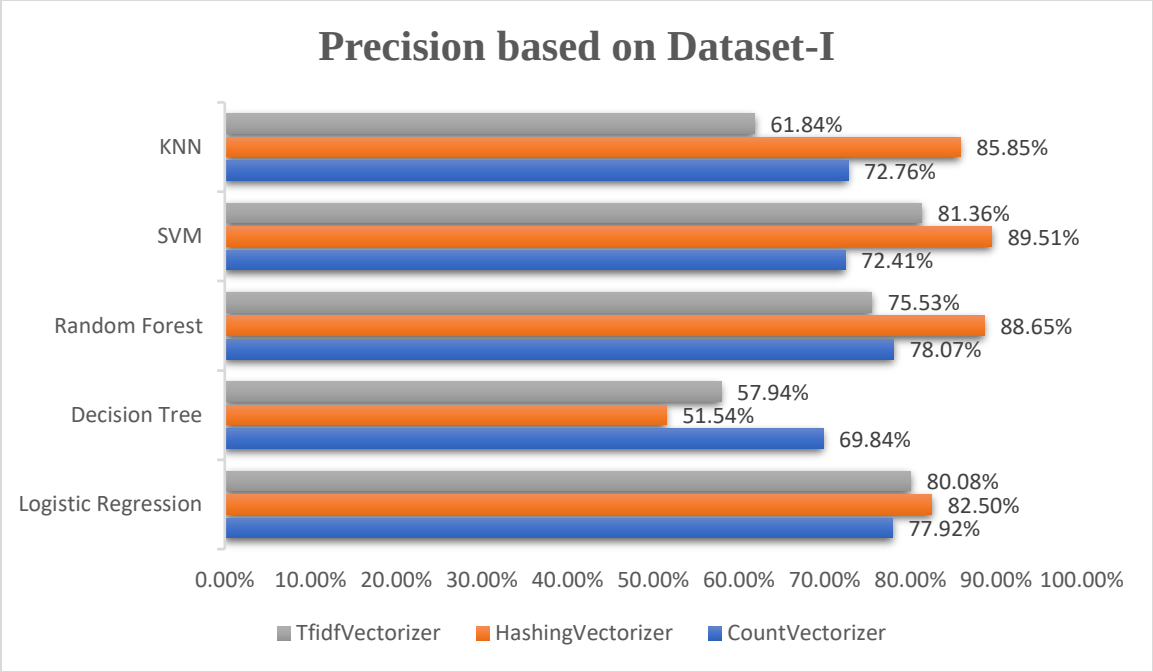


Fig.9.3: Graphical Representation of Precision (Dataset-I)

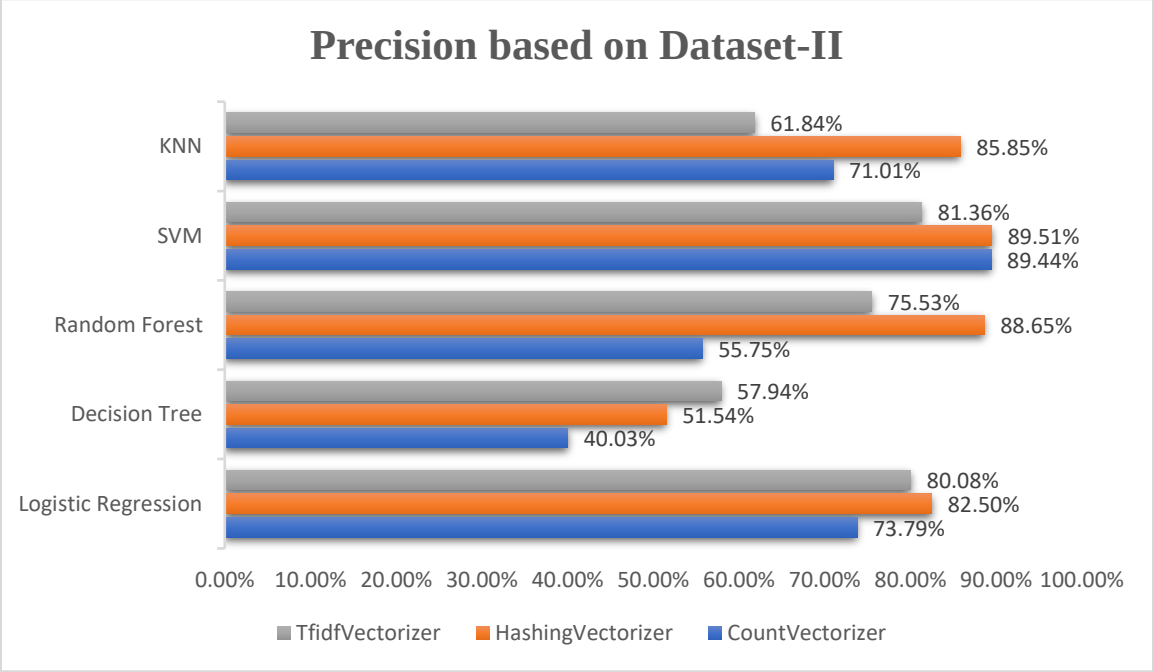


Fig.9.4: Graphical Representation of Precision (Dataset-II)

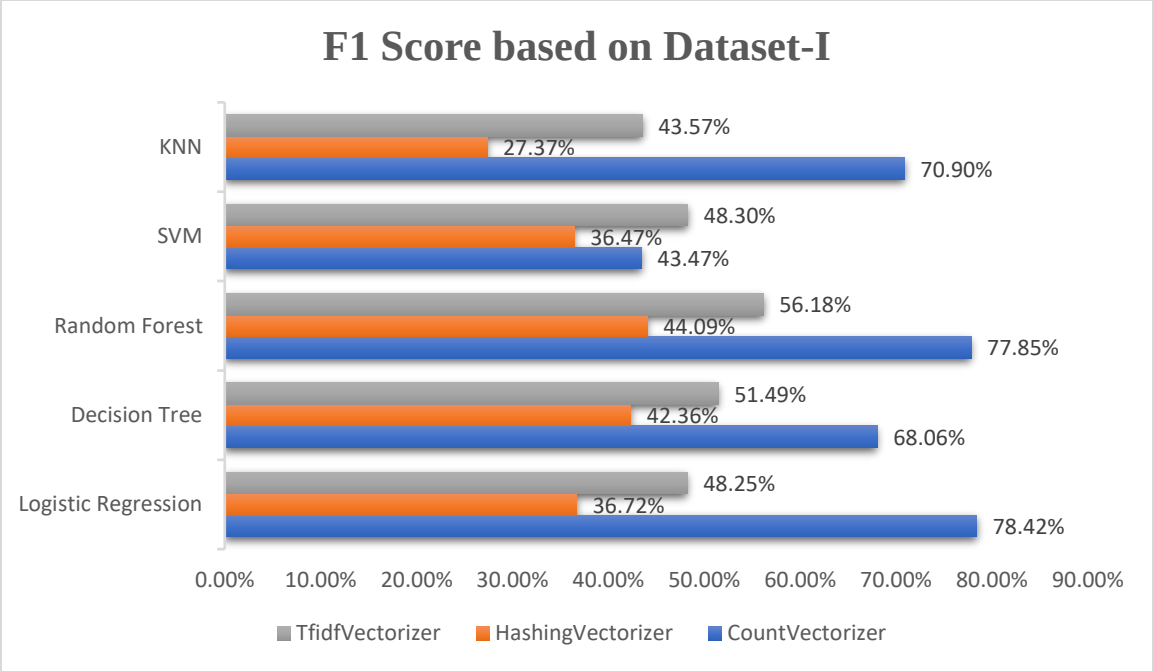


Fig.9.5: Graphical Representation of F1 Score (Dataset-I)

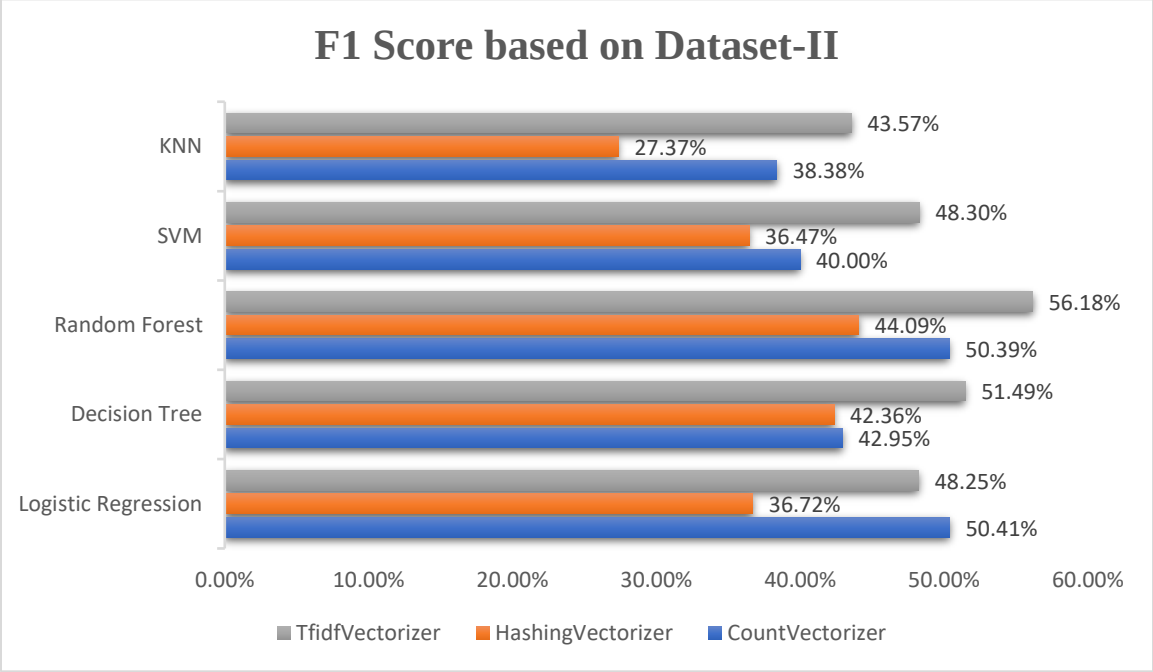


Fig.9.6: Graphical Representation of F1 Score (Dataset-II)

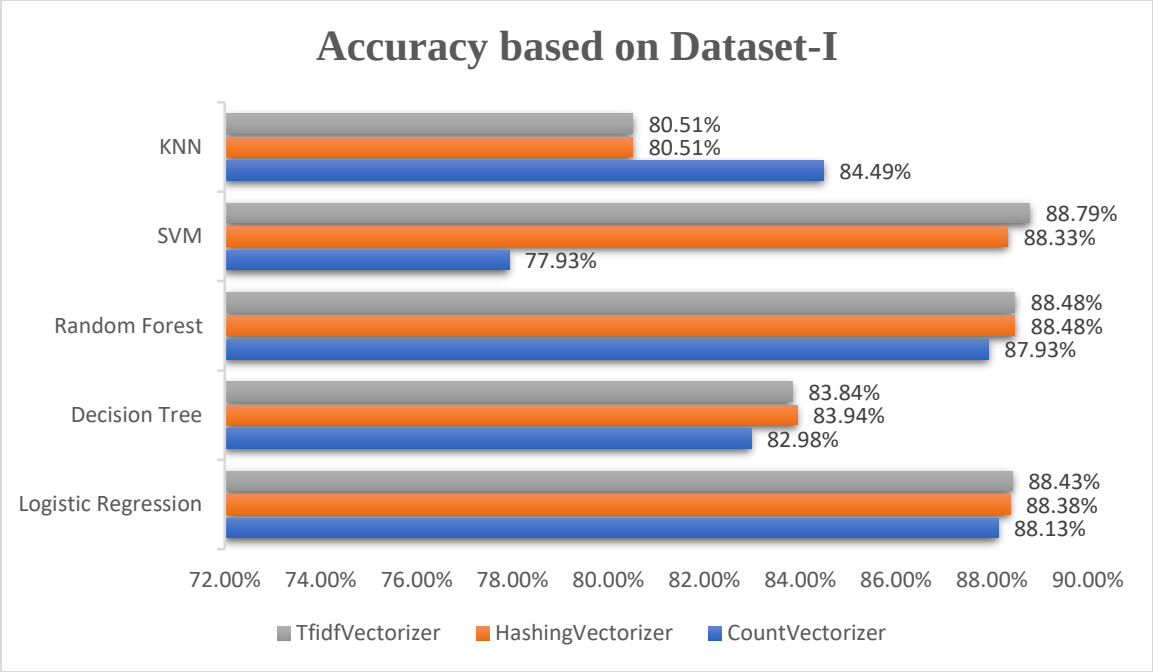


Fig.9.7: Graphical Representation of Accuracy (Dataset-I)

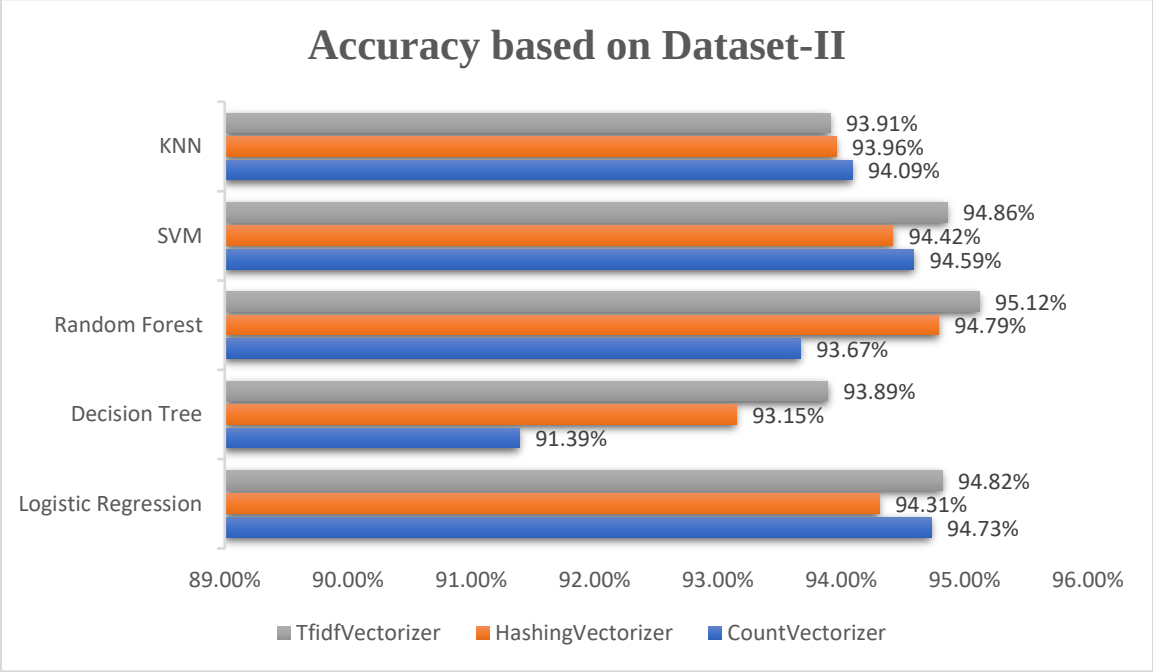


Fig.9.8: Graphical Representation of Accuracy (Dataset-II)

CHAPTER 10

CONCLUSION AND FUTURE WORKS

10.1 Conclusion

Every social networking site like Facebook, twitter, Instagram etc. become one of the key sources of information. It is found that by extracting and analyzing data from social networking sites, a business entity can be benefited in their product marketing. Twitter is one of the most popular sites where people used to express their feelings and reviews for a particular product. To complete this process or in a word we can say that to learn and the machine its very much important to fix that which algorithm should I have to use which will give more accuracy. According to the thesis, logistic regression gives 78.93% recall score in CountVectorizer method. On the other hand, SVM gives 80.22% recall accuracy in HashingVectorizer method. SVM scores 79.67% in TfidfVectorizer. So, finally we can take decision that SVM gives more accurate score in two methods. 77.92% Precision scores Logistic Regression in CountVectorizer. In HashingVectorizer method 80.55% precision score and 80.71% in TfidfVectorizer. So, it's the better algorithm. SVM algorithm is better in F1 score than Logistic Regression and others. Now, in accuracy rate, Logistic Regression gives 94.73% accuracy in CountVectorizer method. On the other hand, HashingVectorizer method 94.79% accuracy gives Random Forest. And Random Forest gives 95.12% accuracy in TfidfVectorizer method. So, finally we can come into fix decision that Logistic Regression and Random Forest algorithms are more effective than other algorithms in their own methods.

10.2 Limitations

We conducted a thorough thesis with the goal of reaching definitive conclusions. However, our research was subject to certain limitations.

Firstly, our study was confined to English language comments, and it's possible that different algorithms might be more effective for other languages.

Secondly, we limited our analysis to five specific algorithms. While these algorithms were distinct, it's likely that other algorithms could yield different results.

Additionally, if we had access to a larger dataset, the machine learning model might have learned more effectively, potentially leading to the superiority of different algorithms.

10.3 Future Works

The study was focused on sentiment analysis only, using the Twitter comments dataset. In further work, it can be extended on other datasets such as those of Facebook or Instagram. This would help us to understand which one gives the highest accuracy. Also, in business applications where supervised learning is used for training machines, research with different datasets can be done. Based on the study of such cases, it will be possible to establish which algorithm for what kind of work is better, for instance, medical applications. Because if the machine will be trained fast, outputs can be received sooner, which will further result in finer results and, therefore, in improved performance for different tasks. Key points incorporated in these options:

c. Dataset diversity: Expanding beyond Twitter to include Facebook and Instagram.
d. Language variation: Introducing Bengali as a target language for sentiment analysis.
e. Algorithm selection: Identifying the best algorithms for different tasks and domains.
f. Business applications: Exploring the use of sentiment analysis in real-world business scenarios.
g. Medical applications: Highlighting the potential of sentiment analysis in healthcare and find which algorithm will be more effective.
h. Performance metrics: Emphasizing the importance of evaluating model accuracy using appropriate metrics.
i. Transfer learning: Suggesting the use of transfer learning to improve model performance on new datasets.

In future research, we intend to focus on developing custom algorithms for Bengali sentiment analysis by creating our own datasets. This will allow us to explore novel approaches and potentially achieve more accurate results.

REFERENCES

[1] Cole Stryker and Eda Kavlakoglu, "What is artificial intelligence (AI)?", p. 9, August, 2024. [Online]. Available: https://www.ibm.com/think/topics/artificial-intelligence. [Accessed August 12, 2024].

[2] Lex Fridman, "What is machine learning?" p. Jan 11, 2019. [Online]. Available: https://www.ibm.com/think/topics/machine-learning. [Accessed August 17, 2024].

[3] Lev Craig, "Importance of Machine Learning (ML)" p. Jan 4, 2024. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML.[Accessed August 22, 2024].

[4] Varun Aina, "Supervised Learning" p. Jun 4, 2019. [Online]. Available: https://www. javatpoint.com/supervised-machine-learning. [Accessed August 25, 2024].

[5] Varun Aina, "Supervised Learning" p. Jun 4, 2019. [Online]. Available: https://www. javatpoint.com/unsupervised-machine-learning. [Accessed August 25, 2024].

[6] Varun Aina,"Unsupervised Learning" p. Jun 4, 2019. [Online]. Available: https://www. javatpoint.com/unsupervised-machine-learning. [Accessed August 25, 2024].

[7] Varun Aina,"Semi - supervised Learning" p.Jun 4, 2019. [Online]. Available: https://www.javatpoint.com/semi-supervised-machine-learning. [Accessed August 25, 2024].

[8] Ranji Raj and Sandeep Jain, "Call for new look at skilled migrants," p. May 21, 2024. [Online]. Available: https://www.geeksforgeeks.org/ml-semi-supervised-learning/. [Accessed May 31, 2024].

[9] Sandeep Jain and Winston Tang, "Reinforcement Learning: An Overview," p. September 4, 2024. [Online]. Available: https://www.geeksforgeeks.org/what-is-reinforcement-learning. [Accessed May 31, 2005].

[10] Tan, J. Y., Chow, A. S. K., & Tan, C. W. (2022). A Comparative Study of Machine Learning Algorithms for Sentiment Analysis of Game Reviews. IEM Journal, 82(3). https://doi.org/10.54552/v82i3.101

[11] Yassine Al Amrani, Mohamed Lazaar, Kamal Eddine El Kadiri, Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis, Procedia Computer Science, Volume 127, 2018, Pages 511-520, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.01.150.

[12] F. Jemai, M. Hayouni and S. Baccar, "Sentiment Analysis Using Machine Learning Algorithms," 2021 International Wireless Communications and Mobile Computing (IWCMC), Harbin City, China, 2021, pp. 775-779, doi: 10.1109/IWCMC51323 .2021.9498965

[13] LI Y, BONTCHEVA K, CUNNINGHAM H. Adapting SVM for data sparseness and imbalance: a case study in information extraction. Natural Language Engineering. 2009;15(2):241-271. doi:10.1017/S1351324908004968

[14] Guia, M., Silva, R. R. and Bernardino, J. (2019). Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis. In Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019) - KDIR; ISBN 978-989-758-382-7; ISSN 2184-3228, SciTePress, pages 525-531. doi: 10.5220/0008364105250531

[15] MohammadRezwanul Huq, Ahmad Ali, and Anika Rahman, " Sentiment Analysis on Twitter Data using KNN and SVM," International Journal of Advanced Computer Science and Applications (ijacsa), vol. 8, June 2017. https://doi.org/10.14569/IJACSA.2017.080603

[16] Sandeep Jain and Winston Tang, " Machine learning is a subset of AI," p. January 3, 2025. [Online]. Available: https://www.geeksforgeeks.org/machine-learning. [Accessed January 6, 2025].

[17] Sandeep Jain and Winston Tang, " Real life applications of machine learning," p. January 3, 2025. [Online]. Available: https://www.geeksforgeeks.org/real-life-applications-of-machine-learning/. [Accessed January 6, 2025].

[18] Varun Aina, "Introduction to supervised machine learning," p. December 2, 2023. [Online]. Available: https://www.javatpoint.com/supervised-machine-learning. [Accessed January 3, 2025].

[19] Yassine Al Amrani, Mohamed Lazaar, " Machine Learning (ML) for NLP," p. March 8, 2024. [Online]. Available: https://www.lexalytics.com/blog/machine-learning-natural-language-processing/#supervised. [Accessed January 7, 2025].

[20] Sandeep Jain and Winston Tang, " Natural language processing overview," p. April 26, 2023. [Online]. Available: https://www.geeksforgeeks.org/natural-language-processing-overview. [Accessed January 5, 2025].