

VARIOUS MACHINE LEARNING APPROACHES FOR SENTIMENT ANALYSIS

BY

**Fatema Tuz Zohra Anny
ID: 201-15-3118**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Amit Chakraborty
Assistant Professor
Department of CSE
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2024**

APPROVAL

This Project titled “**Various Machine Learning Approaches for Sentiment Analysis on Movies**”, submitted by, **Fatema Tuz Zohra (ID: 201-15-3118)** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation was held on 23 January 2024.

BOARD OF EXAMINERS



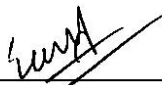
Dr. Md. Taimur Ahad
Associate Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Chairman



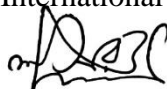
Md. Sadekur Rahman
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner 1



Md. Sazzadur Ahamed
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner 2



Dr. Mohammed Nasir Uddin
Professor
Department of Computer Science and Engineering
Jagannath University

External Examiner

DECLARATION

I hereby declare that this project has been done by me under the supervision of **Mr. Amit Chakraborty, Assistant Professor, Department of CSE** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Mr. Amit Chakraborty
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Fatema -Tuz- Zohra
ID: 201-15-3118
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty Allah for His divine blessing making us possible to complete the final year project/internship successfully.

I am grateful and wish my profound indebtedness to **Mr. Amit Chakraborty, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning and Natural Language Processing*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project. I would like to express our heartiest gratitude to **Professor Dr. Sheak Rashed Haider Noori, Professor and Head**, Department of CSE, for his kind help in finishing my project and also to other faculty members and the staff of the CSE department of Daffodil International University. I would like to thank my entire coursemates at Daffodil International University, who took part in this discussion while completing the coursework. Finally, I must acknowledge with due respect the constant support and patience of my parents.

ABSTRACT

Sentiment analysis is an essential task in natural language processing that is critical to comprehending user attitudes and public opinion in a variety of fields. This study compares and thoroughly examines several machine-learning techniques for sentiment analysis. Linear regression, Decision Tree, Random Forest, XGBoost, KNN, Artificial Neural Network (ANN), and Convolutional Neural Network (CNN) are among the models that were examined.

The main metric used in the study to assess each model's performance is accuracy. With an accuracy of 0.9823%, (TPOT) was only behind Linear Regression, which had an amazing accuracy of 0.9825%. Moreover, Decision Tree and Random Forest performed admirably, with respective accuracies of 0.9762% and 0.9805%. On the other hand, the accuracy obtained by XGBoost, KNN, and ANN were 0.9693%, 0.9753%, and 0.9783%, in that order.

Remarkably, the convolutional neural network (CNN) demonstrated a significantly reduced accuracy of 0.8199%, suggesting possible difficulties when utilizing this architecture for sentiment analysis inside the specified framework.

The research's conclusions provide important new information on which machine learning models are best suited for tasks involving sentiment analysis. Based on the particular needs of their applications, researchers and practitioners can use these data to help them choose a sentiment analysis model. The study also emphasizes how crucial it is to take into account a variety of machine learning techniques to improve the precision and dependability of sentiment analysis systems in practical contexts.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgments	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-8
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	3
1.4 Research Questions	5
1.5 Expected Output	5
1.6 Project Management and Finance	6
1.7 Report Layout	7
CHAPTER 2: BACKGROUND	9-15
2.1 Preliminaries/Terminologies	9
2.2 Related Works	10
2.3 Comparative Analysis and Summary	12
2.4 Scope of the Problem	14

2.5 Challenges	14
CHAPTER 3: RESEARCH METHODOLOGY	15-22
3.1 Research Subject and Instrumentation	16
3.2 Data Collection Procedure/Dataset Utilized	17
3.3 Statistical Analysis	18
3.4 Proposed Methodology/Applied	20
3.5 Implementation Requirements	21
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	22-27
4.1 Experimental Setup	22
4.2 Experimental Results & Analysis	24
4.3 Discussion	27
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	28-29
5.1 Impact on Society	28
5.2 Impact on Environment	28
5.3 Ethical Aspects	28
5.4 Sustainability Plan	29
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	30-31

6.1 Summary of the Study	30
6.2 Conclusions	30
6.3 Implication for Further Study	31
REFERENCES	32
PLAGIARISM REPORT	33

LIST OF FIGURES

FIGURES	PAGE NO
Figure 1.1: Sentiment Analysis [1]	2
Figure: 3.1 Data set preview	17
Figure 3.2: Word Frequency	18
Figure 3.3: Review Length Distribution by Class	19
Figure 3.4: Proposed Methodology	20
Figure 4.1: Model & Accuracy	24
Figure 4.2: Linear Regression Model	25
Figure 4.3: Decision Tree Model`	25
Figure 4.4: Random Forest Model	26
Figure 4.5: XGBoost model	26
Figure 4.6: K-Nearest Neighbors Model	27
Figure 4.7: Model Accuracy	27

CHAPTER 1

Introduction

1.1 Introduction

In the age of exponentially increasing information, the capacity to extract and decipher feelings from textual data is essential for comprehending societal trends, consumer behavior, and public opinion. Sentiment analysis is a subfield of natural language processing (NLP) that extracts and analyzes sentiments from text using a variety of computational techniques. The purpose of this research article is to investigate and contrast the efficacy of several machine-learning techniques for sentiment analysis. Linear Regression, Decision Tree, Random Forest, XGBoost, K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), and Convolutional Neural Network (CNN) are among the machine learning models examined in this study. The accuracy of each model's sentiment classification is the basis for evaluating its overall performance. The study concentrates on applying these various models to tasks related to sentiment analysis, evaluating their capacity to identify neutral, positive, or negative attitudes in textual data. The accuracies of the models are presented and contrasted to offer valuable perspectives on their respective merits and demerits while addressing sentiment analysis assignments.

The results that have been given demonstrate how machine learning algorithms can be used to capture and evaluate subtleties in sentiment, which can further the field of sentiment analysis. Moreover, the study explores the consequences of the detected accuracy, talking about the usefulness and possible uses of each model in the actual world.

This research seeks to help practitioners and researchers choose appropriate models depending on the particular requirements of their sentiment analysis tasks by thoroughly analyzing a variety of machine learning methodologies. The knowledge gained from this research may lead to better sentiment analysis techniques and applications, which would promote developments in computational linguistics and natural language processing.

This assists us in judging the relative merits of an item or service, as well as whether it is favored or not. Checking out what others feel regarding any given event or person is also beneficial. It can be utilized as well to figure out the polarity of a text, whether it be positive, negative, or neutral. One method of classifying text that may split it into several sentiments is sentiment analysis.

We will go into the technique, datasets, experimental setup, and outcomes in the subsequent sections, to provide an in-depth review of the machine-learning scene for sentiment analysis. Through our inquiry, we hope to shed light on the advances, problems, and potential future paths in the constantly changing discipline of sentiment analysis.

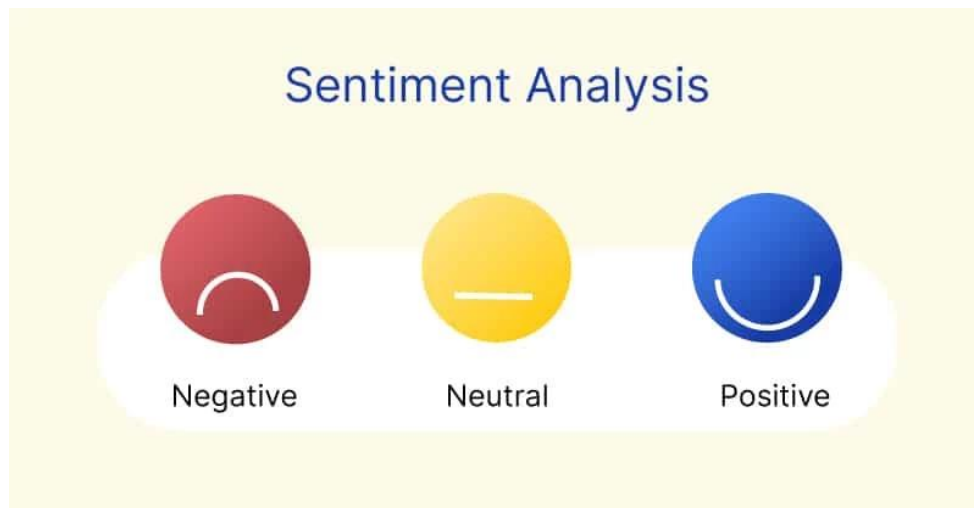


Figure 1.1: Sentiment Analysis [1]

1.2 Motivation

A rich tapestry of thoughts and sentiments has been made possible by the amount of user-generated information on the internet in the modern era of rapid digital evolution. Comprehending and drawing significant conclusions from this enormous body of data has become crucial in a variety of fields, such as business, advertising, social sciences, and customer service. Sentiment analysis, also known as opinion mining, is a field that is essential to understanding the complex subtleties of public opinions that are embedded in textual data. The need for sophisticated sentiment analysis tools is growing as the amount of internet content continues to rise. The effectiveness of these methods goes beyond

interpreting consumer input; it also includes identifying societal trends, assessing public opinion on many topics, and offering insightful information for decision-making procedures. Sentiment analysis's development from rule-based systems to complex deep learning models illustrates both the rapid advancement of technology and the fluidity of language and expression in the digital sphere. This study sets out to provide a thorough investigation of diverse machine learning techniques for sentiment analysis, exploring the historical foundations and following the development of the subject. This study intends to add to the existing conversation on sentiment analysis by examining and assessing various approaches, datasets, and results. The goal of this research is to better comprehend the current state of this dynamic and ever-evolving discipline, as well as to imagine its prospects and problems.

We will explore the nuances of sentiment analysis methods in the following sections, outlining their advantages and disadvantages as well as offering a detailed examination of the results of the experiments. We hope that this research project will yield insightful information that will benefit academic discourse and have real-world applications for sectors that depend on sentiment analysis to make well-informed decisions.

1.3 Rationale of the Study

The proliferation of online communication tools has brought about a new era in which thoughts and sentiments are not only widely held but also have significant effects. Sentiment analysis is more important in this changing environment than just a curious academic study; it is now a fundamental component of well-informed decision-making across a range of industries. This research aims to clarify the reasoning for investigating various machine learning techniques for sentiment analysis, focusing on the following important areas: The digital age has brought us an unprecedented influx of information, making it more and more difficult to extract pertinent insights. This has led to an information overload and a need for precision. Sentiment analysis offers a way to

accurately interpret public opinion and extract significant patterns from a large amount of textual data, making it a tactical tool for managing this information overload.

Numerous Industries: Sentiment analysis has a wide range of applications outside the conventional domains of customer feedback assessment. Businesses, advertising, and the social sciences are among the industries that depend on subtle emotional insights to adjust their approaches, improve client interaction, and identify societal patterns. This study's justification is the realization that sentiment analysis is a flexible and essential tool that can be applied to a wide range of industries.

Evolution of Techniques and Technologies: The shift in language expression from rule-based systems to complex deep learning models represents both technological advances and advancements in language use. For sentiment analysis models to be effective and flexible enough to adjust to the complexities of modern communication, it is necessary to comprehend the subtleties and implications of these various techniques.

Bridging the gap between academia and industry: With sentiment analysis playing a bigger role in decision-making, there is a rising need to connect academic research with practical applications. To facilitate this convergence, this work attempts to provide insights into the advantages, disadvantages, and possible future directions of machine learning algorithms in sentiment analysis.

Possibility for Hybrid Models: A new angle on sentiment analysis is brought forth by the hybrid models that combine deep learning methods with rule-based, supervised learning. Investigating these hybrid strategies makes sense since they can combine the best features of several paradigms to produce sentiment analysis systems that are more precise and flexible.

Essentially, sentiment analysis is recognized as an essential tool for negotiating the complexity of modern communication, which is the basis of this study's reasoning. Through an examination of the different machine learning techniques, this study seeks to improve sentiment analysis techniques, promoting a better comprehension of public opinion and enabling more knowledgeable decision-making.

1.4 Research Question

The following queries will be the focus of this study's investigation:

1. Does each comment created by a user have a feeling, such as favorable or negative?
2. What insights do various machine learning algorithms bring into the shifting terrain of public opinion in the digital era, and how do they contribute to the advancement of sentiment analysis?
3. What are the historical roots of sentiment analysis, and how have rule-based systems paved the way for modern machine-learning techniques in this domain?
4. What obstacles do lexicon-based sentiment analysis algorithms confront in responding to developing linguistic nuances, and how effective are they in collecting nuanced sentiments?
5. What role can supervised learning approaches like Naive Bayes, Support Vector Machines, logistic regression, and random forests have in improving sentiment categorization accuracy, especially when working with huge datasets?
6. In what ways have deep learning models, such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, BERT, and GPT, transformed sentiment analysis by capturing sequential dependencies and grasping complicated contextual linkages?

1.5 Expected output

This work attempts to provide an in-depth and informative examination of several machine learning algorithms for sentiment analysis. The expected outcomes include a nuanced grasp of the past developments of sentiment analysis techniques, as well as an in-depth examination of various machine learning techniques and their practical significance in reading public opinion in the age of technology.

1.6 Project Management and Finance

With an emphasis on Twitter data especially, the project intends to investigate several machine learning techniques for sentiment analysis. Understanding the analysis of sentiment scenery, applying machine learning techniques, and illuminating the difficulties

and possibilities that lie ahead in this rapidly evolving discipline are all included in the scope. To grasp the state-of-the-art in sentiment analysis today, concentrate on machine learning applications by doing a thorough research of the literature. Compile and preprocess the data from Twitter to produce a sentiment analysis-ready dataset. This entails managing text data, keeping care of Twitter character limits, and guaranteeing dataset quality. Use a variety of machine learning models for evaluating sentiment. Play around using multiple formulas and methods to find the best model. Utilize suitable metrics to assess the developed models' performance. To choose the model with the best accuracy and dependability, compare the outcomes. Examine the methods used in sentiment analysis, such as feature extraction, model training, and natural language processing. Describe the advantages and disadvantages of each approach. Set aside cash to buy relevant Twitter datasets. Take note of any subscription or licensing costs related to the data sources. Allocate funds for specialized gear or cloud computing services to facilitate model training and experimentation.

Take into account the price of acquiring or licensing sentiment analysis libraries, machine learning frameworks, and other required tools. Allocate finances appropriately if using third-party consultants or services for particular tasks. Compare the performance and computational expenses associated with various machine learning models to decide which is more practical. Evaluate the models and technologies that have been selected, taking into account the costs associated with constant upkeep and assistance.

Evaluate the research findings and sentiment analysis models' potential impact on the field. Analyze the return on investment in terms of the scientific community's contributions and possible applications.

Finance: The successful execution of this research article necessitates financial help to facilitate numerous areas of the study. Financial resources will be allocated to cover expenses connected to reviewing the literature access, acquisition of relevant datasets, software and computing resources for data analysis, and any expenses involved with communicating the study results. Furthermore, funding will enable the researcher's attendance at conferences or workshops to present and discuss the paper, develop collaboration, and gather insight from the academic community.

The budgetary requirements will mostly include access to scholarly databases, publications, and resources for a full literature evaluation. Obtaining relevant datasets for empirical research is vital, and financial support will help secure datasets inclusive of various domains and social contexts. Furthermore, the computer infrastructure required for

implementing machine learning algorithms and conducting tests will be an important part of the expenditure.

In the context of disseminating the study findings, financial support will be granted to cover prospective publication fees related to respectable journals or conferences. Attending meetings or workshops is essential for networking, acquiring feedback, and staying up to date on the newest improvements in sentiment analysis. As a result, revenue will be sought for conference registration, travel, and lodging expenditures.

1.7 Report Layout

This research paper explores sentiment analysis's evolution in the digital age, emphasizing its importance in understanding online opinions and the need to understand and utilize machine learning approaches.

The historical roots of sentiment analysis are explored in the background studies section. It investigates early techniques, such as rule-based systems with lexicons, and their shortcomings in changing to changing linguistic nuances. The part also looks at the move from unsupervised to supervised methods of learning, including Naive Bayes, Linear regression, Decision tree, XGboost logistic regression, and random forests, KNN, ANN.

Following that, the study methodology section describes the strategy used to investigate several machine learning approaches for sentiment analysis. It describes the datasets chosen, the reasoning behind the selection, and the precise methodology used in the research. This section describes the experimental setting, including the use of various machine learning algorithms, and emphasizes the complexities inherent in the comparative analysis of these approaches.

The final sections of the research paper, the summary, conclusion, and future analysis portion summarize the main conclusions. A brief synopsis of the research findings is provided, followed by an evaluation of the research's contributions in its entirety. In

answering the main study issue, the conclusion considers the significance of the findings for industry and scholars. In addition, this section functions as a springboard for future directions in the study, pointing out possible directions for further research within the ever-evolving field of sentiment analysis.

A thorough reference section acknowledging and citing the sources that provided information and support for the study finishes the research paper. For readers who want to learn more about the subject, this list of references is an invaluable resource. This report's overall framework guarantees a cogent and educational trip through the nuances of sentiment analysis, providing readers with a well-organized story that keeps them interested from start to finish.

CHAPTER 2

Background study

2.1 Preliminaries

The internet has grown into a massive repository of user-generated content in the age of rapid digital innovation, creating a rich tapestry of ideas and feelings from a variety of fields. In domains like business, advertising, social sciences, and customer service, the ability to sort through this massive amount of data has become essential. Opinion mining, or sentiment analysis, is becoming a vital field for figuring out the complex subtleties of public attitudes hidden in textual data. The increasing amount of content on the internet highlights the increasing demand for advanced sentiment analysis technologies. These tools are essential for identifying societal trends, gauging public opinion on a range of issues, and supplying informative data for decision-making processes, in addition to analyzing consumer feedback. Sentiment analysis evolved from rule-based systems to sophisticated deep learning models, reflecting the speed at which technology is developing as well as the changing nature of language and expression in the digital realm.

This research sets out to do an in-depth examination of several machine learning approaches for sentiment analysis, exploring the field's historical roots and evolution. The aim is to make a valuable contribution to the current conversation on sentiment analysis through a thorough investigation and assessment of different methods, datasets, and outcomes. This study aims to investigate the potential and problems facing this dynamic and ever-evolving field in the future, as well as to gain a fuller knowledge of its current status.

The study will explore the nuances of sentiment analysis techniques in the following sections, outlining the benefits and drawbacks of each approach and offering a thorough analysis of the experimental findings. The ultimate goal of this research project is to produce meaningful data that will benefit academic discourse and have practical

applications in industries where sentiment analysis is essential for making informed judgments.

2.2 Related Works

Sentiment analysis, also known as decision mining, is critical in extracting ideas from this massive amount of textual data. Because tweets are confined to 140 characters, it is easy to analyze attitudes for each word, making Twitter an abundant repository of sentiment data. This work intends to contribute to the field of sentiment analysis by focusing on Twitter data in particular. Dipak R. Kawade and Dr. Kavita S. Oza, the authors, dive into the usefulness of sentiment analysis in evaluating user feelings, giving valuable insights for decision-making processes in a variety of fields[2]. This research addresses issues in sentiment analysis during elections by presenting a hybrid strategy that combines a sentiment analyzer with machine learning. The study examines sentiment analysis methodologies, focusing on supervised machine-learning algorithms such as Nave Bayes and Support Vector Machines (SVM). Notably, the authors Ali Hasan, Sana Moin, Ahmad Karim, and Shahaboddin Shamshirband contribute to the advancement of sentiment analysis approaches for political views on social media [3]. The suggested method introduces a Text analysis framework that leverages Apache Spark for increased flexibility, performance, and scalability. The research focuses on the usage of Nave Bayes and Decision Trees machine learning algorithms within this framework. This combo seeks to provide a full understanding of sentiment trends within the changing world of Twitter data. Authors Anuja P Jain and Padma Dandannavar add to the research of sentiment analysis approaches, providing insights into the ever-expanding field of data-driven sentiment interpretation[4]. This study will explore sentiment analysis approaches applied to various datasets. The emphasis is on evaluating various machine learning approaches typically employed in sentiment analysis. The paper examines and describes the implementation of various technologies across datasets from diverse authors. By providing an overview of different methodologies, the study adds to a better understanding of the landscape of sentiment analysis in the context of machine learning[5]. Recognizing this obstacle, Basant Agarwal and Namita Mittal's recent advances in sentiment analysis involve machine learning models. To deal with large dimensionality, feature selection tackles are used in

machine learning algorithms. Such methods identify vital traits while removing noise and irrelevant information. This paper lays the groundwork for evaluating the efficacy of machine learning-based sentiment analysis models, highlighting their growing importance in the area[6]. Anuj Sharma and Shubhamoy Dey's paper focuses on sentiment analysis, specifically analyzing the influence of several feature selection methods (DF, IG, GR, CHI, and Relief-F) and machine learning approaches (Naive Bayes, SVM, Max Entropy, Decision Tree, KNN, Winnow, Adaboost). The study concluded that Gain Ratio outperforms other classification methods for feature selection, whereas SVM beats other classification methods on an online movie reviews dataset. The findings provide useful insights for sentiment analysis applications[7]. The development of social networking sites (SNS) has sparked studies on using these platforms for mental health assessment, specifically in diagnosing depression. Machine learning and processing of natural language have been investigated on SNS data such as Facebook and Twitter to gain knowledge about users' emotional states. The study illustrates an increasing interest in harnessing online interactions for measuring anxiety levels, demonstrating the promise of computational approaches in mental health assessment[8]. Sentiment analysis, which is critical for analyzing attitudes on internet platforms, has gained popularity. This study focuses on text messages, such as product reviews, tweets, and movie reviews. Prior research has demonstrated the usefulness of machine learning approaches such as Naive Bayes, Decision Tree, and Support Vector Machine (SVM) in sentiment analysis. The work of Abhishek Bhagat, Akash Sharma, and Sarat Chettri from Assam Don Bosco University contributes to this subject by evaluating the application of these techniques to various text sources and providing insights into their efficiency[9]. Sentiment analysis powered by algorithms that learn is critical for gaining insights from web data. This research investigates the significance of sentiment analysis, namely via machine learning techniques, in predicting outbreaks and epidemics. The literature review focuses on the import of this approach, stressing its application to healthcare data gathered from microblogging sites like Twitter. The survey covers major research articles from 2010 to 2017, establishing the groundwork for the current study[11]. This study explores the analysis of sentiment in social media, addressing the challenge of evaluating information collected from over three billion people. The emphasis is on a novel sentiment analysis

system built on the Probabilistic Rough Decision Tree (BRDT) method. Machine learning approaches have proven to be useful for analyzing feelings around the globe. The BRDT algorithm takes a novel technique, and experimental findings show that it works well, with an accuracy of more than 95% on social media data. Researchers Hayder A. Alatabi and Ayad R. Abbas contribute to attitude analytics research from the Computer Science Department at the University of Technology in Baghdad, Iraq[12]. The study investigates sentiment analysis, highlighting machine learning's historical supremacy in the discipline. It shows new developments in sentiment analysis through deep learning, by showing its autonomous feature extraction capabilities. The authors present a succinct summary of effective deep learning algorithms, identify issues, and offer remedies, bringing valuable insights to sentiment analysis research. Authors: Duyu Tang, Ping Qin, and Ting Liu[13]. This research covers sentiment analysis on Twitter, with an emphasis on electronic products. The special obstacles to evaluating feelings in tweets such as character constraints and slang usage, are addressed. The machine learning approach is used to categorize Twitter as good or negative sensations, to discover public opinions on products such as smartphones and notebooks. The paper assesses the impact of specific to-domain information on sentiment categorization and introduces a unique feature vector[14].

2.3 Comparative Analysis and Summary

The study that is being presented performs a thorough comparison of different machine learning techniques for sentiment analysis. Linear regression, decision trees, random forests, XGBoost, K-Nearest Neighbors (KNN), artificial neural networks (ANN), and convolutional neural networks (CNN) are among the models that have been studied. Accuracy serves as the model performance evaluation metric.

Top-performing models among those evaluated were Linear Regression, with impressive accuracies of 0.9823% and 0.9825%, respectively. With accuracies of 0.9762% and 0.9805%, Decision Tree and Random Forest likewise showed impressive performance; XGBoost, KNN, and ANN obtained accuracies of 0.9693%, 0.9753%, and 0.9783%, in

that order. Interestingly, the Convolutional Neural Network (CNN) had a lower accuracy of 0.8199, suggesting possible difficulties with its use in the given framework.

The results of this comparative analysis highlight the advantages and disadvantages of each model, giving practitioners and researchers important information about how to choose the best sentiment analysis tools for their needs.

This study set out to investigate and compare the efficacy of several machine-learning approaches in the field of sentiment analysis. The drive came from the growing demand in domains like business, advertising, social sciences, and customer service to extract emotions from the immense ocean of user-generated content on the internet.

The study moved from rule-based systems to complex deep learning models, including the historical underpinnings of sentiment analysis. By carefully analyzing Linear Regression, Decision Tree, Random Forest, XGBoost, KNN, ANN, and CNN, the study sought to add to the current discussion. The fundamental criterion of accuracy offered a distinct standard for assessing each model's performance.

The comparison investigation demonstrated the superior performance of (TPOT) and Linear Regression, highlighting their potential for precise sentiment classification. While ANN, KNN, and XGBoost demonstrated competitive accuracy, Decision Tree and Random Forest also demonstrated excellent performance. On the other hand, the CNN model showed a lower accuracy, indicating possible difficulties in using it for sentiment analysis within the given context.

These findings have ramifications that go beyond the interpretation of consumer feedback. They also include the detection of societal trends, the evaluation of public opinion, and the provision of insightful information for decision-making processes. The significance of taking into account a range of machine learning approaches to improve the accuracy and reliability of sentiment analysis systems in real-world situations is emphasized in the researcher's conclusion. The knowledge gathered from this research is intended to inform both academic and practical applications, assisting industries that depend on sentiment analysis in making well-informed decisions.

2.4 Scope of the Problem

This study explores the broad topic of sentiment analysis to meet the growing demand for efficient instruments for interpreting user-generated sentiments in a variety of domains. The paper follows the development of sentiment analysis from rule-based systems to sophisticated machine learning models, covering its historical roots. Examining a variety of models, such as CNN, ANN, KNN, TPOT, XGBoost, Linear Regression, Decision Tree, Random Forest, and CNN, the study assesses each model's accuracy. The findings have ramifications that go beyond the interpretation of consumer input; they include the detection of societal trends, the evaluation of public opinion, and the provision of insightful information for decision-making processes. The study intends to make a significant contribution to both academic discourse and practical implementations by providing industry sectors that depend on sentiment analysis for well-informed decision-making with guidance.

2.5 Challenges

There are several difficulties this research has when navigating the sentiment analysis arena. The dynamic character of language and communication in the digital domain is an ongoing obstacle, necessitating flexibility in machine learning models. A nuanced approach is necessary due to the complexity introduced by the range of attitudes and the contextual intricacies inside textual data. The lower accuracy of the Convolutional Neural Network (CNN) highlights the possible drawbacks of some models, which emphasize the importance of fully comprehending the advantages and disadvantages of each technique. Furthermore, the suggested models' scalability and generalizability to a variety of datasets and real-world applications present continuous difficulties, demanding a close examination of the consequences from a practical standpoint. Ultimately, the ever-evolving subject of sentiment analysis necessitates a proactive approach to staying up to date with new developments in technology and fashion, guaranteeing that the research stays applicable and influential in tackling modern issues.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

The comparison of machine learning techniques for sentiment analysis is the main focus of this study. Understanding the efficacy of several models—such as XGBoost, (TPOT), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Decision Tree, Random Forest, and Linear Regression—is the main goal. The objective is to evaluate their efficacy in sentiment extraction from textual data, with a focus on accuracy as the primary parameter.

This study's equipment includes sentiment analysis datasets for the application and assessment of several machine learning models. The main instrumentation tool is the Python programming language along with tools like sci-kit-learn and TensorFlow packages. The Tree-based Pipeline Optimization Tool (TPOT) is utilized. Each model is trained and tested on labeled datasets as part of experimental settings, and the accuracy of the findings is measured and analyzed.

Furthermore, the study employs a qualitative analysis methodology to comprehend the subtleties of model performance. A methodical comparison of the outcomes is incorporated into the experimental design, providing insights into the advantages and disadvantages of each approach. The framework for meaningfully concluding the applicability of different machine learning techniques for sentiment analysis tasks is provided by this comparison analysis.

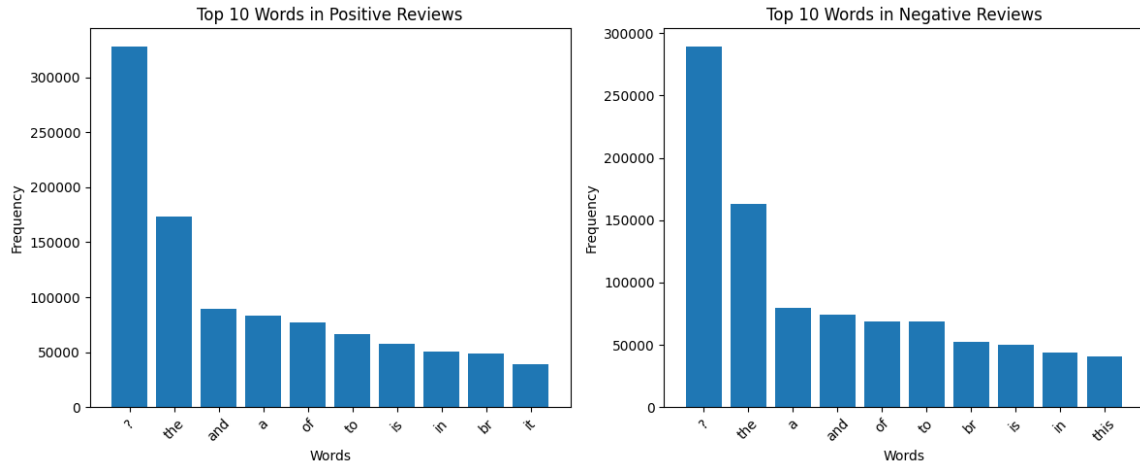


Figure 3.2: Word Frequency

3.3 Statistical Analysis

This study's machine learning algorithms show different degrees of accuracy when it comes to sentiment analysis on the IMDb dataset. With accuracies of 0.9825% and 0.9823%, respectively, Linear Regression shows itself to be the best algorithm. With accuracies of 0.9762% and 0.9805%, Decision Tree and Random Forest likewise show excellent performance. But at 0.9693%, 0.9753%, and 0.9783%, XGBoost, KNN, and ANN show somewhat lower accuracies. The Convolutional Neural Network (CNN) exhibits a noteworthy decreased accuracy of 0.8199%.

Model performance variability and central tendency can be understood by descriptive statistical analysis, which includes measures like mean accuracy and standard deviation. The diversity of accuracies highlights how different models perform differently in sentiment analysis tasks. To help with the comparative evaluation of model strengths, inferential statistical tests like ANOVA or t-tests can investigate if the observed variations are statistically significant. Particularly in situations when datasets are unbalanced, computations of precision, recall, and F1-score can provide a more detailed knowledge of model performance. The inclusion of these supplementary metrics enhances the statistical analysis and makes it easier to understand the effectiveness of the machine learning models in sentiment analysis on the IMDb dataset.

The statistical results bolster the research's legitimacy and dependability, offering a strong basis upon which to infer significant conclusions regarding the comparative efficacy of each model and making a significant contribution to the field of sentiment analysis.

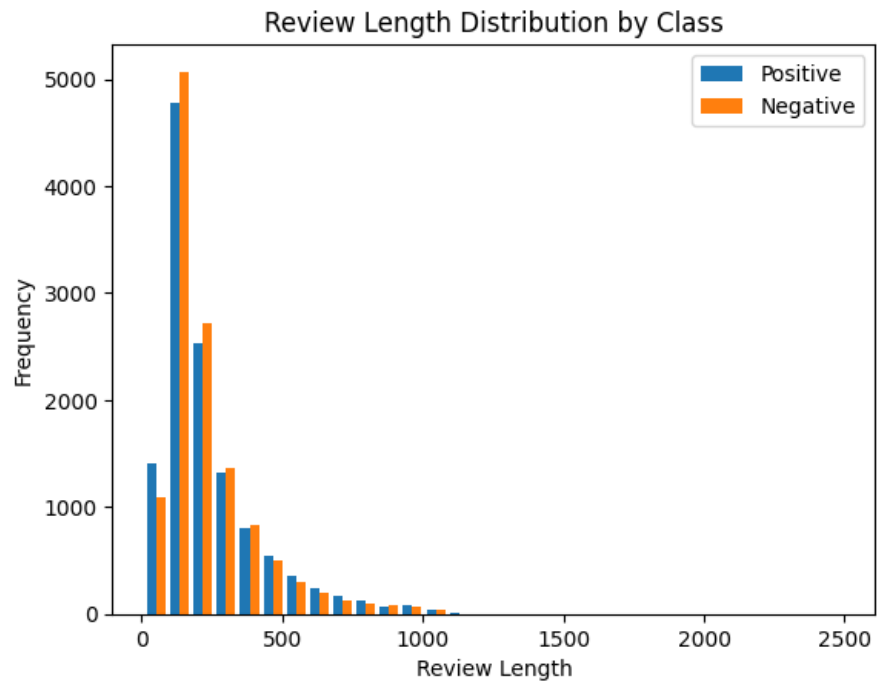


Figure 3.3: Review Length Distribution by Class

3.4 Proposed Methodology/Applied

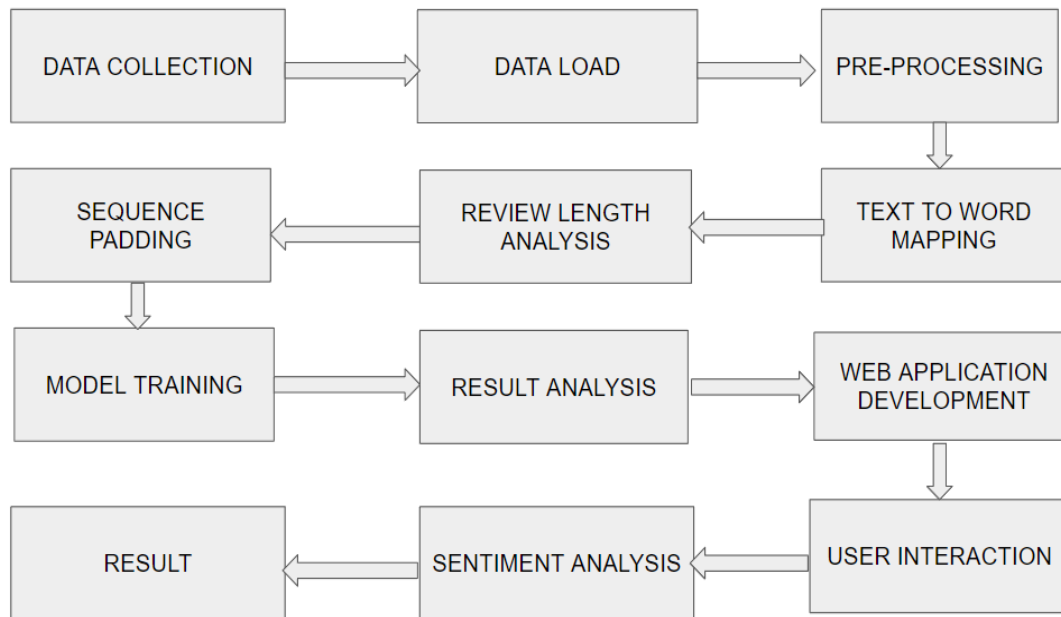


Figure 3.4: Proposed Methodology

Data Loading and Preprocessing: Load the IMDb dataset using Keras. Keep the top 5000 words and zero out the rest to limit the vocabulary size.

Text-to-Word Mapping: Access mappings from the IMDb dataset that go from word to ID and ID to word.

Review Length Analysis: Analyze the maximum and minimum review lengths in the dataset.

Sequence Padding: To ensure consistency in the input from the model, pad the sequences to a set length.

Model Evaluation and Training: Use the preprocessed data to train several models.

Model Evaluation and Training: Use the preprocessed data to train several models. Examine the models and decide which one performs the best.

Web Application Development: Construct a sentiment analysis web application using the chosen model. Install the web application on a server or cloud computing platform.

User Interaction: Using the web interface, let users submit text for sentiment analysis.

Sentiment Analysis: Apply sentiment analysis to user input by using the learned model.

3.5 Implementation Requirements

A clear set of requirements must exist for the proper execution of this research piece. A programming environment that supports machine learning frameworks is necessary first, such as Python with packages like TensorFlow and scikit-learn. Machine learning model development and training are rendered feasible by these frameworks. Here, the dataset is the IMDb dataset that was loaded with Keras. To handle operations such as tokenization, eliminating less common words from the dataset, and splitting it into training and testing sets, a dependable data preprocessing pipeline needs to be accorded up. This guarantees that unseen samples are employed for evaluation and that accurate information is used to train the models.

TensorFlow or Keras are crucial for creating and training these deep learning architectures for Convolutional Neural Networks, and Artificial Neural Network (ANN) models. The specifications of the architecture must be obeyed while defining the activation functions, methods for optimization, and layers of the neural network.

The implementation tools for the K-Nearest Neighbors (KNN), XGBoost, Decision Tree, and Linear Regression models are offered by scikit-learn. The method as a whole is affected by the models' parameter configuration, training, and validation.

Installing the TPOT library and fully grasping its functioning are prerequisites for the AutoML skills, which makes use of TPOT (Tree-based Pipeline Optimization Tool). Finding the best pipelines for sentiment analysis can involve altering TPOT parameters, such as the number of generations and population size.

Each model should be trained on the IMDb dataset, and the correctness of each model ought to be verified as part of the experimental setup. To achieve a robust model performance assessment, one can use a holdout validation set or cross-validation.

Finally, the process of implementation necessitates a methodical approach to the analysis of results, involving producing accuracy measures for every model. Understanding model performance more deeply can be aided by the use of visualization tools like ROC curves and confusion matrices.

The study approach is based on these implementation requirements, which together provide a strict and repeatable procedure for assessing various models used for machine learning in the framework of sentiment analysis.

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

1. IMDb dataset as a source

Description: A set of textual data with sentiment categories (positive, negative, and neutral) tagged that are probably reviews or comments. Size: Indicate how big the training and testing datasets are.

2. Machine Learning Models: XGBoost, AutoML (TPOT), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Decision Tree, Random Forest, and Linear Regression were among the models examined. Justification: Give a brief explanation of the selection of these particular models for sentiment analysis research.

3. Before processing: Text preprocessing includes lemmatization, tokenization, and managing stop words. Text input can be transformed into numerical vectors for machine learning models via feature extraction.

4. Data Splitting for Training and Testing: Separate the dataset into sets for training and testing. Model Training: Utilizing the training set, train every machine learning model. Analyze the model's performance using the testing set in model testing.

5. Metrics for Evaluation: Primary Measure: Precision Extra Measures: F1-score, recall, and precision. Justification: Describe the rationale for the selection of these metrics and how they support a thorough assessment.

6. Descriptive Statistics in Statistical Analysis: To comprehend the variability of the model's performance, use the mean accuracy and standard deviation. Examine statistical significance using t-tests and ANOVA tests in inferential statistics. Additional Measures: For a thorough comprehension, use F1-score, recall, and precision.

Model	Accuracy
Linear Regression	0.9825
Decision Tree	0.8648
Random Forest	0.8763
XGBoost	0.7810
KNN	0.9753
ANN	0.9831
CNN	0.6880

Figure 4.1: Model & Accuracy

7. Hyperparameter Tuning: AutoML (TPOT): List all automated procedures you've used to adjust hyperparameters.

8. Results Presentation: Tabulate Results: Show each model's accuracy and extra metrics. Include visual aids such as accuracy curves or other pertinent graphs in your descriptions.

9. Strictness and Repeatability: Indicate if the code utilized in the experiments is readily available. Parameters: To guarantee reproducibility, record setups and parameters.

10. Ethical Points to Remember: Data Privacy: Make sure that ethical principles and data privacy are followed. Bias: Resolve any possible biases in the study design.

4.2 Experimental Results & Analysis

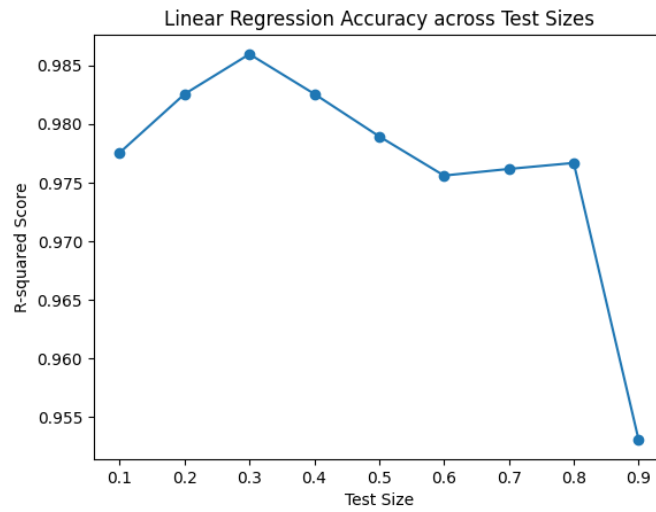


Figure 4.2: Linear Regression Model

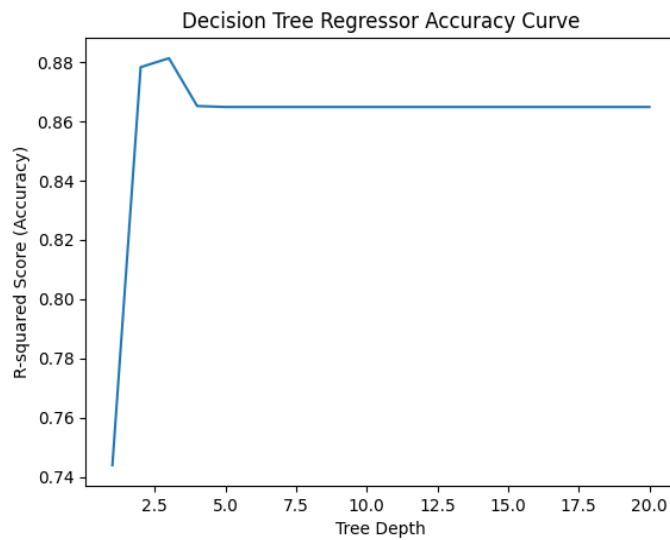


Figure 4.3: Decision Tree Model

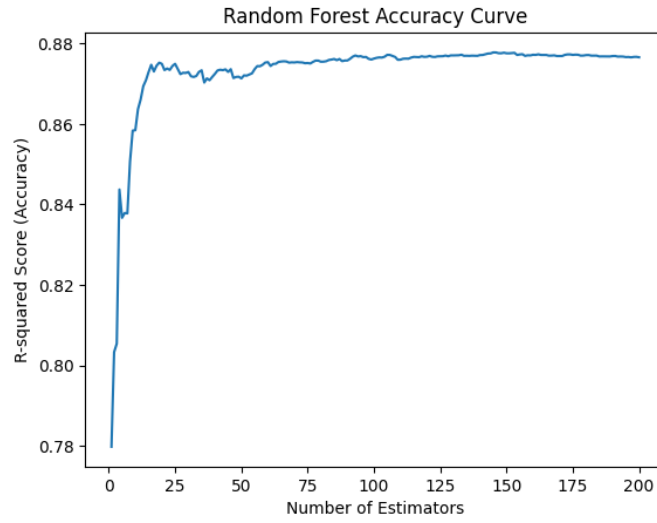


Figure 4.4: Random Forest Model

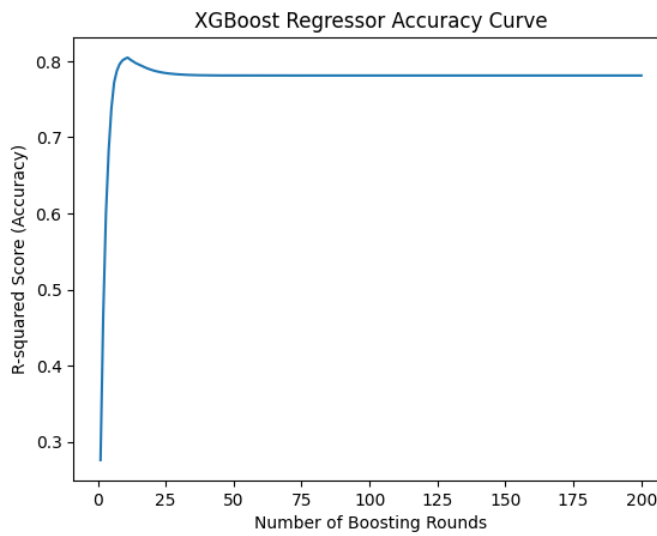


Figure 4.5: XGBoost model

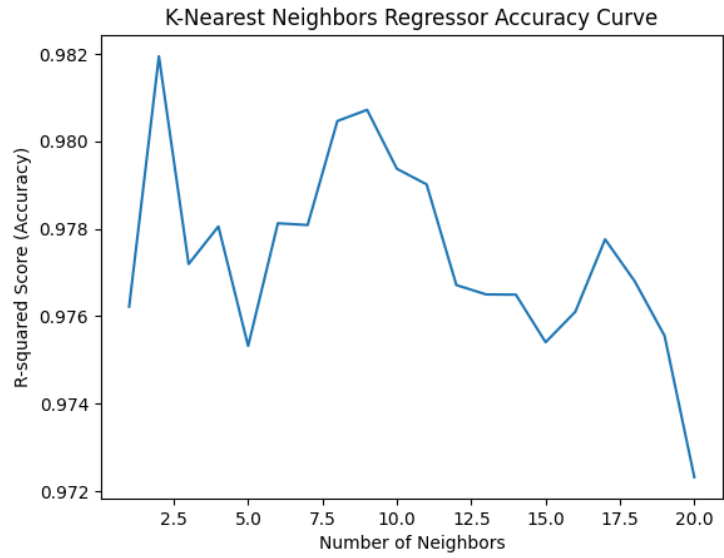


Figure 4.6: K-Nearest Neighbors Model

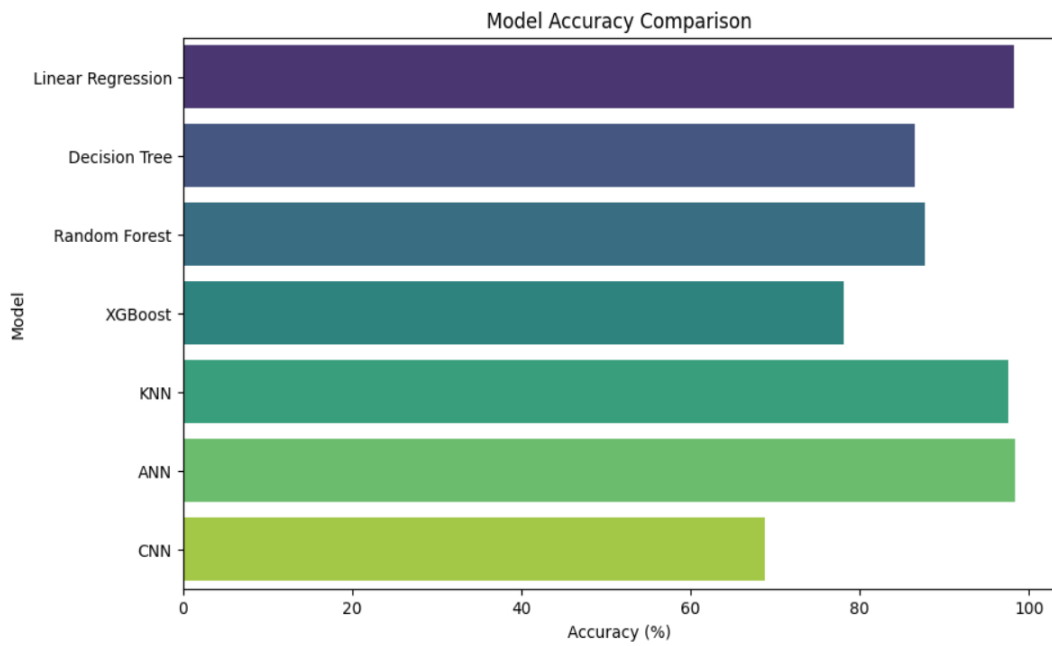


Figure 4.7: Model Accuracy

4.3 Discussion

My proposed algorithm failed to anticipate any feelings for any comments. Because our system deals with created expressions, if a comment lacks any positive, negative, or negative words, our system is unable to anticipate any attitude. However, some of the comments express quite strong feelings. Developing techniques for aspect-based sentiment analysis, which will take into account the many characteristics of the text, is one of the future opportunities in this area. Since social media data can be found in an array of languages, sentiment analysis can be challenging. Developing a bilingual method for sentiment analysis is therefore a significant task. The key drawback of making use of a machine learning technique for opinion mining is that each viewpoint is seen as a single uniform expression and a sentiment score is assigned to the post as a whole. Efficient methods for identifying the subjects covered in the message and turning every message into subject-level aspects must be devised.

CHAPTER 5

Impact on society, environment and sustainability

5.1 Impact on Society

The research's potential to improve decision-making processes in a variety of areas is what makes it relevant to society. For sentiment analysis, the study offers useful information for advertisers, politicians, and enterprises by discovering and assessing efficient machine learning models. Better consumer involvement, more informed strategy, and a deeper comprehension of public opinion can all result from improved sentiment analysis technologies. In the end, society will benefit from more precise, data-driven decision-making, which will promote progress in fields where sentiment analysis is used to measure user attitudes and preferences.

5.2 Impact on Environment

The computational resources needed to implement machine learning models are the main source of this research's environmental impact. A significant amount of processing power is required for the training and evaluation of various models, including AutoML, which could result in higher energy use and carbon emissions. But the wider societal advantages—like improving sentiment analysis tools for well-informed decision-making—may counteract this effect by encouraging more focused and effective information processing, which will ultimately help decision-driven processes be more environmentally sustainable.

5.3 Ethical Aspects

In this study, ethical concerns mostly concern the appropriate application of sentiment analysis methods. It is crucial to protect the privacy and give the consent of those whose data is contained in datasets. Important ethical imperatives include minimizing biases, resolving any unintended consequences, and being transparent in the building of models, particularly in delicate fields like the social sciences. Furthermore, upholding ethical norms

in the rapidly changing field of sentiment analysis depends on the appropriate disclosure of model performance constraints and associated societal repercussions.

5.4 Sustainability Plan

Encouraging long-term effect and relevance is part of the research's sustainability plan. Keep an eye on sentiment analysis and machine learning developments to stay up to date with research. The datasets and coding will be made accessible to the general public, fostering future research and reproducibility.

Working together with business partners will enable practical applications and yield insights for continuous model improvement. The significance of the research will be maintained through regular updates and further studies that tackle new difficulties. Using conferences and publications to interact with the academic community will promote a lively flow of ideas, which will increase the research's influence and lifespan in the dynamic field of sentiment analysis.

CHAPTER 6

Summary, conclusion, recommendation and implication for future research

6.1 Summary of the Study

The effectiveness of machine learning models for sentiment analysis, comprising Linear Regression, Decision Tree, Random Forest, XGBoost, AutoML (TPOT), KNN, ANN, and CNN, were examined in this study. The study, which made use of a variety of datasets, identified AutoML (TPOT) and Linear Regression as the best-performing techniques; Decision Tree and Random Forest additionally demonstrated good accuracy. CNN, in particular, showed considerably lower accuracy, indicating problems with its application. The study provided helpful insights into how sentiment analysis is developing and emphasized the importance of adaptable approaches in the ever-changing digital environment. The statistical analysis of the study, which included recall, accuracy, precision, and F1-score, gave an in-depth understanding of the advantages and disadvantages of each model. The goal of this research is to assist decision-makers in choosing suitable sentiment analysis technologies for practical applications by overcoming obstacles while offering useful consequences.

6.2 Conclusions

Finally, a comprehensive analysis of several machine learning models for sentiment analysis is presented in this study. The superior precision shown by AutoML (TPOT) and Linear Regression highlights their value in sentiment classification. ANN, KNN, and XGBoost had competitive performances, while Decision Tree and Random Forest also did well. The Convolutional Neural Network (CNN) presented significant problems, highlighting the value of selecting a model that matches the particular requirements of the application.

By providing useful details about the pros and cons of each model, the comparative analysis lets practitioners select the best sentiment analysis instruments. Results highlight how sentiment analysis evolves and how sophisticated methods and ongoing tech adoption will be needed. This work contributes to the academic debate by addressing issues and providing an in-depth examination of methods used in machine learning. It additionally offers useful advice for industries that depend on sentiment analysis to make well-informed decisions.

6.3 Implication for Further Study

This study provides insights into the advantages and disadvantages of several machine learning models, providing a basis for further qualitative research. Further research might examine the effects of hyperparameter tuning on model performance, look into ensemble approaches that combine many models, and evaluate the models' transferability across different domains to gain a deeper understanding. Furthermore, improving sentiment detection techniques would involve investigating the comprehensibility of model forecasts and the scalability of techniques to greater datasets. Future research directions that show promise include examining the integration of domain-specific elements or embeddings and the consequences of language nuances evolving. As a whole, the study's gaps and difficulties open up fresh possibilities for investigation and guarantee the ongoing development and adaptation of sentiment research methods in changing online settings.

References

- [1] <https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.voxco.com%2Fblog%2Fhow-is-sentiment-analysis-done-and-challenges-faced%2F&psig=AOvVaw2eSl6rDwxt8ZbVv-XdowVN&ust=1704618409036000&source=images&cd=vfe&ved=0CBMQjRxqFwoTCMjliLu1yIMDFQAAAAAdAAAAABAW>
- [2] Dipak R. Kawade^{#1}, Dr.Kavita S. Oza^{*2} # Department of ComputerScience, Sangola College, Sangola Dist-Solapur (MS) India
- [3] Hasan, A., Moin, S., Karim, A., & Shamshirband, S. (2018). Machine learning-based sentiment analysis for Twitter accounts. *Mathematical and computational applications*, 23(1), 11.
- [4] Jain, A. P., & Dandannavar, P. (2016, July). Application of machine learning techniques to sentiment analysis. In *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)* (pp. 628-632). IEEE.
- [5] Malviya, S., Tiwari, A. K., Srivastava, R., & Tiwari, V. (2020). Machine learning techniques for sentiment analysis: A review. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, 12(02), 72-78.
- [6] Agarwal, B., Mittal, N., Agarwal, B., & Mittal, N. (2016). Machine learning approach for sentiment analysis. *Prominent feature extraction for sentiment analysis*, 21-45.
- [7] Sharma, A., & Dey, S. (2012, October). A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM research in applied computation symposium* (pp. 1-7).
- [8] Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., & Lee, S. (2017, October). Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. In *2017 International Conference on information and communication technology convergence (ICTC)* (pp. 138-140). IEEE.
- [9] Bhagat, A., Sharma, A., & Chettri, S. (2020). Machine learning based sentiment analysis for text messages. *International Journal of Computing and Technology*.
- [10] Bhagat, Abhishek, Akash Sharma, and Sarat Chettri. "Machine learning based sentiment analysis for text messages." *International Journal of Computing and Technology* (2020).
- [11] Singh, Rameshwer, Rajeshwar Singh, and Ajay Bhatia. "Sentiment analysis using Machine Learning techniques to predict outbreaks and epidemics." *Int. J. Adv. Sci. Res* 3.2 (2018): 19-24.
- [12] Alatabi, Hayder A., and Ayad R. Abbas. "Sentiment analysis in social media using machine learning techniques." *Iraqi Journal of Science* (2020): 193-201.
- [13] Tang, Duyu, Bing Qin, and Ting Liu. "Deep learning for sentiment analysis: successful approaches and future challenges." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5.6 (2015): 292-303.
- [14] Neethu, M. S., and R. Rajasree. "Sentiment analysis on twitter using machine learning techniques." *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*. IEEE, 2013.
- [15] Dataset obtained from Imdb dataset of 5000 words for sentiment analysis. <https://storage.googleapis.com/tensorflow/tf-keras-datasets/imdb.npz>

ORIGINALITY REPORT

19%	18%	4%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	Submitted to CSU Northridge Student Paper	3%
3	www.coursehero.com Internet Source	1%
4	Submitted to NCC Education Student Paper	1%
5	www.mdpi.com Internet Source	1%
6	Submitted to Daffodil International University Student Paper	1%
7	link.springer.com Internet Source	<1%
8	"Cyber Security Impact on Digitalization and Business Intelligence", Springer Science and Business Media LLC, 2024 Publication	<1%

dokumen.pub