IDENTIFYING THE RESEARCH FIELD OF A SCIENTIFIC PAPER FROM THE ABSTRACT USING DEEP LEARNING APPROACHES

BY

Sudipto Sarker
ID: 201-15-3271

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

Supervised By

Mr. Partha Dip Sarkar
Lecturer
Department of CSE
Daffodil International University

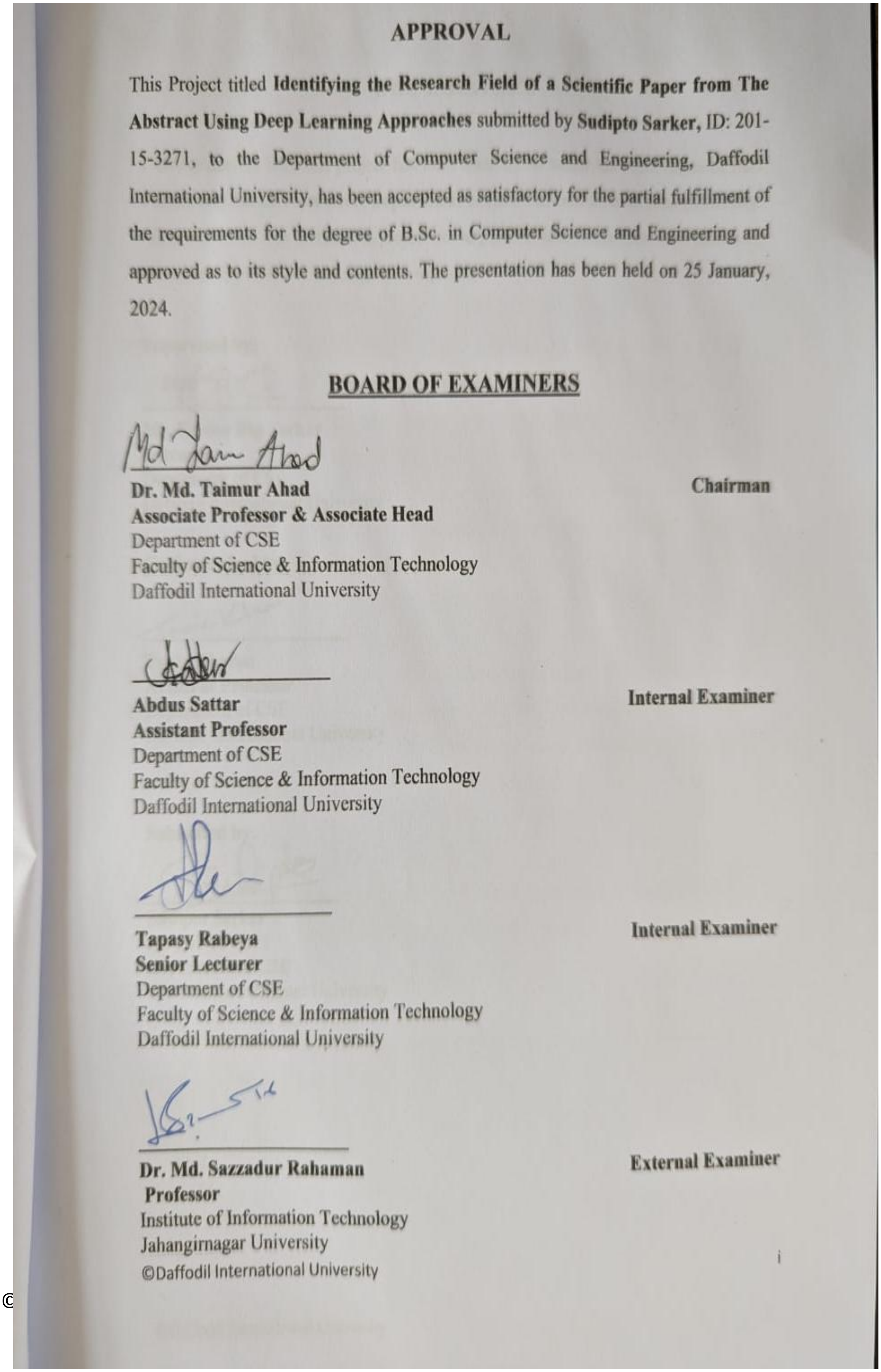
Co-Supervised By

Tanina Khatun
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 25, 2024



APPROVAL

This Project titled **Identifying the Research Field of a Scientific Paper from The Abstract Using Deep Learning Approaches** submitted by **Sudipto Sarker,** ID: 201-15-3271, 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 has been held on 25 January, 2024.

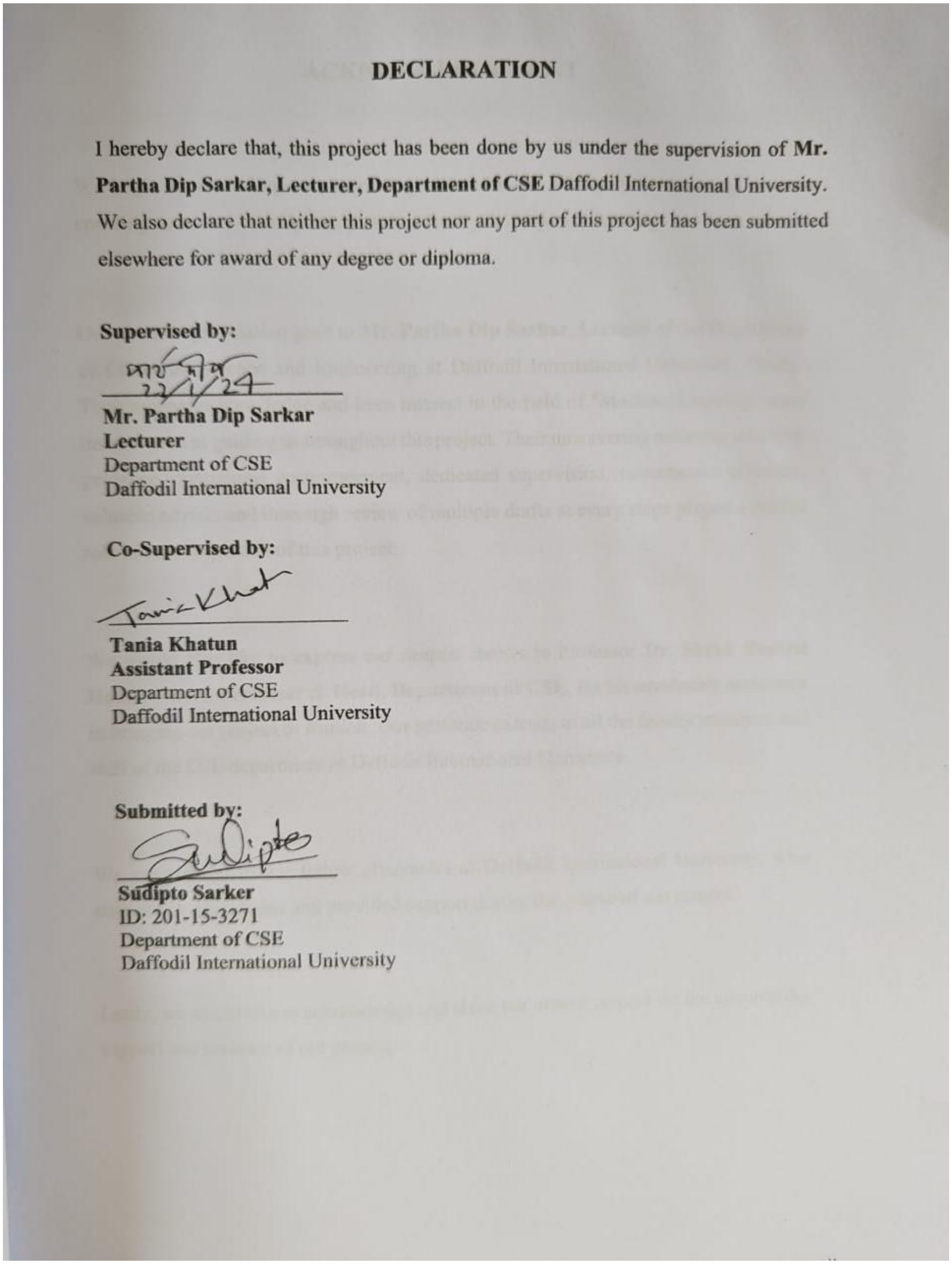
BOARD OF EXAMINERS

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

Abdus Sattar — **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Tapasy Rabeya — **Internal Examiner**
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Sazzadur Rahaman — **External Examiner**
Professor
Institute of Information Technology
Jahangirnagar University



DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Mr. Partha Dip Sarkar, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

পার্থ দীপ
22/1/24

Mr. Partha Dip Sarkar
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Tania Khat

Tania Khatun
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Sudipto

Sudipto Sarker
ID: 201-15-3271
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

We begin by extending our heartfelt gratitude to the Almighty for blessing us and enabling us to successfully complete our final year project/internship.

Our sincere appreciation goes to Mr. Partha Dip Sarkar, Lecture of the Department of Computer Science and Engineering at Daffodil International University, Dhaka. Their extensive knowledge and keen interest in the field of "Machine Learning" were instrumental in guiding us throughout this project. Their unwavering patience, scholarly guidance, continuous encouragement, dedicated supervision, constructive criticism, valuable advice, and thorough review of multiple drafts at every stage played a crucial role in the completion of this project.

We would also like to express our deepest thanks to Professor Dr. Syed Akhter Hossain, the Head of the Department of CSE, for his invaluable assistance in bringing our project to fruition. Our gratitude extends to all the faculty members and staff of the CSE department at Daffodil International University.

We are grateful to our fellow classmates at Daffodil International University, who engaged in discussions and provided support during the course of our project.

Lastly, we would like to acknowledge and show our utmost respect for the unwavering support and patience of our parents.

ABSTRACT

This study explores the application of deep learning to automate the identification of research fields within scientific paper abstracts. The goal is to create a resilient model that effectively categorizes the primary subject matter discussed in abstracts, enhancing precision and efficiency. The dataset undergoes preprocessing, tokenization, and transformation into sequences suitable for input into various models, including Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Bidirectional Long Short-Term Memory (BLSTM) cells. The trained model incorporates techniques such as word embedding and dropout, and its performance is evaluated using metrics like accuracy and the AUC-ROC score. The research addresses challenges in identifying research fields within English language abstracts, employing language-specific preprocessing and data augmentation. The results highlight the efficacy of deep learning in accurately categorizing diverse research fields within English abstracts, showcasing its potential applicability beyond English contexts. The findings contribute to advancing automated techniques for recognizing research themes, streamlining the comprehension and classification of scientific papers. Various algorithms, including ANN, CNN, BLSTM, DT, GB, ABC, RF, SVC, XGB, MNB, PA, RC, and LR were employed. Notably, the Gradient Boosting (GB) model demonstrated exceptional performance with an 83.82% accuracy rate, and the Support Vector Classification (SVC) yielded impressive results with an 83.50% accuracy rate. These outcomes were achieved through meticulous hyperparameter tuning, enhancing the overall robustness of the model.

Keywords: Abstract, Detection, Deep Learning, Algorithms, Ensemble Model, Dropout, Embedding.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Introduction

In recent world, landscape of scientific research, the ability to efficiently identify and categorize the research field of a scientific paper is crucial for scholars, academics, and researchers. The abstract of a scientific paper serves as a concise summary, encapsulating the essence of the research and its contributions to a particular field. Traditionally, this categorization process has been undertaken manually, requiring extensive domain expertise and time-consuming efforts. However, with the advent of deep learning approaches, there exists an unprecedented opportunity to automate and enhance this process, bringing about a paradigm shift in the way we comprehend and navigate scholarly literature. This study focuses on the innovative application of deep learning methodologies to address the challenging task of automatically identifying the research field of a scientific paper based on its abstract. While there has been considerable progress in natural language processing (NLP) and machine learning, adapting these techniques to the nuanced complexities of scientific language and diverse domains remains an intricate challenge. Deep learning, that excels in capturing intricate patterns within large datasets, emerges as a potent solution for tackling this nuanced task. The motivation behind this research stems from the recognition of the growing volume of scientific literature across diverse disciplines and the need for an efficient, scalable, and accurate system to navigate and comprehend this vast intellectual terrain. As scholars and researchers grapple with an ever-expanding corpus of knowledge, an automated approach to identify the research field from abstracts becomes not only a convenience but a necessity. The methodology employed in this study centers on the development and training of a deep learning model specifically tailored for the intricacies of scientific abstracts. Leveraging advanced architectures, such as recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM) cells, the model aims to discern subtle contextual cues and linguistic nuances that indicate the research field. The utilization of carefully curated datasets, incorporating abstracts from diverse scientific disciplines, enhances the model's adaptability and generalizability across varied domains. Furthermore, the study delves into the exploration of techniques to mitigate potential challenges, including the incorporation of domain-specific preprocessing methods and the consideration of cultural and contextual factors that might influence the abstraction of scientific content. In essence, this research aspires to contribute to the ongoing evolution of scholarly communication by proposing an automated, deep learning-driven paradigm for identifying the research field from abstracts. The outcomes of this study not only hold the promise of revolutionizing the way we navigate scientific literature but also offer a glimpse into the transformative potential of deep learning approaches in the broader realm of scientific inquiry.

1.2 Motivation

The motivation behind this research lies at the intersection of the escalating volume of scientific literature and the need for advanced, automated tools to efficiently navigate and comprehend this expanding intellectual landscape. In the current era, where knowledge production is rapidly accelerating across diverse disciplines, traditional manual approaches to categorizing and understanding the research field of scientific papers are proving increasingly cumbersome. As scholars and researchers confront an ever-growing corpus of knowledge, the quest for an automated system capable of swiftly and accurately identifying the research field from abstracts becomes paramount. The motivation is rooted in the recognition that such a system not only offers convenience but is fundamentally necessary for staying abreast of the vast and intricate web of scholarly information. Deep learning, as a subset of machine learning, emerges as a compelling solution for this challenge. The motivation to employ deep learning methodologies stems from their proven ability to discern intricate patterns within large datasets. Scientific abstracts, often laden with nuanced language and discipline-specific terminologies, present a unique set of challenges that demand the sophistication and adaptability of deep learning models. The overarching goal is to streamline and revolutionize scholarly communication by introducing an automated approach that significantly enhances the efficiency and accuracy of identifying the research field from abstracts. This research recognizes the transformative potential of deep learning approaches in addressing the complexities of scientific language and diverse domains, ultimately contributing to the ongoing evolution of how knowledge is accessed, comprehended, and disseminated. Moreover, the motivation extends beyond the immediate benefits for researchers and academics. A more automated and efficient system for research field identification has the potential to democratize access to knowledge, making scholarly information more accessible to a broader audience. This inclusivity aligns with the broader goals of fostering collaboration, interdisciplinary research, and the democratization of knowledge across global communities. In summary, the motivation behind this research is deeply rooted in the imperative to adapt to the evolving landscape of scientific inquiry. By leveraging the capabilities of deep learning, this study aspires to not only address the challenges posed by the expanding volume of scientific literature but also to pave the way for a more accessible, efficient, and transformative scholarly communication process.

1.3 Rationale of the Study

In this research, the rationale of this study lies in filling the gap in abstract class detection research for scientific contexts, addressing the cultural and linguistic specificities of the science papers, empowering research communities, exploring the potential of deep learning in natural language processing, and promoting a safer digital environment for all individuals communicating in the research field.

1.4 Research Question

What is the likelihood that abstract is predicted well?

How difficult is this task?

What benefits does our proposed model offer?

How can we evaluate our model for detecting abstract?

How well do the algorithms in this proposed model work?

What was the precautions of the study?

How many ways to detect abstract?

What credentials are required for this position?

1.5 Expected output

This study on abstract classification using deep learning aims to fill the gap in research pertaining to English contexts and address the specific challenges faced by the scientific research community. By focusing on abstract classification in research, we strive to provide language-specific approaches to effectively detect and combat classification. The cultural and linguistic nuances unique to the English language require tailored strategies and interventions. Leveraging the power of deep learning techniques, specifically RNNs with LSTM cells, we seek to develop a model capable of accurately identifying instances of abstract classification. The outcomes of this study will contribute to the development of language-specific tools and policies to prevent and mitigate abstract classification incidents, fostering a safer and more inclusive online environment for researchers.

1.6 Project Management and Finance

Our proposed model offers cost-effective utility for everyday use and has the potential to be a valuable asset in the context of abstract detection for scientific research. The practical implementation of this detection process necessitates the use of common tools, although optimal performance and exceptional results can be achieved with high-configuration tools. However, even with simpler tools, our model can still effectively serve its purpose, providing seamless operation and aiding in the support to research.

1.7 Report Layout

The structure of the report encompasses the following key sections:

Background Study: Providing an in-depth exploration of the context and relevant research in the field of hypothyroidism.

Research Methodology: Detailing the approach, tools, and techniques used to conduct the study and develop the proposed model.

Experimental Results and Discussion: Presenting the findings, outcomes, and a comprehensive discussion of the research results.

Summary, Conclusion, and Future Analysis: Summarizing the key takeaways, drawing conclusions, and outlining potential directions for future research.

References: Citing the sources and literature used to support the study and its findings

Chapter 2

BACKGROUND STUDY

2.1 Preliminaries

The precise analysis of abstract patterns employs deep learning techniques. In this segment, we investigate research related to the assessment of published research paper reports. Various computational models, including Bidirectional Long-Term Memory (BLSTM), Conventional Nural Network (CNN), Artificial Nural Network (ANN), Decision Tree (DT), Gradient Boosting (GB), Adaboost Classifier (ABC), Random Forest (RF), Support Vector Classifier (SVC), XGBoost Classifier (XGB) Multinomial Naive Bayes (MNB), Passive Aggressive (PA), Ridge Classifier (RC) and Logistic Regression (LR) are utilized. Deep learning models are implemented in this section to conduct the research. The section also highlights multiple researchers who have employed various models in their respective studies to enhance our understanding of abstract detection.

2.2 Related Works

In the domain of abstract classification, our proposed methodologies showcase a tailored approach to effectively discern and categorize scientific paper abstracts. Addressing the intricacies of this task involves an exploration of diverse research studies and methodologies that contribute to the advancement of abstract classification techniques. These investigations encompass the deployment of an array of deep learning algorithms and natural language processing (NLP) methodologies, each presenting unique strengths and perspectives. First and foremost, deep learning models stand out as a powerful tool in the endeavor to identify abstract. These models, which leverage algorithms based on decision trees, often referred to as "tree structures," enable effective decision-making processes. The application of deep learning algorithms is crucial in this context.

In the contemporary landscape, the proliferation of scientific articles across diverse disciplines is on the rise [1]. Notably, these articles often span multiple disciplines, as exemplified by Akshai and Anitha's study on early plant disease detection, which seamlessly integrates agricultural and computer science [2]. The challenge for researchers lies in efficiently identifying articles pertinent to their specific discipline amidst this vast and varied scientific literature. The existing manual categorization process is both time-consuming and prone to errors, with some articles not aligning with the designated disciplines of journal portals.

Automated classification of scientific articles into relevant disciplinary categories, coupled with insights into the interdisciplinarity within articles, emerges as a valuable solution for researchers and journal portals. While prior research has delved into text classification encompassing abstracts, sentiment, and news categories, there remains room for improvement. Notably, existing text classification approaches predominantly rely on traditional word embedding models like Word2Vec, FastText, and TF-IDF [1], [3]–[9].

The advent of advanced word embedding models, particularly the pre-trained BERT model, offers a promising avenue for enhancing natural language processing (NLP) tasks [10]. However, previous BERT-based text classifications primarily focus on English datasets, centering around sentiment or news category classification [11]–[17]. Notably, these studies often overlook the crucial text preprocessing stage, acknowledged to impact model accuracy [9], [12], [13]. Moreover, a majority of past research merely conducts a single round of hyperparameter testing, neglecting the significance of fine-tuning hyperparameters for optimal model accuracy [3], [12], [13], [16], [17].

In advancing text classification methodologies, future research should address these gaps, exploring the potential of BERT models across various languages and incorporating robust text preprocessing. Additionally, a more comprehensive exploration of hyperparameter combinations through fine-tuning is essential for achieving superior accuracy. Furthermore, acknowledging the possibility of a text belonging to more than one category introduces a nuanced perspective that warrants consideration in future text classification endeavors [1], [16], [18]–[21].

In summary, these various approaches and methodologies highlight the diversity and effectiveness of techniques employed to detect abstract type. From deep learning algorithms to NLP methods and innovative word embedding models, researchers have explored a wide array of tools and strategies to address this pressing issue. While each approach has its unique strengths and limitations, collectively, they contribute to the ongoing effort to combat abstract type effectively.

2.3 Comparative Analysis and Summary

Deep learning models are currently in high demand, and our journey to identify our specific role demanded a significant undertaking. Many related projects experienced subpar model results and limited accuracy. To attain the highest accuracy in dataset detection, we had to adopt a distinct deep learning model. This endeavor required the utilization of top-tier hardware to run these models efficiently. The process involved performing individual computations to gauge categorization rates. The integration of expensive GPUs allowed for the execution of intricate models, even though they could extend the runtime considerably.

2.4 Scope of the Problem

Abstract class prediction in the English language is multifaceted, encompassing language-specific challenges, cultural factors, diverse online platforms, and individuals of different age groups. Understanding and addressing these aspects are crucial in developing effective strategies to detect and combat abstracts. By considering the linguistic nuances, cultural dynamics, and the wide range of online platforms, we can create comprehensive solutions to mitigate abstract classification incidents and promote a safer digital environment for individuals communicating in research.

2.5 Challenges

The difficulties in detecting abstract class in research include linguistic problems, cultural quirks, changing internet infrastructures, and the constantly evolving nature of abstract practices. Overcoming these challenges requires developing language-specific algorithms, accounting for cultural contexts, adapting to platform changes, and staying updated with emerging abstract class trends. Addressing these challenges is essential to effectively detect and combat abstract class in the English language and create a safer online environment for individuals research in science.

Chapter 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

To achieve optimal accuracy with our dataset, we harnessed a variety of algorithms and hybrid models. Critical to our success were the use of cutting-edge hardware tools, particularly top-tier GPUs, and the employment of Python programming language and its associated resources. These tools, including Jupyter Notebook, Google Colab, and Anaconda, played a pivotal role in our workflow. Python, known for its versatility and rich ecosystem of libraries, enabled us to seamlessly develop and execute code, facilitating efficient data processing and model training. Furthermore, platforms like Jupyter Notebook, Google Colab, and Anaconda provided the flexibility of browser-based coding, enhancing our collaborative efforts and accessibility. Through this comprehensive toolkit and resource integration, we were able to push the boundaries of our dataset's accuracy and deliver robust results in our pursuit of excellence.

3.2 Data Collection Procedure

The dataset was collected manually from IEEE research papers, it was nearly prepared for implementation, comprising two essential components: the textual data and corresponding labels. This supervised dataset encompassed five distinct categories, facilitating a comprehensive understanding of the content. To facilitate model evaluation, the dataset underwent a division into two segments: a test set and a training set. This partitioning allocated 20% of the data for the test section, while the remaining 80% was designated for in-depth analysis and training purposes.

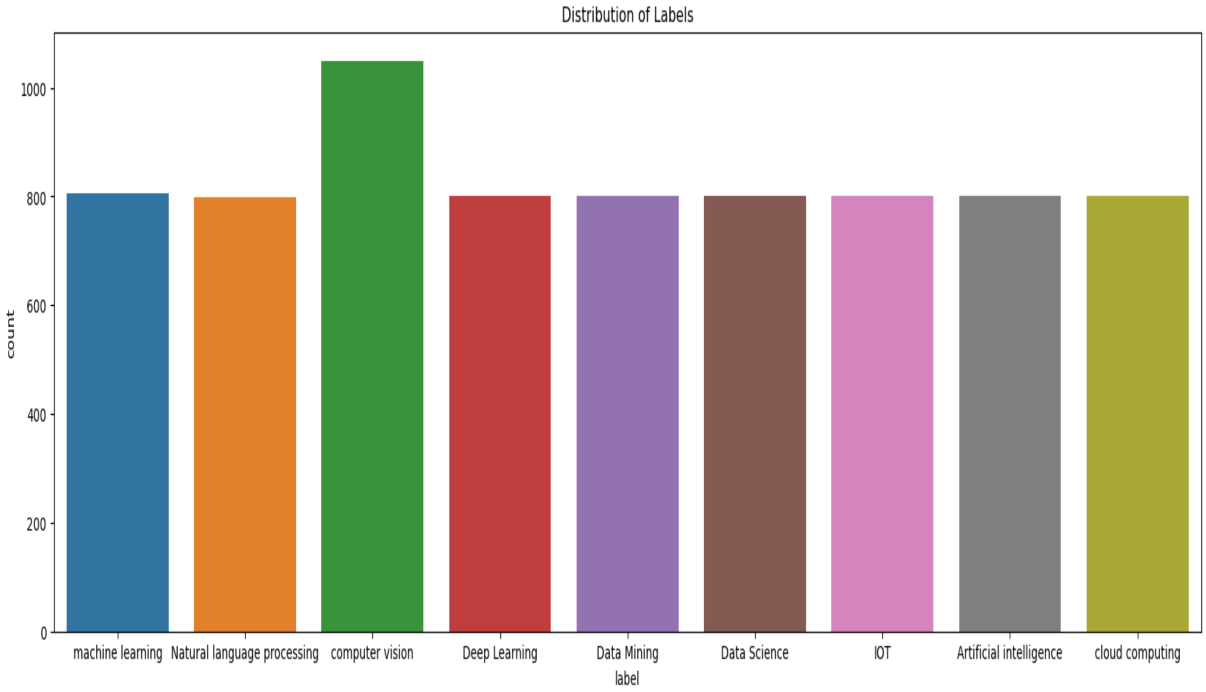


Figure 3.1: Number of target values dataset

The dataset exhibited an absence of inaccurate or missing values, including the category values. For a comprehensive understanding of the dataset's characteristics the count plot is shown in Figure 3.1.

3.2.1 Categorical Data Encoding

The transformation of categorical variables into numerical values is referred to as categorical encoding. This encoding method played a pivotal role in our research, as machine learning algorithms exclusively handle numeric data for input and output. To implement this categorical encoding approach effectively, we focused on the "Labels" column within our dataset.

3.2.2 Missing Value Imputation

This process involves replacing imputed values with missing or incomplete data identified through the analysis of data from other datasets. However, it's worth noting that our dataset had the advantage of containing no missing values, obviating the need for such imputation.

3.2.3 Handling Imbalanced Data

This technique pertains to the manipulation of the class distribution within a dataset. It accomplishes this by systematically introducing new instances into the dataset, effectively balancing the data. The primary objective is to bolster the representation of minority data, all while utilizing the entire dataset as input.

3.3 Statistical Analysis

In every research project, an essential component is the analysis section, which hinges on the development and evaluation of the algorithms utilized. In our case, since we've opted to work with an Excel file format, several preparatory steps were necessary to ensure the dataset's usability. These measures encompassed data collection and meticulous pre-processing, all of which were integral to the successful execution of our research.

In this research, we used different types of algorithms, including BLSTM, CNN, ANN, DT, GB, ABC, RF, SVC, XGB, MNB, PA, RC and LR. Among these algorithms, the GB model demonstrated exceptional performance, achieving an accuracy rate of 83.82%. The SVC also delivered impressive results, with an accuracy rate of 83.50%. These high-performance outcomes were achieved through the meticulous application of hyperparameter tuning, further enhancing the robustness of our model.

3.4 Proposed Methodology

In our research, we have applied a diverse range of classifiers, including Bidirectional Long-Term Memory (BLSTM), Artificial Nural Network (ANN), Conventional Nural Network (CNN) Decision Tree (DT), Gradient Boosting (GB), Adaboost Classifier (ABC), Random Forest (RF), Support Vector Classifier (SVC), XGBoost Classifier (XGB), Multinomial Naive Bayes (MNB), Passive Aggressive (PA), Ridge Classifier (RC) and Logistic Regression (LR) algorithms. These classifiers were instrumental in our data analysis and algorithm evaluation process.

Artificial Nural Network (ANN)

Artificial Neural Networks (ANNs) are a fundamental component of machine learning, inspired by the structure and function of the human brain. Comprising interconnected nodes, or artificial neurons, organized into layers, ANNs process information through a series of mathematical operations. The layers typically include an input layer, one or more hidden layers, and an output layer. Each connection between neurons has an associated weight, representing the strength of the connection. During training, the network learns to adjust these weights by minimizing the difference between predicted outputs and actual targets. This is achieved through a process called backpropagation, where the error is propagated backward through the network, and the weights are updated accordingly using optimization algorithms like gradient descent. The strength of ANNs lies in their ability to capture complex patterns and relationships within data, making them versatile for various tasks such as image recognition, natural language processing, and regression analysis. Deep learning, a subfield of machine learning, extends the capabilities of ANNs by introducing deep neural networks with multiple hidden layers. This depth enables ANNs to automatically extract hierarchical features from data, empowering them to tackle intricate and high-dimensional problems with remarkable efficacy. The flexibility and adaptability of ANNs contribute to their widespread application and continual evolution in the field of artificial intelligence.

Conventional Nural Network (CNN)

The Convolutional Neural Network (CNN) is a powerful deep learning architecture specifically designed for processing structured grid data, such as images. At its core, a CNN comprises layers that systematically learn hierarchical representations of the input data. The fundamental building blocks include convolutional layers, pooling layers, and fully connected layers. In the initial layers, convolutional operations are performed to detect spatial patterns and features within the input data. Filters, also known as kernels, slide across the input, extracting local patterns through convolutional operations. Subsequent pooling layers reduce spatial dimensions, preserving essential features while enhancing computational efficiency. The extracted features are then flattened and fed into fully connected layers, enabling the network to understand complex relationships and make predictions. Activation functions, like Rectified Linear Units (ReLU), introduce non-linearity, enhancing the model's capacity to capture intricate patterns. CNNs are renowned for their ability to automatically learn hierarchical representations, making them highly effective in tasks like image classification, object detection, and feature extraction. Their inherent capability to recognize spatial hierarchies in data makes CNNs indispensable in various domains, ranging from computer vision to natural language processing. Despite being initially developed for image-related tasks, CNNs have demonstrated adaptability and success across a wide array of structured grid data applications.

Bidirectional Long- Short Term Memory (BLSTM)

The architecture of BLSTM consists of two LSTM layers, each processing the input sequence in a different direction—forward and backward. This bidirectional flow of information allows the model to maintain a comprehensive understanding of the temporal relationships within the data, contributing to its ability to grasp intricate patterns and dependencies. Consequently, BLSTM finds wide application in various fields, including natural language processing, speech recognition, and time series analysis, where capturing contextual information is paramount. The bidirectional nature of BLSTM facilitates a more holistic comprehension of sequential patterns, empowering the algorithm to excel in tasks that demand a nuanced understanding of the temporal dynamics inherent in sequential data. [23].

Long Short-Term Memory (LSTM)

Since its inception in 1995, the LSTM (Long Short-Term Memory) architecture within recurrent neural networks has undergone several refinements. LSTM models have proven to be effective and widely adopted for learning from sequential data, incorporating hidden components [24]. LSTM, an advanced version of the RNN (Recurrent Neural Network), enhances the system's ability to capture context when processing inputs and generating outputs [25]. Within LSTM blocks, various sub-components, including outputs, input gates, inputs from previous blocks, and activation functions, contribute to its complexity and functionality [26]. The name "Long Short-Term Memory" reflects its capability to extend short-term memory techniques into the realm of longer-term memory, making it a valuable tool for sequence-based learning.

Adaboost Classifier

The AdaBoost (Adaptive Boosting) classifier is a powerful ensemble learning method designed to enhance the performance of weak classifiers by combining them into a robust and accurate model. AdaBoost operates iteratively, sequentially adjusting the weight of each training instance based on the accuracy of the previous weak classifiers. This means that instances that are misclassified receive higher weights, allowing subsequent weak classifiers to focus on them and improve their classification accuracy. The final prediction is then made by combining the weighted outputs of these weak classifiers. One of AdaBoost's strengths lies in its adaptability to different classification problems, as it can work with a wide range of base classifiers, typically decision stumps or shallow decision trees. It's particularly effective in addressing complex datasets and overcoming issues such as overfitting, as it gives more emphasis to challenging data points during training. Moreover, AdaBoost is known for its ability to handle high-dimensional feature spaces effectively. While AdaBoost is a powerful algorithm, it's not immune to outliers or noisy data, which can adversely affect its performance. However, its capacity to mitigate these issues is strengthened by its sequential learning process. By leveraging AdaBoost's combination of weak learners, it often results in a strong and accurate classifier that is widely used in various fields, including face detection, text classification, and bioinformatics, where high performance and adaptability are essential [27]. The graph is shown in Figure 3.2.

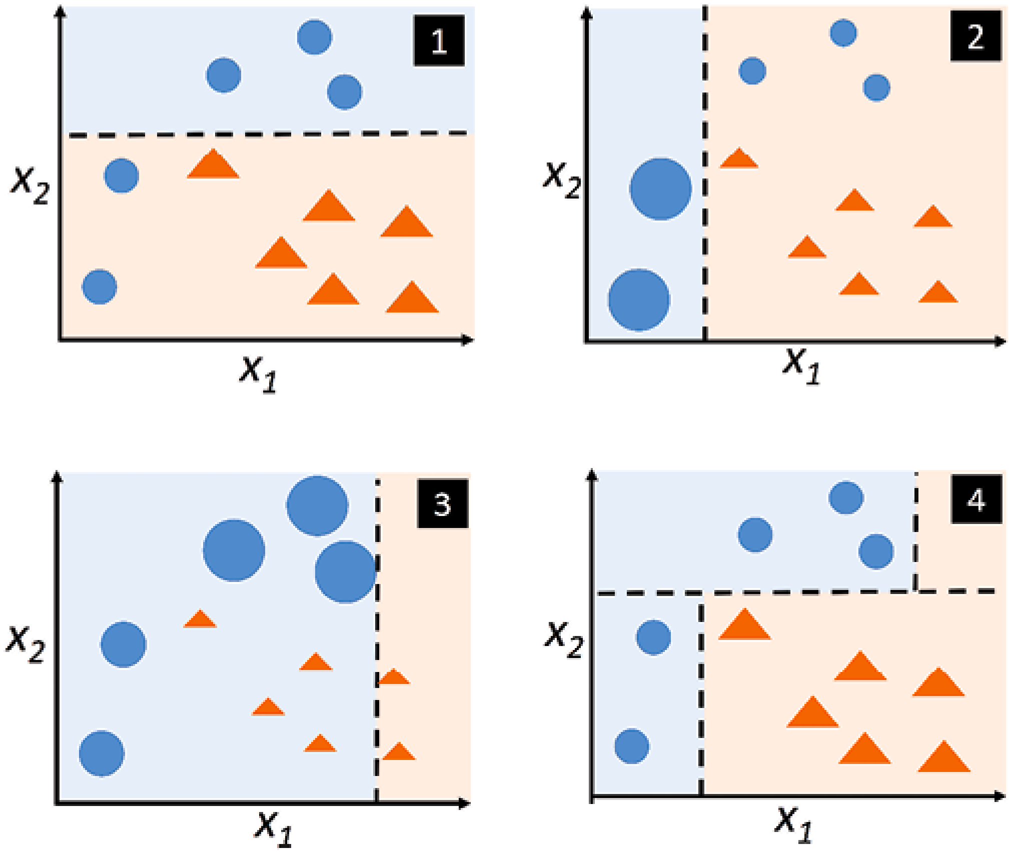


Figure 3.2: Adaboost Classifier

Decision Tree (DT)

Decision trees are a powerful method for categorizing occurrences by branching them out from a central node to leaf nodes that represent different categories. This methodology, known as the Decision Tree approach, is particularly effective for prediction tasks involving binary classes. Each inner node in a decision tree represents an attribute test, and the hierarchical structure of the tree culminates in leaf nodes that indicate separate classes. The DT (Decision Tree) structure is commonly used based on decision trees, and it can be applied to both classification and regression problems. As the tree grows from the root node, the "splitting" process is employed to select the "Best Feature" or "Best Attribute" from the available pool of potential characteristics. Two metrics, namely "Entropy" and "Information Gain," are often used to determine the best attribute. Entropy (formula indicated as (1)) is a measure of dataset homogeneity, indicating the impurity or disorder within a set of instances. On the other hand, Information Gain (formula indicated as (2)) quantifies the rate of change in entropy when splitting the dataset based on different attributes. The attribute that provides the most valuable information is considered the "best attribute." In summary, decision trees offer a hierarchical structure to classify instances, with attribute tests represented by inner nodes. The selection of the "best attribute" is determined by metrics like entropy and information gain, which help measure the dataset's homogeneity and the usefulness of attributes in reducing uncertainty shown in Figure 3.3.

$$E(D) = -P\ (positive)log2P\ (positive) - P(negative)\log 2P\ (negative) \ldots\ldots (2)$$

$$Gain\ (Attribute\ X) = Entropy\ (decision\ Attribute\ Y) - Entropy\ (X, Y) \ldots (3)$$

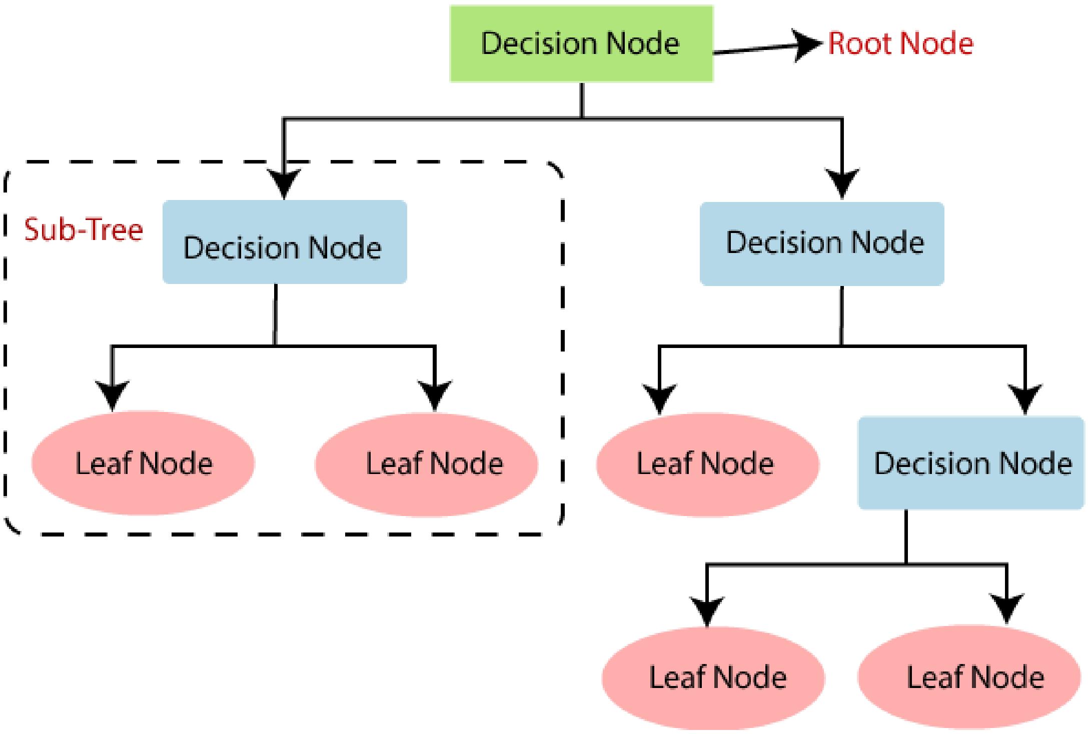


Figure 3.3: Decision Tree Classifier

Logistic Regression (LR)

Logistic Regression is a widely utilized and interpretable machine learning classifier that excels in binary and multiclass classification tasks. Unlike linear regression, which predicts continuous values, logistic regression models the probability of an instance belonging to a particular class using the logistic function (sigmoid). It estimates the odds of an event occurring and maps them to a range between 0 and 1, allowing it to provide clear class separation. The model is trained by minimizing the logistic loss or cross-entropy loss through iterative optimization techniques like gradient descent. Logistic Regression is advantageous for its simplicity, quick training, and ease of interpretation. It can handle both linear and non-linear relationships between features and the target variable through polynomial or interaction terms. While primarily a binary classifier, it can be extended to multiclass problems through techniques like one-vs-rest or softmax regression [28]. One limitation is its susceptibility to overfitting when dealing with high-dimensional data or complex relationships, which can be mitigated through regularization techniques like L1 (Lasso) or L2 (Ridge) regularization. Despite its simplicity, logistic regression is a valuable tool in various domains, including healthcare (predicting disease outcomes), finance (credit risk assessment), and natural language processing (text classification), and it serves as a foundational model in many machine learning pipelines due to its transparency and effectiveness shown in Figure 3.4.

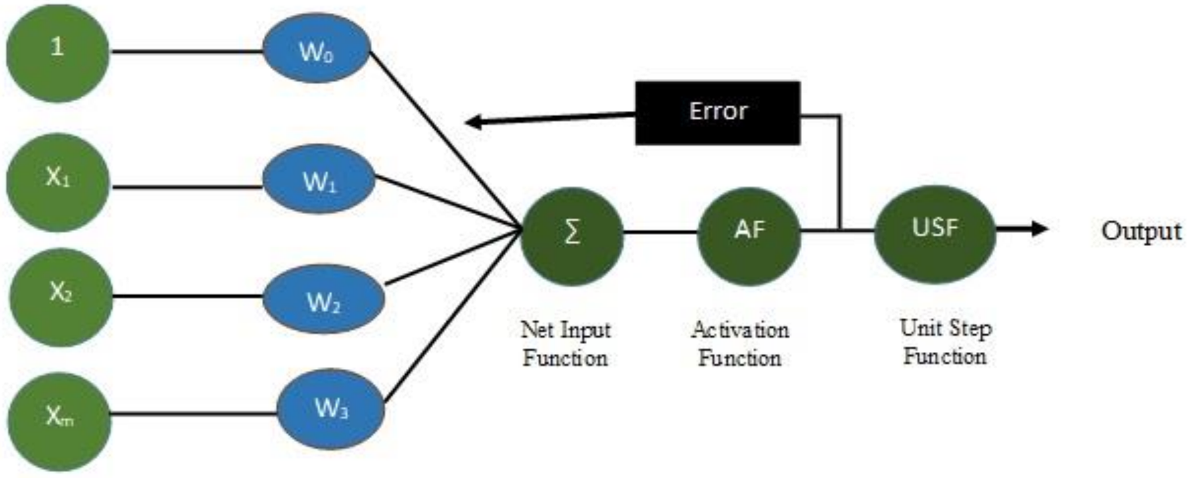


Figure 3.4: Logistic Regression

Random Forest (RF)

The Random Forest classifier is a powerful and versatile machine learning algorithm that has gained immense popularity for both classification and regression tasks. It operates by creating an ensemble of decision trees, where each tree is constructed using a random subset of the training data and a subset of the available features. This technique introduces variability and decorrelates the individual trees, mitigating overfitting and improving the model's generalization performance. In classification, the Random Forest combines the results from these decision trees through a majority vote, while in regression, it computes the average of the individual tree predictions. One of the key advantages of Random Forest lies in its ability to handle high-dimensional data, maintain robustness against outliers, and provide feature importance for model interpretability. The algorithm is less prone to overfitting compared to single decision trees, thanks to its inherent bagging (Bootstrap Aggregating) and feature bagging components. Random Forest is particularly useful when dealing with complex and noisy datasets, and it's less sensitive to hyperparameter tuning than other algorithms. Additionally, the Random Forest can identify influential features and provide insights into their contribution to the model's predictive power. Its robust performance, scalability, and flexibility have made it a popular choice across various domains, including finance, healthcare, and image analysis. However, the trade-off for its power and versatility is increased computational cost and complexity, which can be a consideration for real-time or resource-constrained applications. Nonetheless, the Random Forest remains a reliable workhorse in machine learning, delivering accurate predictions and valuable insights for diverse problem-solving scenarios [29]. The methods shown in Figure 3.5.

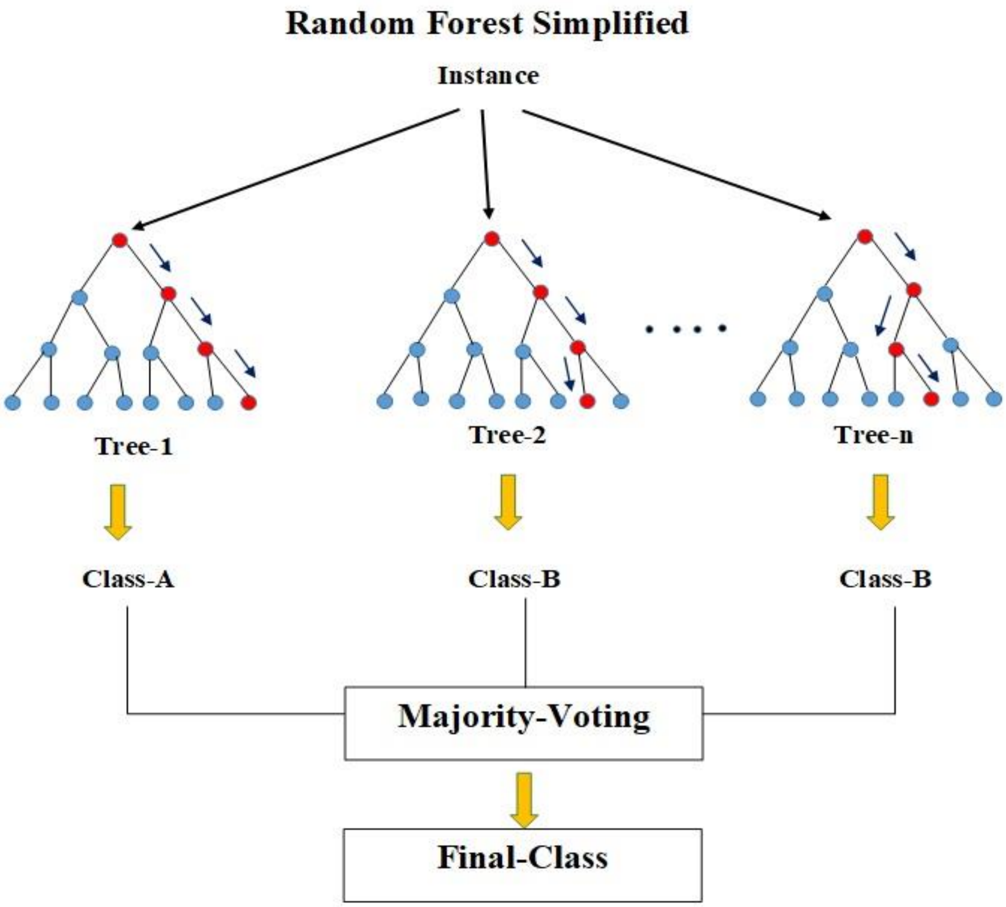


Figure 3.5: Random Forest

Gradient Boosting (GB)

The Classifier is a powerful and versatile ML algorithm that excels in predictive modeling, particularly in classification tasks. It operates by iteratively building a strong predictive model through the combination of multiple weak models, typically decision trees, in a sequential manner. At each iteration, the algorithm focuses on the misclassified data points from the previous stage, assigning them greater importance. This iterative process allows the algorithm to continuously refine its predictions, ultimately creating a robust ensemble model. One of the key advantages of the Gradient Boosting Classifier is its ability to handle complex, high-dimensional data and capture intricate relationships between variables. By combining the outputs of multiple weak learners, it can achieve superior predictive performance. However, this power comes at a computational cost, and training a Gradient

Boosting model can be more time-consuming compared to some other algorithms. To mitigate the risk of overfitting, careful hyperparameter tuning and cross-validation are essential when implementing Gradient Boosting. The choice of the learning rate, the number of boosting iterations (trees), and the maximum depth of trees are critical factors that influence the model's performance. In practice, Gradient Boosting is widely used in various fields, including data mining, finance, and biology, due to its effectiveness in addressing complex classification challenges and producing accurate results. Its versatility and robustness make it a valuable tool for both beginners and experienced data scientists aiming to tackle a wide range of classification tasks shown in Figure 3.6.

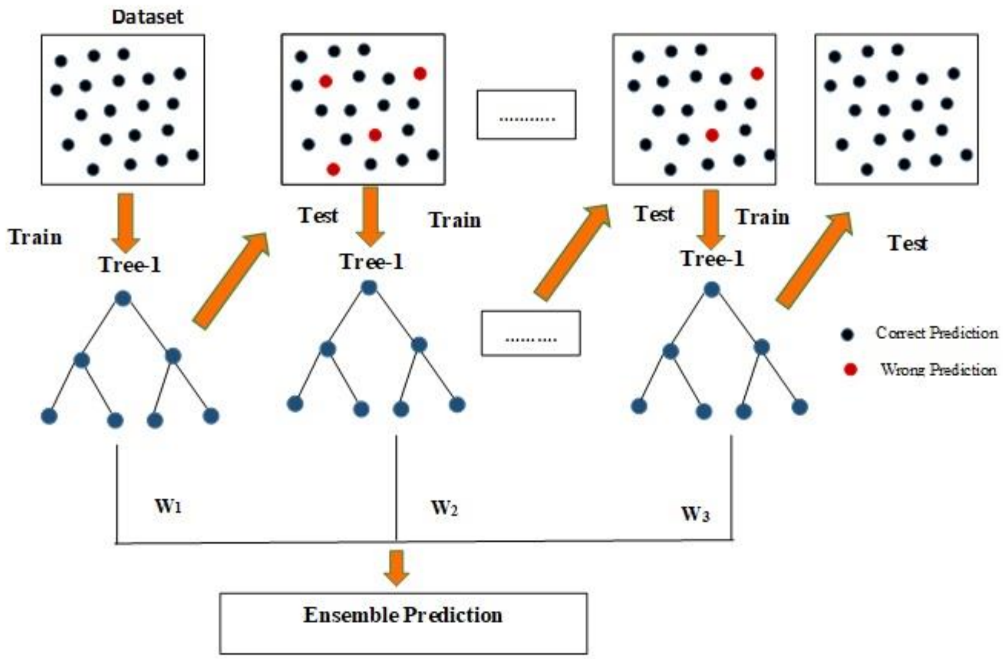


Figure 3.6: Gradient Boosting model

XGB

The XGB Classifier is a gradient-boosted decision tree. Decision trees are constructed one after the other using this procedure. Weights are critical when utilizing the XGB classifier. The decision tree, which predicts based on several criteria, is fed information about the weights assigned to the independent variables. Factors that the tree did not examine at first are given more weight and utilized to train a second decision tree. By integrating numerous separate classifiers, a more robust and accurate model is created. It is capable of doing regression, classification, ranking, and bespoke prediction tasks. The concept is shown in below Figure 3.7.

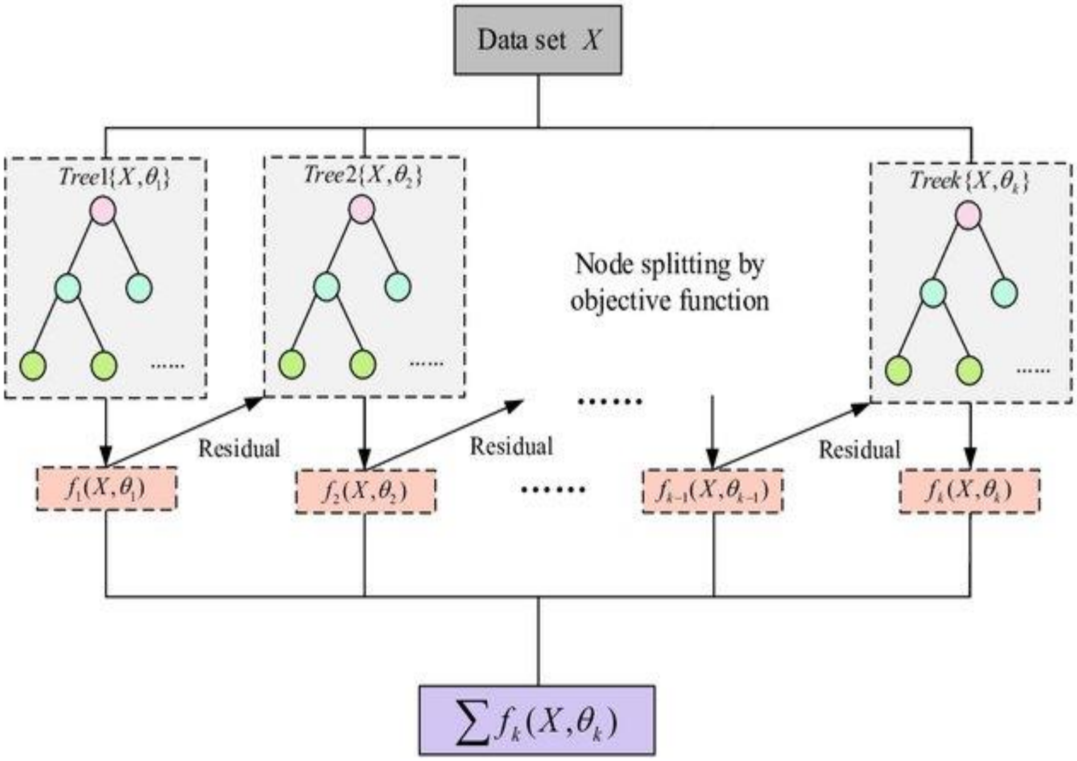


Figure 3.7: XGB model

Support Vector

Regression and classification problems may both be resolved using the Support Vector Classifier (SVC). However, categorization issues are where machine learning is most frequently applied. The SVM approach looks for a straight line, or judgment boundary, that divides the area into categories in all n variables in order to successfully classify fresh data points. A hyperplane is this highest utility bound. Using SVM, which chooses the most extreme points and vectors, the hyperplane may be created. As a result, the word "support vector," which is used to describe these severe situations, is where the technique's name, "support vector machine," comes from. Figure 3.8 shows the procedure.

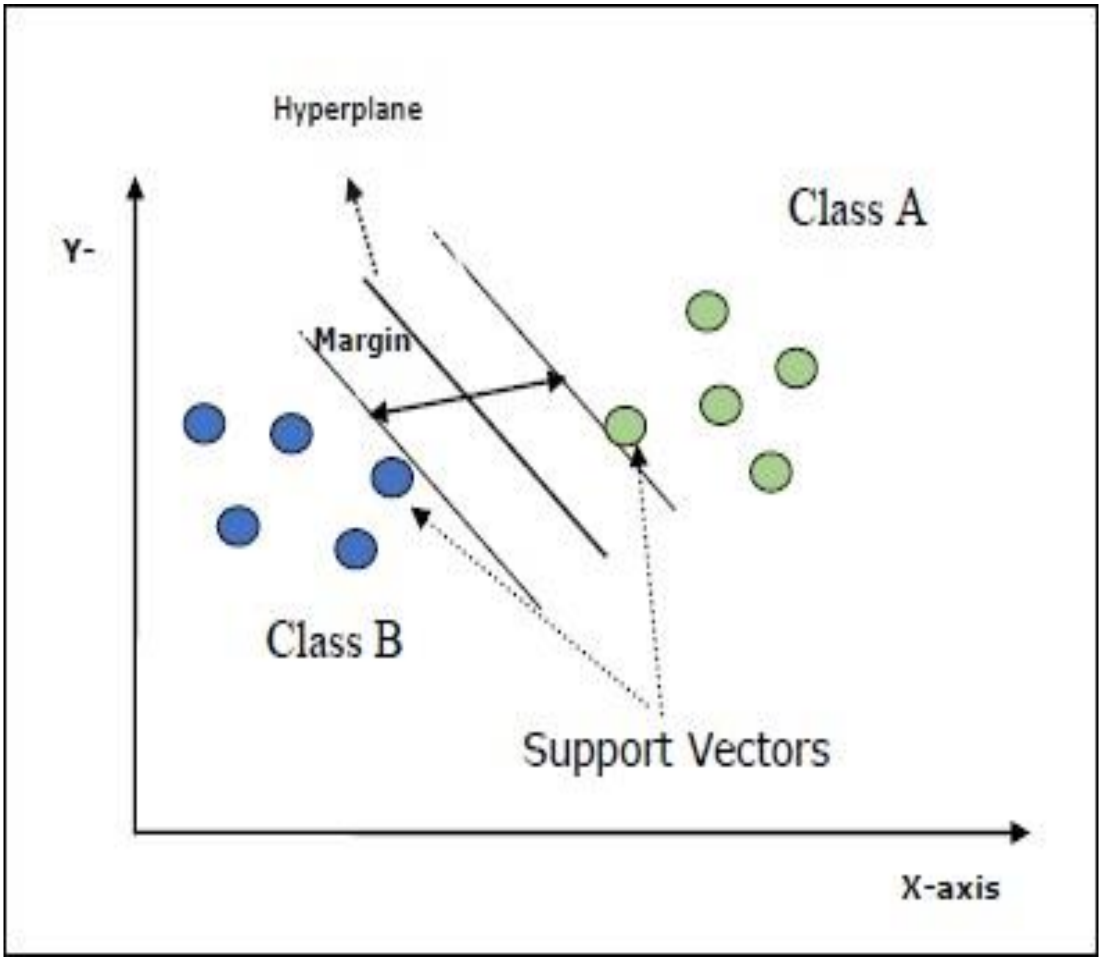


Figure 3.8: SVC Classifier

Multinomial Naive Bayes (MNB)

Multinomial Naive Bayes (MNB) is a probabilistic classification algorithm designed for text and document-based data. It is particularly effective for tasks involving discrete features, such as word frequencies in natural language processing. MNB models the likelihood of observing a set of words in a document given its class, assuming independence between features. This algorithm is widely used for tasks like text categorization and spam filtering, where the focus is on counting occurrences of words and making predictions based on the frequency distribution. MNB's simplicity, efficiency, and ability to handle large datasets make it a popular choice for various text-based applications.

Passive Aggressive (PA)

The Passive-Aggressive (PA) algorithm is a machine learning classification algorithm known for its efficiency and simplicity. It belongs to the family of online learning algorithms and is particularly suited for large-scale, real-time scenarios. The "passive" aspect refers to its conservative approach during correct predictions, where the model does not undergo significant updates. However, when faced with misclassifications, the algorithm becomes "aggressive," adapting swiftly to the new information. This characteristic makes PA well-suited for scenarios where data distribution may change over time. Common applications include text classification, financial fraud detection, and spam filtering. Due to its adaptability and low computational complexity, Passive-Aggressive is valuable in situations requiring swift adjustments to evolving data patterns.

Ridge Classifier (RC)

The Ridge Classifier (RC) is a linear classification algorithm that operates on the principles of ridge regression. It is commonly used in machine learning tasks where the relationship between the features and the target variable is linear. The Ridge Classifier addresses the issue of multicollinearity in the input features by introducing a regularization term, known as the ridge or L2 penalty, into the linear regression equation. In the context of classification, the Ridge Classifier aims to find the optimal hyperplane that separates different classes in the feature space. The regularization term helps prevent overfitting by penalizing large coefficients. This regularization is particularly useful when dealing with high-dimensional data. The Ridge Classifier is employed in scenarios where a balance between model simplicity and predictive accuracy is crucial. Its effectiveness is notable in tasks such as text classification, sentiment analysis, and medical diagnosis, where linear relationships are prevalent and feature interpretability is valuable.

Proposed Methodology Flow chart:

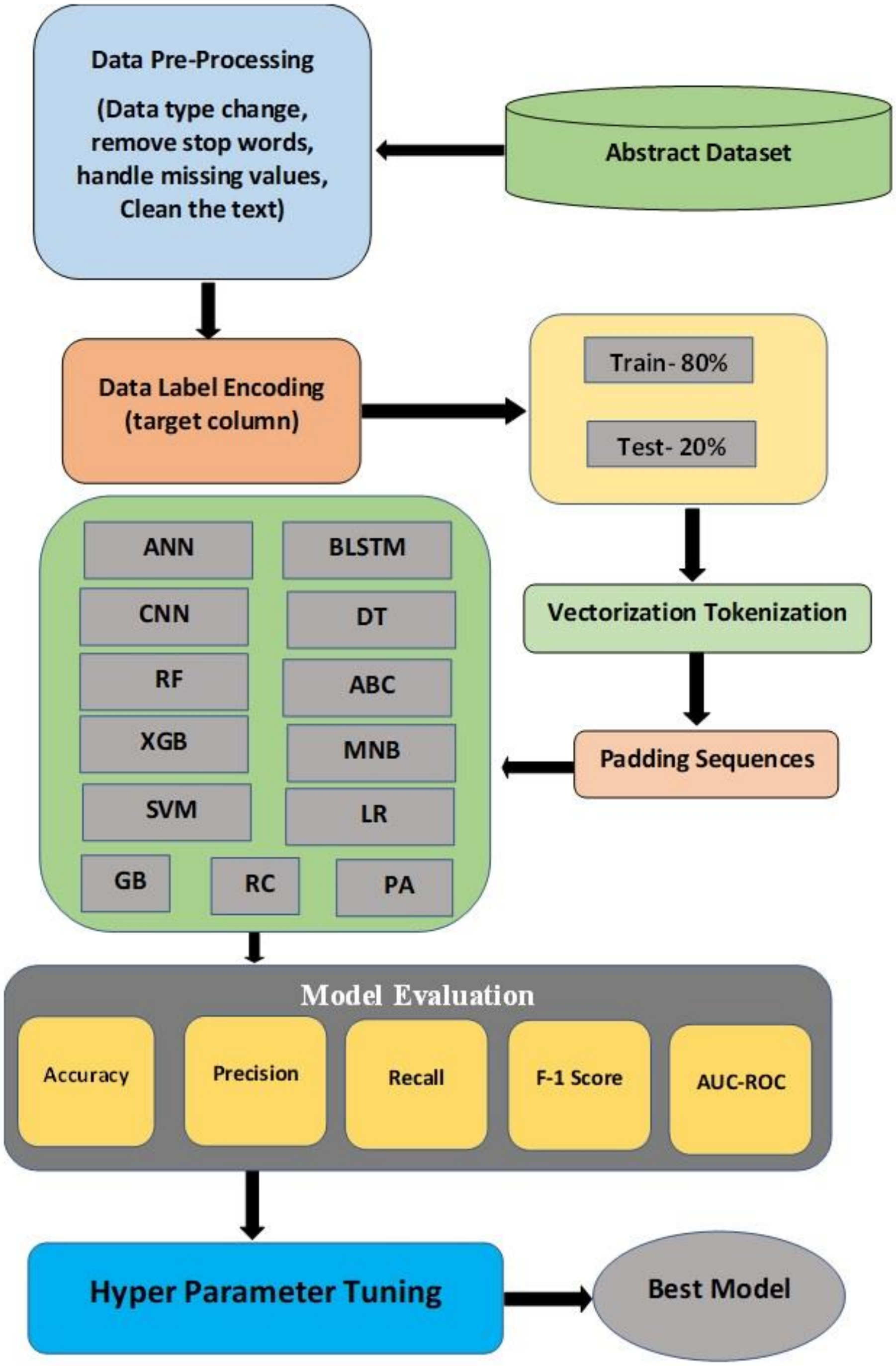


Figure 3.9: Methodology of Abstract Classification

The methodology for detecting abstract from scientific research dataset involves several steps. Firstly, a dataset is collected. The dataset is preprocessed by removing irrelevant columns, handling missing values, and cleaning the text. The target column is then mapped to numerical values for model training. There are sets for testing and training created from the dataset. Transforming categorical data to numerical form is a vital step in machine learning. We opted for an 80-20 split, utilizing 80% of the resources for training and allocating the remaining 20% for testing. The text is tokenized and padded to prepare it for model input. Various models such as CNN, ANN, BLSTM, DT, GB, ABC, RF, SVC, XGB, MNB, PA, RC and LR are trained on the data. The models are evaluated using metrics like accuracy, precision, recall, F1 score, and AUC-ROC. For models with probability estimates, the ROC curve is analyzed. Finally, the best-performing model is selected based on the evaluation results [13]. This methodology provides a comprehensive approach to detect abstract in the dataset and can guide researchers in developing effective abstract detection systems shows in Figure 3.9.

3.5 Implementation Requirements

To effectively train and assess our proposed model, we began by sourcing the required datasets. Ensuring the dataset's cleanliness was of utmost importance, achieved through a series of filtering methods. Subsequently, data preprocessing techniques like the Scaler Transform were applied, translating categorical data into numerical values. An 80-20 split was employed for training and testing, allowing us to thoroughly evaluate our chosen algorithms. We then put our selected algorithms into practice and conducted a comprehensive evaluation of the results. Employing ensemble techniques, we sought to maximize our detection accuracy. The outcomes of these ensemble algorithms were rigorously assessed, and the results were further validated through hyperparameter tuning. This process involved a meticulous examination of the mathematical frameworks utilized. Following these steps, the data analysis phase was initiated, culminating in the implementation of model learning and a tailored detection strategy. The best model was selected based on key metrics, including accuracy, precision, recall, and the F-1 score. Additionally, we measured the AUC-ROC score and implemented the corresponding curve for a holistic evaluation of our model's performance. This structured approach ensured a thorough and well-informed model selection and validation process.

Chapter 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

In this study, a supervised learning approach was adopted, which relies on the training and testing phases. The deep learning model was constructed using the training dataset, and the model's performance was assessed by applying it to the testing dataset. The subsequent sections will provide a concise and swift illustration of the deep learning algorithm employed in this research.

4.2 Experimental Results & Analysis

At this juncture, our focus turned to the evaluation of existing models and assessing the efficiency of our proposed model. To gauge the overall performance, we employed various performance evaluation methods on unseen data. This section entails presenting an analytical report based on our experimental findings pertaining to deep learning models designed for the specific context of abstract datasets. Our methodology began with the initial implementation of our selected dataset. We meticulously handled missing or incorrect values, ensuring dataset integrity. Next, we introduced a range of diverse algorithms and conducted a thorough analysis of their performance. To evaluate the effectiveness of these algorithms, we employed key metrics, including Confusion Matrices, Accuracy, Precision, Recall, and the F-1 Score. These metrics allowed us to comprehensively assess both our proposed models and traditional algorithms. In our analysis, we employed a variety of algorithms, including BLSTM, ANN, CNN, DT, GB, ABC, RF, SVC, XGB, MNB, PA, RC and LR algorithms, specifically tailored to our dataset. To further enhance our evaluation, we explored different ensemble techniques, leveraging Confusion Matrices for both datasets. This comprehensive approach aimed to provide a holistic understanding of the performance and effectiveness of the various models under consideration.

Accuracy

This section explores accuracy, denoting the proportion of accurate predictions made with testing data. It measures the percentage of predictions that precisely match the actual outcomes. Accuracy is a measure of the model's correctness, comparing its predictions to the actual real-world measurements. It focuses on a single variable and primarily addresses intentional errors, making it one of the most straightforward and widely used evaluation techniques for any model. Ensuring the accuracy of our models is a crucial aspect of model validation and performance assessment.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

Precision

This section addresses precision, which measures the proportion of positively predicted observations that actually occurred. Precision reflects the true positive rate, highlighting the actual percentage of instances when the model correctly predicted true positive outcomes. It's important to note that while a strong recall is desirable for many models, it can sometimes be misleading if not considered in the context of precision and other performance metrics.

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

Recall

This section discusses recall, which is the proportion of actual positive data points correctly predicted by the model. Recall is crucial in determining the model's ability to capture true positive instances, and it establishes the ratio of all positive labels to the predicted positives. While high accuracy is generally desirable, it's essential to recognize that it can sometimes be misleading if not assessed alongside other important metrics like recall.

$$Recall = \frac{TruePositive}{TruePositive + \ FalseNegative}$$

F-1 Score

This section discusses the evaluation metrics of accuracy and recall, emphasizing their relevance in assessing a model's performance. Key metrics to consider are the recall and accuracy ratios, which provide insights into the model's ability to correctly identify relevant instances and overall accuracy. It's important to note that if the mean of the harmonic mean of these metrics is relatively low, it may indicate that the model's performance is not optimal, warranting further improvements.

$$F - 1\ Score = 2 * \frac{Recall *\ Precision}{Recall + Precision}$$

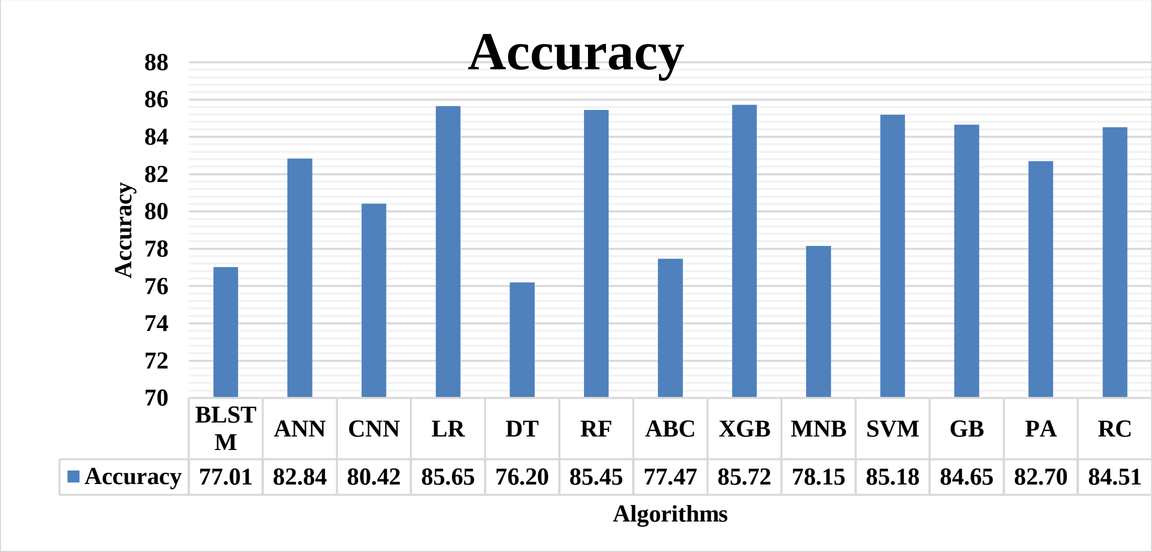
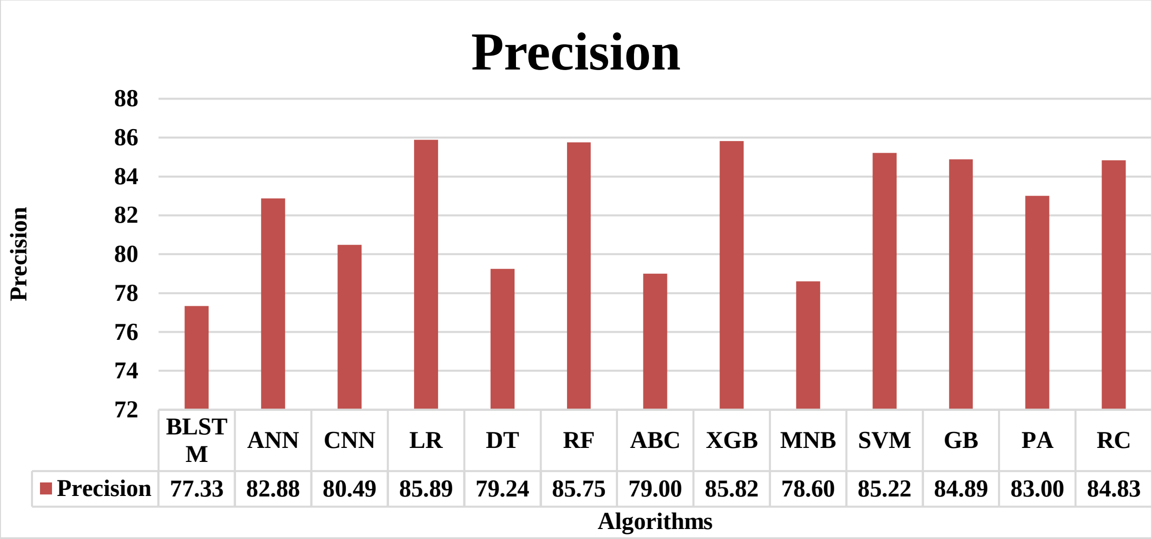


Figure 4.1: Experimental Results of Accuracy

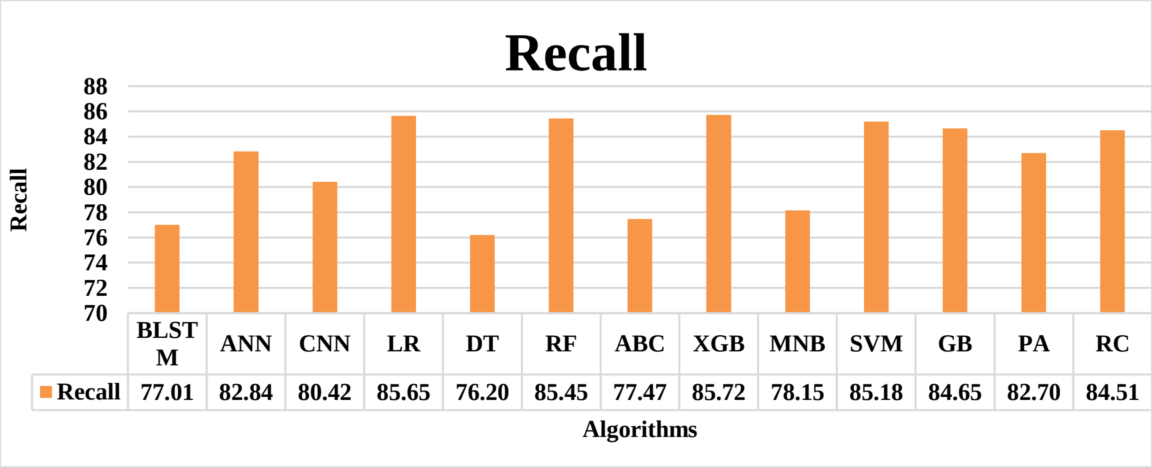
First, we looked at algorithmic accuracy performance. Using XGB, the best accuracy was 85.72%. SVM got 85.18%, BLSTM got 77.01%, DT got 76.20%, ANN got 82.84%, ABC got 77.47%, RF got 85.45%, GB got 84.65%, MNB got 78.15%, CNN got 80.42%, PA got 82.70%, RC got 84.51% and LR got 85.65%. The visualization is shown in Figure 4.1.



	BLSTM	ANN	CNN	LR	DT	RF	ABC	XGB	MNB	SVM	GB	PA	RC
■ Precision	77.33	82.88	80.49	85.89	79.24	85.75	79.00	85.82	78.60	85.22	84.89	83.00	84.83

Figure 4.2: Experimental Results of Precision

Second, we looked at the algorithmic precision performances. Using LR, the best precision was 85.89%. GB got 84.89%, BLSTM got 77.33%, DT got 79.24%, CNN got 80.49%, ABC got 79%, RF got 85.75%, XGB got 85.82%, MNB got 78.60%, ANN got 82.88%, PA got 83%, RC got 84.83% and SVM got 85.22%. The visualization is shown in Figure 4.2.



	BLSTM	ANN	CNN	LR	DT	RF	ABC	XGB	MNB	SVM	GB	PA	RC
■ Recall	77.01	82.84	80.42	85.65	76.20	85.45	77.47	85.72	78.15	85.18	84.65	82.70	84.51

Figure 4.3: Experimental Results of Recall

We looked at algorithmic recall performance, and using XGB, the best recall was 85.72%. SVM got 85.18%, BLSTM got 77.01%, DT got 76.20%, ANN got 82.84%, ABC got

77.47%, RF got 85.45%, GB got 84.65%, MNB got 78.15%, CNN got 80.42%, PA got 82.70%, RC got 84.51% and LR got 85.65%. The visualization is shown in Figure 4.3.

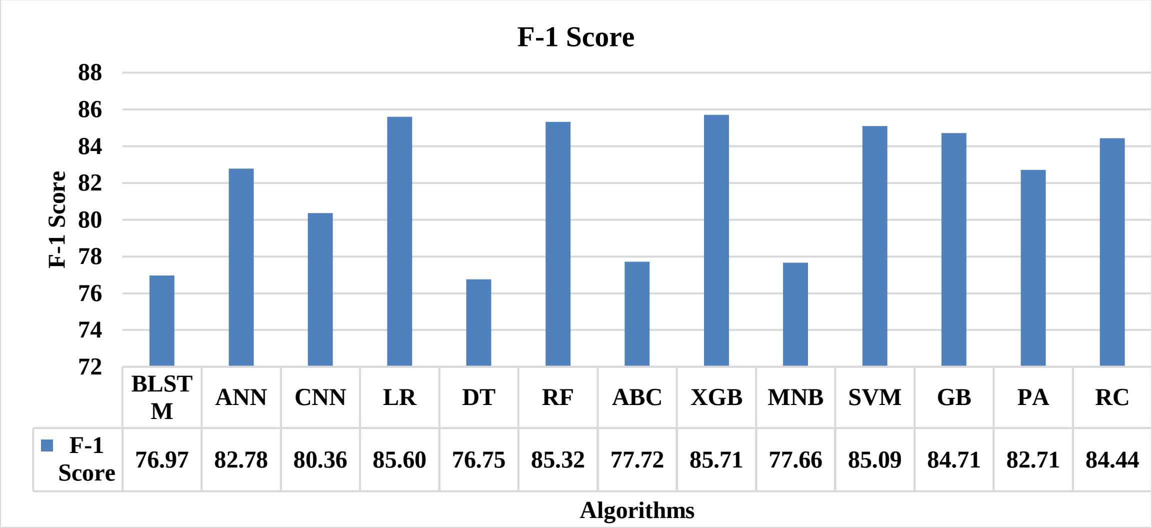


Figure 4.4: Experimental Results of F-1 Score

F-1 Score, the best F-1 Score had obtained at 85.71% using XGB. SVM got 85.09%, BLSTM got 76.97%, DT got 76.75%, ANN got 82.78%, ABC got 77.72%, RF got 85.32%, GB got 84.71%, MNB got 77.66%, CNN got 80.36%, PA got 82.71%, RC got 84.44% and LR got 85.60%. The visualization is shown in Figure 4.4.

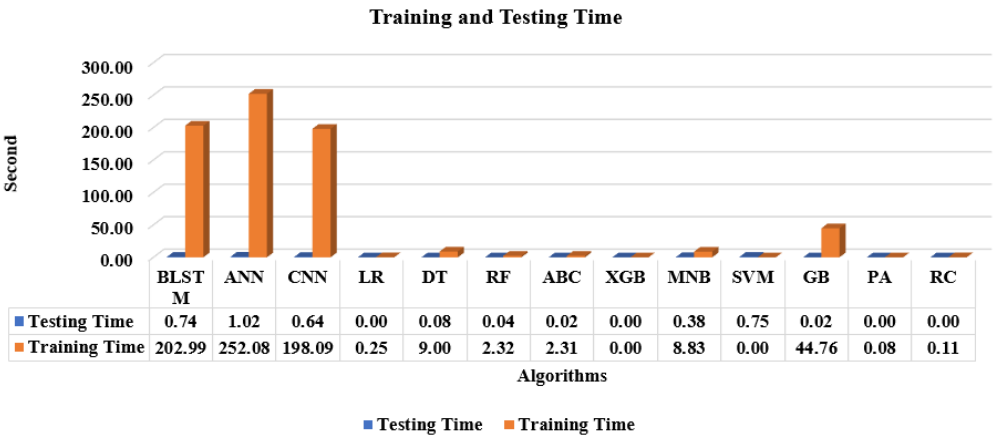


Figure 4.5: Experimental Results of Training and Testing time

We considered the performances of algorithmic training and testing time. The highest time was assumed for training is 0.74 second for BLSTM, ANN took 1.02 second and CNN took 0.64 second. The visualization is shown in Figure 4.5.

Table 4.1: Details of the algorithms results

Algorithm	Accuracy	Precision	Recall	F-1 Score	Testing Time	Training Time
BLSTM	**77.01**	**77.33**	**77.01**	**76.97**	**0.74**	**202.99**
ANN	**82.84**	**82.88**	**82.84**	**82.78**	**1.02**	**252.08**
CNN	**80.42**	**80.49**	**80.42**	**80.36**	**0.64**	**198.09**
LR	**85.65**	**85.89**	**85.65**	**85.60**	**0.00**	**0.25**
DT	**76.20**	**79.24**	**76.20**	**76.75**	**0.08**	**9.00**
RF	**85.45**	**85.75**	**85.45**	**85.32**	**0.04**	**2.32**
ABC	**77.47**	**79.00**	**77.47**	**77.72**	**0.02**	**2.31**
XGB	**85.72**	**85.82**	**85.72**	**85.71**	**0.00**	**0.00**
MNB	**78.15**	**78.60**	**78.15**	**77.66**	**0.38**	**8.83**
SVM	**85.18**	**85.22**	**85.18**	**85.09**	**0.75**	**0.00**
GB	**84.65**	**84.89**	**84.65**	**84.71**	**0.02**	**44.76**
PA	**82.70**	**83.00**	**82.70**	**82.71**	**0.00**	**0.08**
RC	**84.51**	**84.83**	**84.51**	**84.44**	**0.00**	**0.11**

Table 4.1 shows the overall evaluation result each algorithms. Here we also represent the training and testing time for each algorithms.

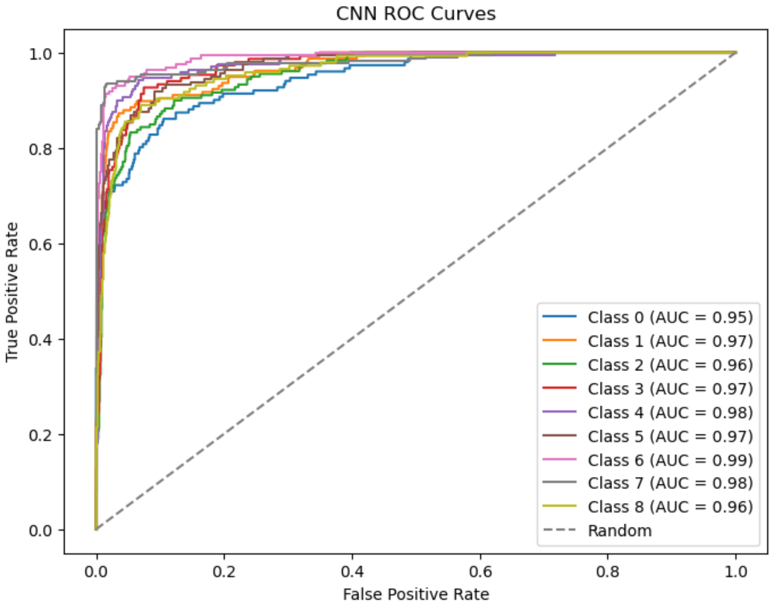


Figure 4.6: AUC-ROC Curve Analysis for CNN

Figure 4.6, CNN had multiple classes in target column. Here the graph was drawn for each class.

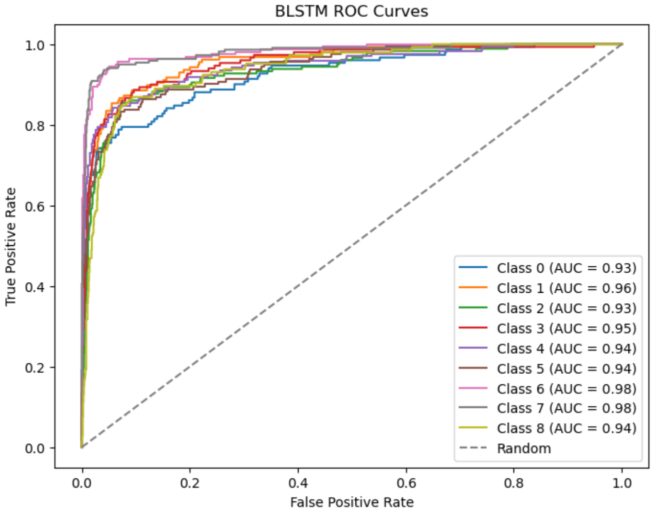


Figure 4.7: AUC-ROC Curve Analysis for BLSTM

Figure 4.7, BLSTM had multiple classes in target column. Here the graph was drawn for each class.

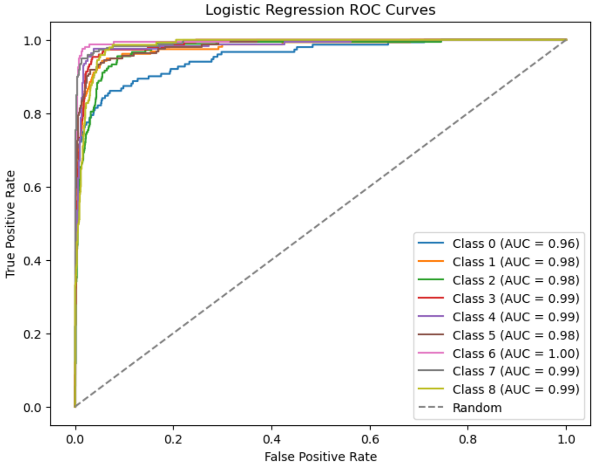


Figure 4.8: AUC-ROC Curve Analysis of for LR

Figure 4.8, LR had multiple classes in target column. Here the graph was drawn for each class.

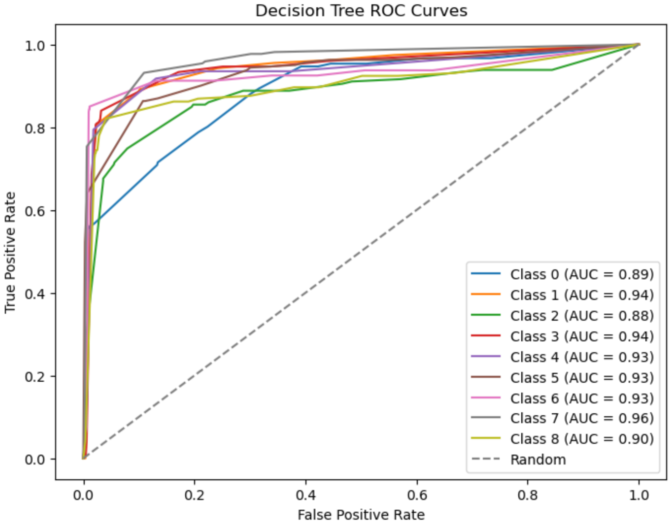


Figure 4.9: AUC-ROC Curve Analysis for DT

Figure 4.9, DT had multiple classes in target column. Here the graph was drawn for each class.

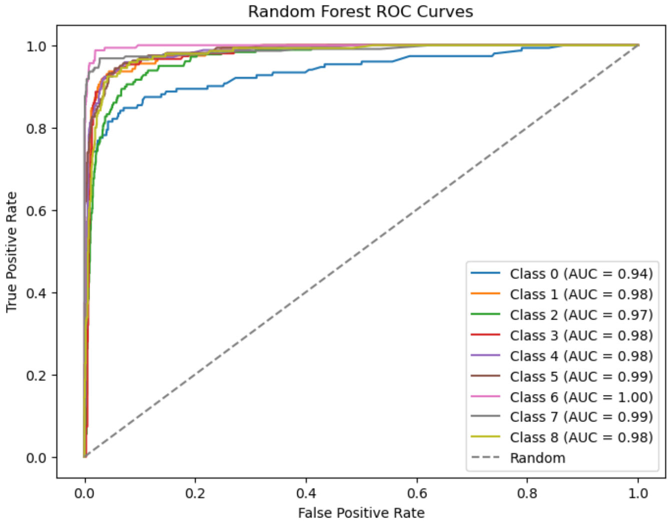


Figure 4.10: AUC-ROC Curve Analysis for RF

Figure 4.10, RF had multiple classes in target column. Here the graph was drawn for each class.

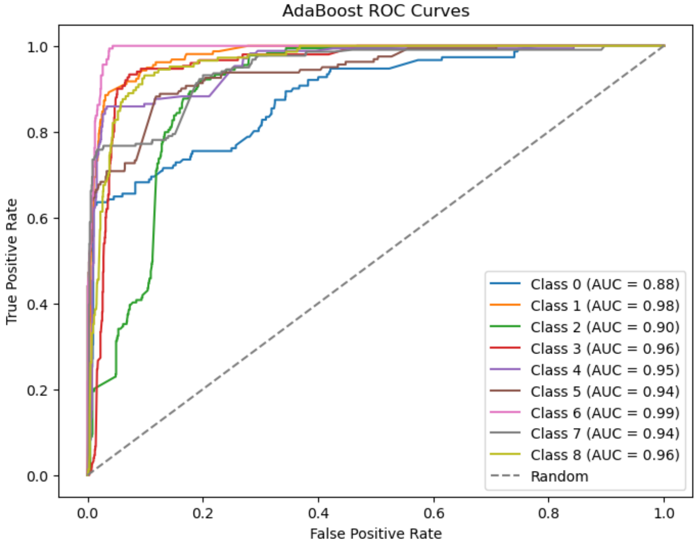


Figure 4.11: AUC-ROC Curve Analysis for ABC

Figure 4.11, ABC had multiple classes in target column. Here the graph was drawn for each class.

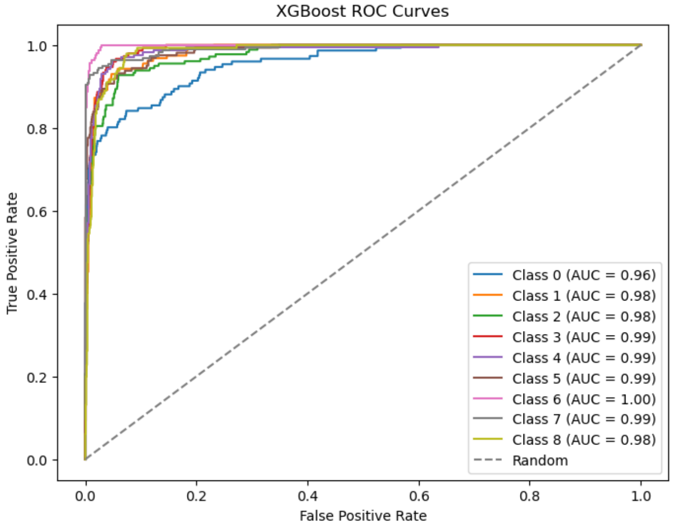


Figure 4.12: AUC-ROC Curve Analysis for XGB

Figure 4.12, XGB had multiple classes in target column. Here the graph was drawn for each class.

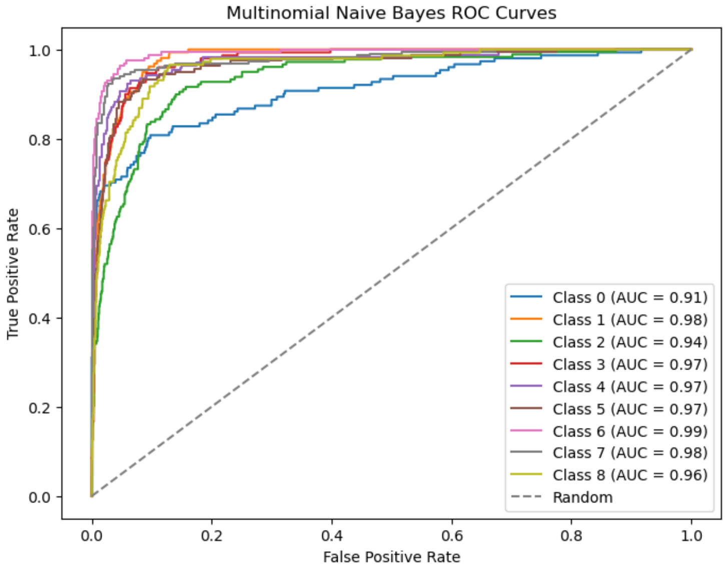


Figure 4.13: AUC-ROC Curve Analysis for MNB

Figure 4.13, MNB had multiple classes in target column. Here the graph was drawn for each class.

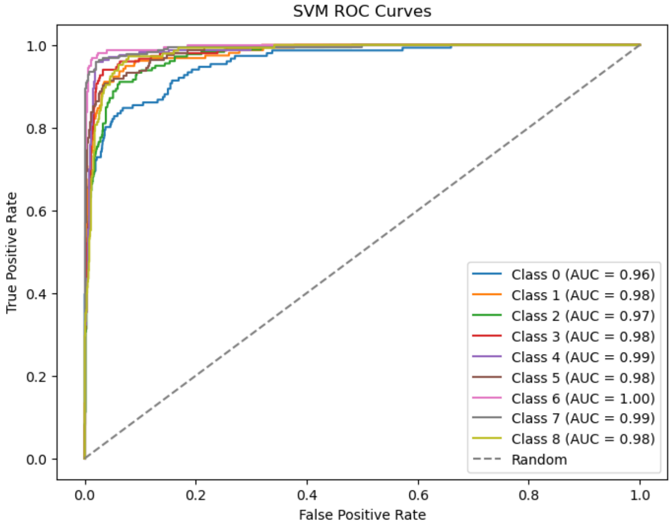


Figure 4.14: AUC-ROC Curve Analysis for SVM

Figure 4.14, SVM had multiple classes in target column. Here the graph was drawn for each class.

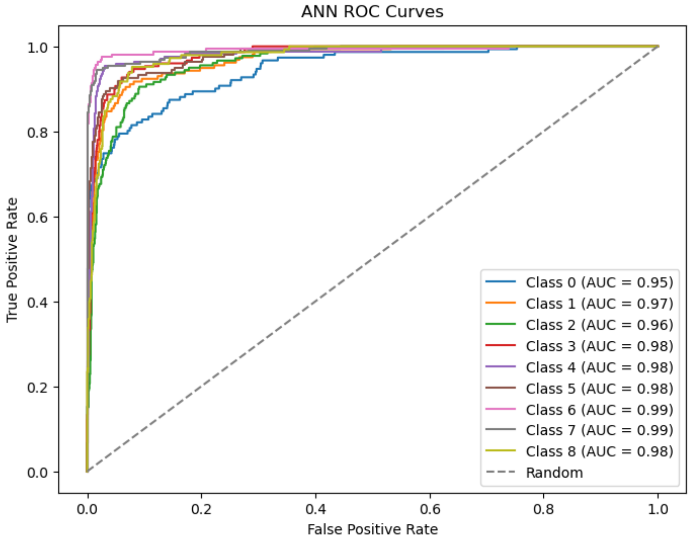


Figure 4.15: AUC-ROC Curve Analysis for ANN

Figure 4.15, ANN had multiple classes in target column. Here the graph was drawn for each class.

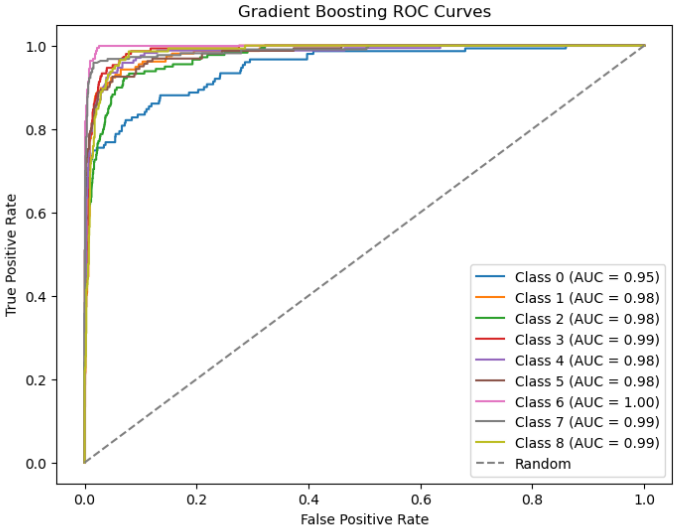


Figure 4.16: AUC-ROC Curve Analysis for GB

Figure 4.16, GB had multiple classes in target column. Here the graph was drawn for each class

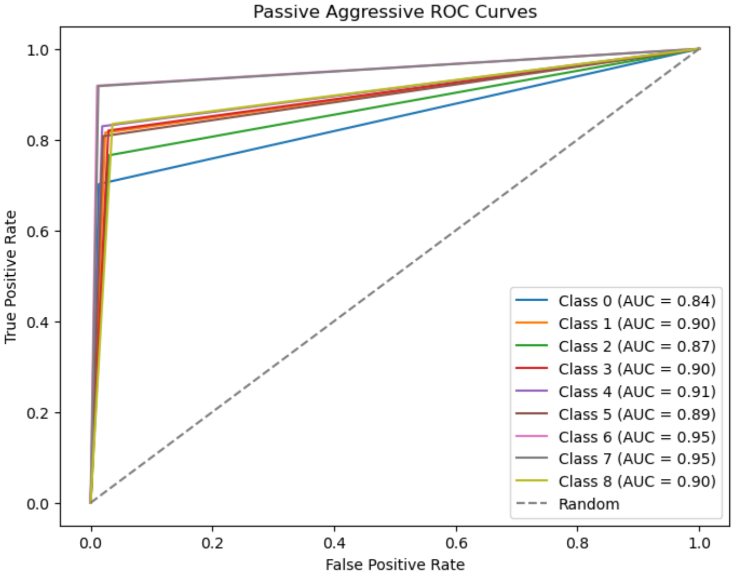


Figure 4.17: AUC-ROC Curve Analysis for PA

Figure 4.17, PA had multiple classes in target column. Here the graph was drawn for each class.

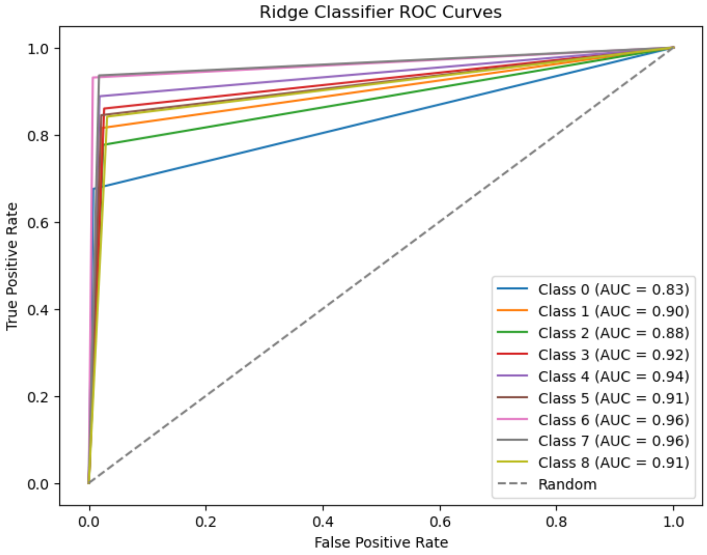


Figure 4.18: AUC-ROC Curve Analysis for RC

Figure 4.18, RC had multiple classes in target column. Here the graph was drawn for each class.

4.3 Discussion

In this stage, we will elucidate the evaluative framework for our proposed model. Our assessment criteria encompassed [22]. Within this study, we harnessed four distinct types of algorithms, which included BLSTM, CNN, ANN, DT, GB, ABC, RF, SVC, XGB, MNB, PA, RC and LR. Among these algorithms, the GB model demonstrated exceptional performance, achieving an accuracy rate of 83.82%. The SVC also delivered impressive results, with an accuracy rate of 83.50%. These high-performance outcomes were achieved through the meticulous application of hyperparameter tuning, further enhancing the robustness of our model.

Chapter 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

The impact of research abstract classification in the English language is profound, encompassing various aspects. This process has detrimental effects on the systematic organization of academic knowledge, leading to challenges in information retrieval, comprehension, and scholarly advancement. Academic performance and educational attainment are compromised, as researchers may experience difficulty navigating through diverse abstracts, potentially hindering their ability to extract relevant information. Scholarly relationships suffer, as individuals may find it challenging to connect with the most pertinent research, resulting in feelings of academic isolation and limited collaboration. Moreover, the reputation and visibility of scholarly works can be affected, impacting their recognition within the academic and professional realms. The pervasive challenge of abstract classification in the English language necessitates proactive measures to enhance the organization, accessibility, and understanding of research abstracts, thereby fostering a more efficient scholarly environment.

5.2 Impact on Environment

It is essential to specify that the ramifications of abstract classification predominantly manifest in the digital realm rather than the physical environment. Nevertheless, the impact of abstract classification on the online environment is significant. It establishes an atmosphere that influences how abstracts are perceived, evaluated, and categorized. This process can shape the digital discourse and the way scholarly information is accessed and utilized. The effective implementation of abstract classification strategies is crucial for promoting a positive online scholarly community, facilitating the efficient organization and retrieval of information, and enhancing the overall accessibility and impact of academic knowledge in the digital space.

5.3 Ethical Aspects

Ethical considerations surrounding abstract classification in the English language are crucial. It is essential to individuals involved in the research process, safeguarding their personal information and sensitive data. Ethical guidelines must be followed when collecting and analyzing academic abstracts, respecting the rights and consent of the authors whose data is being utilized. Additionally, fairness and impartiality should be upheld to avoid potential biases or stigmatization of certain research areas. Transparency in the use of algorithms and models is necessary, allowing for scrutiny and accountability. Ethical awareness and responsible practices are vital to ensure the well-being and rights of all individuals impacted by abstract classification, while advancing the understanding and organization of scholarly information in the English-speaking society.

5.4 Sustainability Plan

A sustainable approach to abstract classification in the English language involves long-term strategies and measures to ensure ongoing effectiveness and impact. This includes establishing collaborations with relevant stakeholders such as academic publishers, research institutions, and scholarly organizations to implement awareness campaigns, policies, and support systems. Continuous monitoring and evaluation of the classification methods and algorithms are essential to adapt to evolving research topics and trends. Regular updates and improvements to the classification model, incorporating user feedback and advancements in technology, can enhance its accuracy and relevance. Additionally, fostering a culture of academic integrity, resilience, and responsible research behavior through education and awareness programs can contribute to sustainable organization efforts and create a more robust and transparent scholarly environment for the English-speaking community.

Chapter 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

This study concentrates on abstract classification using deep learning techniques for the English language. The objective is to fill the research gap in abstract classification within non-English contexts and specifically cater to the linguistic and cultural nuances of the English-speaking community. The study aims to develop an effective model capable of identifying instances of abstracts in English text, particularly in scholarly articles. The study acknowledges the unique challenges associated with abstract classification in the English language, including language-specific complexities and academic factors. The outcomes of this research have the potential to contribute to the development of language-specific tools, policies, and interventions to enhance the accuracy and relevance of abstract classification, thereby fostering a more efficient and transparent scholarly environment for individuals communicating in English.

6.2 Conclusions

In summary, this study on abstract classification using deep learning for the English language offers valuable insights into addressing the prevalent challenge of abstract classification in non-English contexts. By devising a deep learning model tailored for the English language, the study takes into account the linguistic and cultural nuances that influence abstract classification dynamics. The research underscores the significance of language-specific methodologies and the efficacy of deep learning techniques, particularly recurrent neural networks with LSTM cells, in accurately classifying abstracts in English text. The study's findings fill a research gap on abstract classification detection in non-English languages, emphasizing the necessity for inclusive approaches that cater to diverse language communities. By concentrating on the English language, the study aims to empower individuals communicating in English, ensuring they receive equitable attention and protection in abstract classification. The research outcomes have practical implications for developing language-specific tools, policies, and interventions to enhance the accuracy and relevance of abstract classification in the English-speaking community. The study stresses the importance of ongoing monitoring, evaluation, and improvement of classification models to adapt to evolving abstract classification tactics. Furthermore, it underscores the significance of cultivating a culture of digital empathy, resilience, and responsible online behavior through education and awareness programs. Overall, this study establishes a foundation for future research and initiatives aimed at enhancing abstract classification in non-English languages, emphasizing collaborative efforts among researchers, policymakers, and educational institutions to create a more efficient and transparent scholarly environment for individuals communicating in English.

6.3 Implication for Further Study

The implications for further study lie in the expansion and refinement of the developed deep learning-based abstract classification model. Future research could explore the extension of this approach to multilingual contexts, assessing its applicability to languages beyond Bangla to enhance universality. Additionally, investigating the adaptability of the model across diverse scientific domains would provide valuable insights into its robustness and potential for cross-disciplinary application. These avenues for further study aim to broaden the scope and effectiveness of automated abstract classification, contributing to the development of more versatile and comprehensive systems for information categorization and retrieval.

Reference:

[1] Lumbanraja, F. R., Fitri, E., Junaidi, A., & Prabowo, R. (2021). Abstract classification using support vector machine algorithm (case study: abstract in a Computer Science Journal). In Journal of Physics: Conference Series (Vol. 1751, No. 1, p. 012042). IOP Publishing.

[2] Akshai, K. P., & Anitha, J. (2021, May). Plant disease classification using deep learning. In 2021 3rd International Conference on Signal Processing and Communication (ICPSC) (pp. 407-411). IEEE.

[3] Khasanah, I. N. (2021). Sentiment classification using fasttext embedding and deep learning model. Procedia Computer Science, 189, 343-350.

[4] Fadhil, I. M., & Sibaroni, Y. (2022, July). Topic classification in indonesian-language tweets using fast-text feature expansion with support vector machine (SVM). In 2022 International Conference on Data Science and Its Applications (ICoDSA) (pp. 214-219). IEEE.

[5] Adipradana, R., Nayoga, B. P., Suryadi, R., & Suhartono, D. (2021). Hoax analyzer for Indonesian news using RNNs with fasttext and glove embeddings. Bulletin of Electrical Engineering and Informatics, 10(4), 2130-2136.

[6] Kusumaningrum, R., Nisa, I. Z., Nawangsari, R. P., & Wibowo, A. (2021). Sentiment analysis of Indonesian hotel reviews: from classical machine learning to deep learning.

[7] David, M. S., & Renjith, S. (2021, September). Comparison of word embeddings in text classification based on RNN and CNN. In IOP Conference Series: Materials Science and Engineering (Vol. 1187, No. 1, p. 012029). IOP Publishing.

[8] Alfarizi, M. I., Syafaah, L., & Lestandy, M. (2022). Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory). JUITA: Jurnal Informatika, 10(2), 225-232.

[9] Boonchuay, K. (2019). Sentiment classification using text embedding for Thai teaching evaluation. Applied Mechanics and Materials, 886, 221-226.

[10] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[11] Munikar, M., Shakya, S., & Shrestha, A. (2019, November). Fine-grained sentiment classification using BERT. In 2019 Artificial Intelligence for Transforming Business and Society (AITB) (Vol. 1, pp. 1-5). IEEE.

[12] Rauf, S. A., Qiang, Y., Ali, S. B., & Ahmad, W. (2019). Using bert for checking the polarity of movie reviews. International Journal of Computer Applications, 975(8887).

[13] Maharani, W. (2020, June). Sentiment analysis during Jakarta flood for emergency responses and situational awareness in disaster management using BERT. In 2020 8th International Conference on Information and Communication Technology (ICoICT) (pp. 1-5). IEEE.

[14] Ravi, J., & Kulkarni, S. (2023). Text embedding techniques for efficient clustering of twitter data. Evolutionary Intelligence, 1-11.

[15] Khadhraoui, M., Bellaaj, H., Ammar, M. B., Hamam, H., & Jmaiel, M. (2022). Survey of BERT-base models for scientific text classification: COVID-19 case study. Applied Sciences, 12(6), 2891.

[16] Chen, X., Cong, P., & Lv, S. (2022). A long-text classification method of Chinese news based on BERT and CNN. IEEE Access, 10, 34046-34057.

[17] Mantas, J. (2021). The classification of short scientific texts using pretrained BERT model. Public Health and Informatics: Proceedings of MIE 2021, 281, 83.

[18] Rabbimov, I. M., & Kobilov, S. S. (2020, May). Multi-class text classification of uzbek news articles using machine learning. In Journal of Physics: Conference Series (Vol. 1546, No. 1, p. 012097). IOP Publishing.

[19] Bogdanchikov, A., Ayazbayev, D., & Varlamis, I. (2022). Classification of Scientific Documents in the Kazakh Language Using Deep Neural Networks and a Fusion of Images and Text. Big Data and Cognitive Computing, 6(4), 123.

[20] Barua, A., Sharif, O., & Hoque, M. M. (2021). Multi-class sports news categorization using machine learning techniques: resource creation and evaluation. Procedia Computer Science, 193, 112-121.

[21] Ali, D., Missen, M. M. S., & Husnain, M. (2021). Multiclass event classification from text. Scientific Programming, 2021, 1-15.

[22] Akter, S., Shamrat, F. J. M., Chakraborty, S., Karim, A., & Azam, S. (2021). COVID-19 detection using deep learning algorithm on chest X-ray images. Biology, 10(11), 1174.

[23] Hasib, K. M., Habib, M. A., Towhid, N. A., & Showrov, M. I. H. (2021, February). A novel deep learning based sentiment analysis of twitter data for us airline service. In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD) (pp. 450-455). IEEE.

[24] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232.

[26] Devi, U. M., & Marimuthu, A. (2021). Donor-Recipient for Liver Transplantation Using CNN and LSTM Deep Learning Techniques (No. 4923). EasyChair.

[27] Islam, R., Beeravolu, A. R., Islam, M. A. H., Karim, A., Azam, S., & Mukti, S. A. (2021, December). A Performance Based Study on Deep Learning Algorithms in the Efficient Prediction of Heart Disease. In 2021 2nd International Informatics and Software Engineering Conference (IISEC) (pp. 1-6). IEEE.

[28] Han, J., & Moraga, C. (1995, June). The influence of the sigmoid function parameters on the speed of backpropagation learning. In International workshop on artificial neural networks (pp. 195-201). Berlin, Heidelberg: Springer Berlin Heidelberg.

[29] Russell, S. J., & Norvig, P. (2010). Artificial intelligence a modern approach. London.

IDENTIFYING THE RESEARCH FIELD OF A SCIENTIFIC PAPER FROM THE ABSTRACT USING DEEP LEARNING APPROACHES

ORIGINALITY REPORT

18%	14%	8%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to Daffodil International University Student Paper	2%
3	sciforum.net Internet Source	<1%
4	danilocvieira.github.io Internet Source	<1%
5	fastercapital.com Internet Source	<1%
6	Submitted to Higher Education Commission Pakistan Student Paper	<1%
7	Submitted to University of Greenwich Student Paper	<1%
8	Dr. Amina NAIDJA, Dr. Saif Eddine CHETTAH, Dr. Dalal FARID, Dr. Meriem BENMACHICHE, Pr. Alkama Djamel. "A Comprehensive Review	<1%