A Machine Learning and Deep Learning Approach for Bengali News Headline Categorization

BY

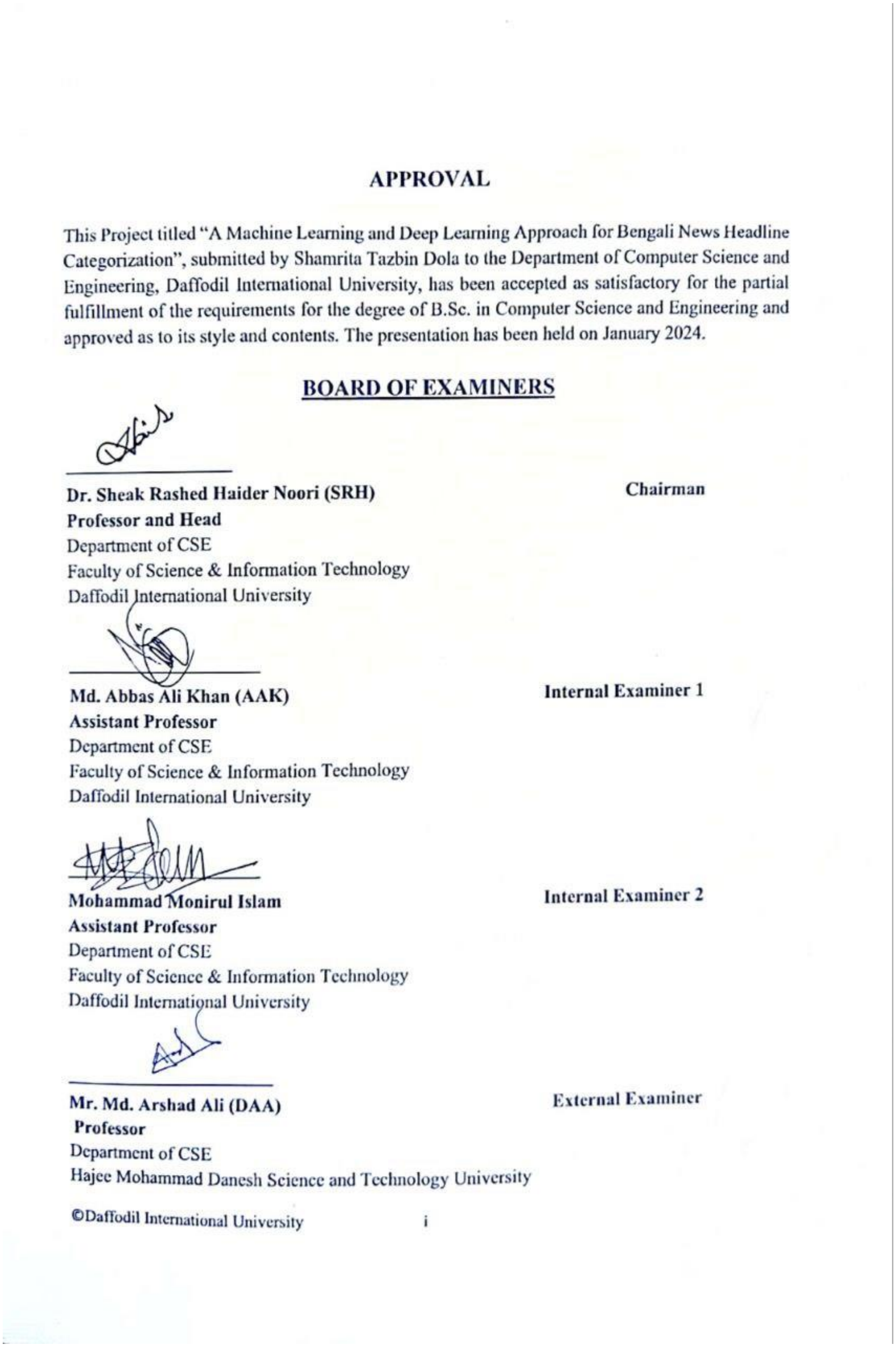
SHAMRITA TAZBIN DOLA
ID: 201-15-14087

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

Supervised By

Shayla Sharmin
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2024



APPROVAL

This Project titled "A Machine Learning and Deep Learning Approach for Bengali News Headline Categorization", submitted by Shamrita Tazbin Dola 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 January 2024.

BOARD OF EXAMINERS

Dr. Sheak Rashed Haider Noori (SRH)
Professor and Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Chairman

Md. Abbas Ali Khan (AAK)
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

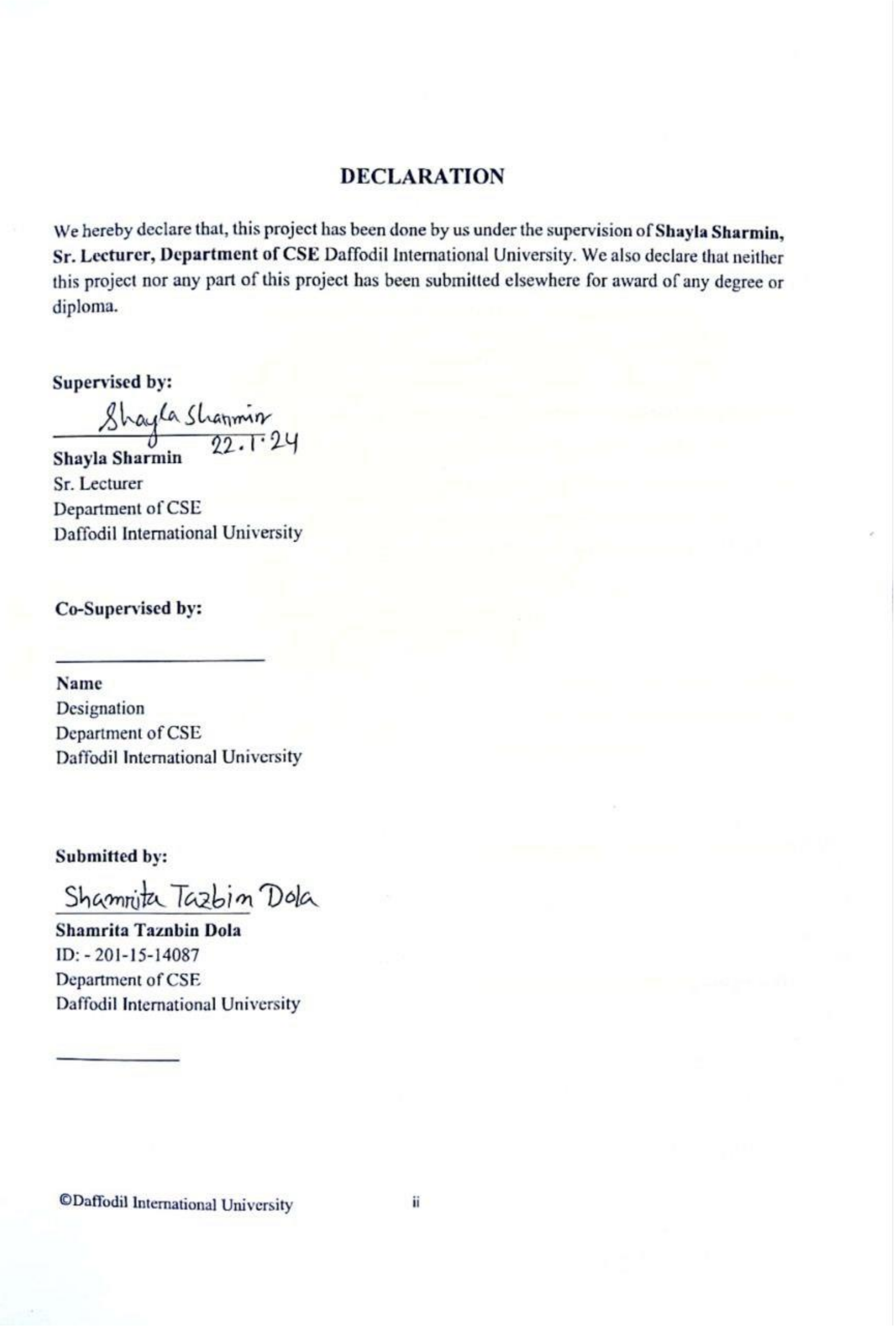
Internal Examiner 1

Mohammad Monirul Islam
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner 2

Mr. Md. Arshad Ali (DAA)
Professor
Department of CSE
Hajee Mohammad Danesh Science and Technology University

External Examiner



DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Shayla Sharmin, Sr. 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:

Shayla Sharmin
22.1.24

Shayla Sharmin
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

____

Name
Designation
Department of CSE
Daffodil International University

Submitted by:

Shamrita Tazbim Dola

Shamrita Taznbin Dola
ID: - 201-15-14087
Department of CSE
Daffodil International University

____

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Shayla Sharmin**, **Sr. Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Field name*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor & Head,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

The internet world is called a repository of information and data. Where there is a huge amount of information and data collection. Through internet people can access any kind of information and data from any place at any time. Current technology has made information and data readily available, due to which the amount of online news on the Internet has increased tremendously. Furthermore, because the internet is so widely available, people are growing increasingly eager to read news articles from news websites that use direct data. In general, online news portals are the terms used to describe Facebook, Twitter, WhatsApp, Telegram, Instagram, blogs, and other services. The quantity of news available on internet news portals is growing daily, and this growth is being matched by an increase in readers. All this online news are digital data, and with the volume of digital data is growing, so is the requirement for data categorization. Numerous methods, including machine learning, deep learning, transfer learning, and other data mining techniques, may be used to classify data. These algorithms classify data such that readers may deduce the news story's primary idea from the headlines alone. To address such issues, data in any language may be classified using natural language processing techniques. This article divides Bengali news stories into six categories: Politics, entertainment, sports, national, international, and IT. It does this by using deep learning and machine learning techniques. Numerous techniques, including BiLSTM, GRU, and Uni-Gram, as well as conventional machine learning algorithms, including SVM, MNB, RF Classifier, and LR, are used to select these classifications. The accuracy rates for these models are as follows: GRU achieves 84.01% accuracy, BiLSTM attains 83.42% accuracy, Logistic Regression performs at 64%, Multinomial Naive Bayes scores 61%, Random Forest Classifier achieves 65% accuracy, and Support Vector Machine also achieves 65% accuracy.

TABLE OF CONTENTS

CONTENTS PAGE

Board of examiners i

Declaration ii

Acknowledgements iii

Abstract iv

List of figures ix

List of tables x

CHAPTER

CHAPTER 1: INTRODUCTION 1-3

1.1. Introduction 1

1.2. Motivation 3

1.3. Relational of the Study 4

1.4. Research Questions 4

1.5. Expected Outcome 4

1.6. Report Layout 5

CHAPTER 2: BACKGROUND 4-9

2.1. Terminology 6

2.2. Related work 6

2.3. Comparative Analysis and Summary 9

2.4. Scope of the Problem 11

2.5. Challenges 12

CHAPTER 3: RESEARCH METHODOLOGY **10-35**

3.1. Dataset Preparation 13

3.2. Dataset Description 13

3.3. Statical Analysis 14

3.4. Design Approach 14

3.5. Proposed Methodology 15

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION **36-37**

4.1. Discussion 39

4.2. Experimental Results and Analysis 41

CHAPTER 5: IMPECT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY **39-40**

5.1. Impact on Society 42

5.2. Impact on Environment 42

5.3. Ethical Aspects 43

5.4. Sustainability 43

CHAPTER 6: SUMMARY, CONCLUSION, RECOMENDATION AND IMPLEMENTATION FOR FUTURE RESEARCH **41-42**

6.1. Summary of the Study 44

6.2. Conclusions 45

6.3. Implication for Further Study 45

REFERENCE **47**

LIST OF FIGURES

FIGURES	PAGE NO
Figure 1: The structure of the work process	15
Figure 2: Distribution of Data for Bi-LSTM	15
Figure 3: Statistics of All Category Data for Bi-LSTM	16
Figure 4: Bi-LSTM of Data Distribution Length Frequency	17
Figure 5: Bi-LSTM of Training and Validation Accuracy vs Epochs	18
Figure 6: Bi-LSTM of Training and Validation Loss vs Epochs	19
Figure 7: Bi-LSTM Algorithm of Confusion Matrix	19
Figure 8: Distribution of Data for GRU	21
Figure 9: Statistics of All Category Data for GRU	22
Figure 10: GRU of Data Distribution Length Frequency	22
Figure 11: GRU of Training and Validation Accuracy vs Epochs	24
Figure 12: GRU of Training and Validation Loss vs Epochs	25
Figure 13: GRU Algorithm of Confusion Matrix	25
Figure 14: Distribution of Data for ML	27
Figure 15: Statistics of All Category Data for ML	28
Figure 16: ML of Data Distribution Length Frequency	28
Figure 17: ROC Curve Graph	37
Figure 18: Cross-Validation ROC Curve Graph	38

LIST OF TABLES

TABLES	PAGE NO
Table 1: Summary of Comparatives Analysis Paper	11
Table 2: Description of Each Category	14
Table 3: Bi-LSTM of Validation Accuracy and Training accuracy	17
Table 4: Bi-LSTM of Validation Loss and Training Loss	18
Table 5: Bi-LSTM Algorithm of Classification Report	20
Table 6: GRU of Validation Accuracy and Training Accuracy	23
Table 7: GRU of Validation Loss and Training Loss	24
Table 8: GRU Algorithm of Classification Report	26
Table 9: Logistic Regression Algorithm of Classification Report	29
Table 10: Multinomial Naive Bayes of Classification Report	30
Table 11: Random Forest Classifier of Classification Report	31
Table 12: SVM of Classification Report	32
Table 13: Decision Tree Classifier of Classification report	33
Table 14: Gradient Bosting Algorithm of Classification report	34
Table 15: SVM of Classification report	34
Table 16: Logistic Regression of Classification report	35
Table 17: Random Forest Classifier of Classification report	36
Table 18: Accuracy Table of All Classifiers	36
Table 19: Decision Tree Classifier of Classification report	37
Table 20: Description Table of All Classifiers	40
Table 21: Accuracy, Recall and Precision Score of All Classifiers	41

CHAPTER 1
INTRODUCTION

1.1. Introduction

Currently, linguistic problems are tried to be solved by NLP within the data science and artificial intelligence fields. Transfer learning, machine learning and deep learning algorithms are the most widely used and beneficial algorithms for comprehending NLP-type issues. All these types of learning are automatic learning systems. There are several approaches such as supervised (data with input and output labels), unsupervised (data with input and output unlabeled) and semi-supervised (both mixed labeled and unlabeled data) learning [2].

Ever since the world became modern, research has been going on about everything. Over the past few decades, researchers have conducted many studies based on text classification. Doc2Vec, Word2Vec, TF-IDF are quite popular in finding meaningful features for language classification in that research. There are several additional well-liked supervised models for text classification such as SVM, KNN or NB etc. Also, there are several popular supervised models for text classification such as SVM, KNN or NB etc. [3].

An important aspect of data science is text mining, which helps the user to find meaningful and interesting information. Searching the internet for specific things is time consuming for anyone. So, searching news through news sorting can cut down on time. That is, it helps people to find relevant news. Many have researched it and found success by categorizing headlines into different categories. That is why researchers are indicated to work on news headlines in a more computer-based manner [4].

Daily hacks for news publishers can positively impact news development if news headlines can be sorted based on the level of news severity. This can be done through sentiment analysis, which is based on assumptions. Emotions after reading news i.e., based on the news can be divided into different categories. These days, researchers are concentrating on sentiment analysis, It uses text analysis, NLP, and computational linguistics to categorise and extract news [5].

Internet is the main source of information in today's world. Today, people's lives are not possible without the internet. Reading news portals from online news sources has been more popular in recent times due to the accessibility of the internet. In today's digital era, no one is interested in reading news in newspapers anymore, everyone is reading news through internet. Internet has become so accessible in today's world that people now prefer to read news from online news portals instead of reading news from offline papers. Facebook, Twitter, WhatsApp, Telegram, Instagram, and so on are examples of online news portals. Any company needs a certain amount of space to print news that cannot be written in a newspaper and that certain amount of space requires a cost. However, writing for websites, blogs, online news portals, Facebook, Twitter, WhatsApp, Telegram, Instagram, and other platforms is limitless. So, there is an opportunity to write big news online. As a result, the amount of news on onwebsites, blogs, social media, and online news portals are growing daily. That is, because of all these reasons, people of the world are becoming interested and dependent on online news portals instead of reading news from offline papers. There is a plethora of important and irrelevant news on virtual entertainment timelines or newsfeeds. Social media is the means of connection for 70% of internet users globally. It is almost 90% among youth. Approximately 80% of Bangladeshi online users use Facebook, according to research. The degree of social communication has increased overall due to the usage of the web more than in the past. People exchange information and documents in many ways via innovation, including photographs, films, thoughts, and more about themselves.

In present, social media and the internet provide access to a vast array of information. We receive information of any sort in a matter of seconds on events occurring anywhere on Earth. We receive online reports regarding climate within a short time from the Meteorological Department. We may access a wide range of information over the internet, such as the locations of cyclones, floods, and twisters as well as global environmental conditions. We are able to quickly ascertain the current weather conditions in any nation or location. We don't base our decisions on what should happen sooner or later on the climate or weather forecast.

Nowadays it is not enough to get news from online, the news is highly refined and classified as only refined and classified news can be consumed. Filtering and categorizing news is not a difficult task in this modern age. Data science through artificial intelligence has many algorithms such as machine learning algorithm, transfer learning algorithm, deep learning algorithm etc. which can be used to classify any text. Also, many other tools like ChatGPT, google birds etc. also take its help. However, algorithms and tools are used to classify news. That is, by looking at the news headlines through sentiment analysis, people can understand what kind of news or what kind of news.

The number of these e-news and online news portals in Bangladesh and the world is increasing day by day. These e-news and online news portals are available on various social sites. This means providing regular breaking and updated news on our favorite sites and the sites the team spends most time on. Because the news of the site is refined and categorized, which is very useful and easy to understand for the readers.

1.2. Motivation

The field of automatic text categorization is still developing. Text categorization is becoming more and more necessary to handle the world's expanding digital data, which is developing at an exponential rate. In addition to other data mining procedures, machine learning techniques are utilised for text classification. Applications like content tagging, spam filtering, and business analytics also make use of text categorization. Websites in the Bengali language contain a vast quantity of info, most of which is hard to discover. However, classifying readers and hash-tagging keywords become crucial tasks if you wish to post on a forum. Status categorization has become crucial for the Bengali language since there has never before been any application platform that uses text classification, which is a major issue. Once more, because text mining is becoming more and more necessary, academics from Bangladesh and India are concentrating on developing text mining applications. There are a lot of Bengali studies on sentiment analysis and text mining that demonstrate effective removal. In order to quickly grasp the tone and agenda of the news, the majority of people first scan the headlines before reading the actual news. According to NLP, every kind of linguistic issue pertaining to categorization may be resolved. The best algorithms for comprehending NLP difficulties are machine learning algorithms since these challenges try to produce concepts from human language problems and articulate them. Machine learning may be divided into three categories: semi-supervised, unsupervised, and supervised. Learning that is supervised, unsupervised, and semi-supervised. Levelled data and inputs and outputs are essential components of supervised learning. Level-free data must be provided together with input and output for unsupervised learning to take place. Labelled and unleveled data mixed with supervised and unsupervised learning is semi-supervised learning. But categorising that much information is demanding, complicated, and time-consuming. These time-consuming, tough, and complex algorithms are defeated by machine learning algorithms. With the advancement of machine learning over the past several years, text categorization has been used extensively. Text classification in the form of news headline classification may be broadly broken down into three stages: feature extraction, classifier extraction, and assessment. Another feature of e-news in Bangladesh is that viewers favour websites that provide frequent updates and breaking news. Pratham Alo, Bangladesh Pratid, Nayadigant, Jugantar, Samakal, and others are among the five well-known websites. It is evident from an analysis of Google trend data that fewer people are reading "Daily News" online on a daily basis.

1.3. Relational of the Study

To determine which category a news item belongs in, our study paper is divided into six sections. Six categories— Politics, IT, International, National, Sports, and Entertainment —are used in our research to group the news. We discovered other such studies; they classified the news similarly to us and used various algorithms, such as DL & ML, to extract the security.

1.4. Research Questions

This research may include a wide range of question kinds. As an illustration:

i. What is the aim of this study?
ii. What is the purpose of this study?
iii. Is there any way this study can help us?
iv. How can the results of this research help us?
v. What are the investigation's findings?
vi. What is the study's conclusion?

1.5. Expected Outcome

News is a kind of information that may be utilised for any kind of activity, including futures. Since the goal of our research is to categorise news headlines, after it is finished, we may use any kind of data to determine which category the news belongs in. Furthermore, a great deal of useless unclassified data may be categorised and turned into useful data, saving a great deal of data from being lost.

1.6. Report Layout

However, the news portals are wasting a lot of potential data because they are not properly categorising, sorting, and analysing the data that they collect from internet and social media sources. This categorization can be a big potential answer if with machine learning (ML), deep learning (DL), and natural language processing (NLP), the data may be automatically classified more quickly, more flexible, more affordable, and more trustworthy manner. The remaining portion of the essay is structured as follows: Part 2: Background; Part 3: Methodology; Part 4: Experimental result and debate; and Part 5: Impact on sustainability, society, and the environment. The paper will finally be concluded in Section 6.

CHAPTER 2
BACKGROUND

2.1. Terminology

Because of the world's perpetual expansion, events take place everywhere. And because of the internet, the aforementioned occurrences are quickly going viral on social media. In this sense, the volume of news keeps growing over time. However, the purpose is not to classify the news—that is, not to specify which news items go into which categories. As a result, a significant amount of internet data is lost and is never recovered. Therefore, a variety of machine learning techniques are used to classify the data in order to make all of the documents or data useful. in order for the facts of today to be useful for tomorrow.

2.2. Related Works

Adrita Barua and colleagues [1] have presented a research paper focusing on the classification of language recognition within the realm of natural language processing in machine learning. They have explored six machine learning algorithms, including SVC, TF-IDF, DT, LoLR, MNB, and RF. This classification effort involves four categories: Football, Tennis, Cricket, and Athletics. The highest accuracy, 97.60%, was achieved using the SVC

Md. Majedul Islam and his team [2] have introduced a research paper employing various supervised learning algorithms in the context of NLP to detect the sentiment of Bengali words. Five machine learning algorithms were examined, including KNN, DT, SVM, RF and NB. Among these, the SVM algorithm achieved the highest accuracy of 75%.

Ettilla Mohiuddin Eumi et al. [3] have proposed a research paper with a primary focus on the classification of Bengali news headlines using the BiGRU. They have experimented with several algorithms, including Scikit Learn Library, LibSVM without stop words, LibSVM with stop words, LR, RF, SVM, NB, BiGRU model, and Sequential Deep Learning model, while categorizing headlines into six distinct groups: Amusement, IT, International, Politics, National, and Sports. The BiGRU model demonstrated the highest accuracy at 84%.

Md. Rafiuzzaman Bhuiyan and his team [4] have presented a research paper concentrating on the classification of news headlines using the LSTM model. With a dataset of 4580 training data points, the headlines were categorized into four different segments: Science, Sports, Rational, and International. The LSTM model achieved an accuracy of 91.22%.

Hoda Ahmed Galal Elsayed et al. [5] have proposed a research paper trying to use a variety of machine learning algorithms to determine the psychological effects of Arabic news headlines on readers. There were seven distinct emotional groups created from the headlines: Feelings: Surprise, Happiness, Neutral, Sadness, Anger, Disgust, Fear. Among the six machine learning algorithms examined, the multilevel CNN algorithm exhibited the highest accuracy of 89.3%.

Ronald Tudu and colleagues [6] have presented a research paper that focuses on using machine learning methods to the classification of Bengali text titles. The headlines were divided into ten distinct categories. Several models, such as Stochastic Gradient Descent (SGD), LR, SVM, MNB classifier, were employed. The SVM algorithm demonstrated the highest accuracy at 87.5%.

Fatema Jahara et al. [7] have introduced a research paper discussing the use of the Multilayer Perception (MLP) classifier framework to classify newspaper headlines based on deep learning. The headlines were categorized into four segments: Entertainment, Sports, Accidents, and Crime. Among all the algorithms explored in the paper, the MLP algorithm achieved an accuracy of 98.18% for news articles and 94.53% for news headlines.

Raghad Bogery and her team [8] have proposed a research paper comparing and contrasting different machine learning algorithms, including those related to NLP, for classifying a substantial number of news headlines. Five machine learning classifiers, including KNN, SVM, MNB, NB and GB, were evaluated. Among these, the MNB algorithm displayed the best performance accuracy, reaching 90.12%, with a recall of 90%. The headlines were divided into three different categories: Travel, Style & Beauty, and Parenting.

Md. Ferdouse Ahmed Foysal and colleagues [9] have proposed a research paper that employs the LSTM algorithm to classify news headlines, achieving an accuracy of 84%. These headlines were categorized into five distinct groups: entertainment, national, sports, city, and state news.

Ke Yahan et al. [10] have presented a research paper that classifies a dataset comprising 18 years of news using machine learning algorithms, including SVC, NN, DT, RF. The highest accuracy of 86.22% was achieved using the NN algorithm.

Prakash Kumar Sing and his team [11] have introduced a research paper focusing on sentiment analysis with a particular emphasis on negation handling. Deep neural network models, including LSTM, SVM, HMM, and CRF, were employed. Among these, the BiLSTM model attained the highest accuracy of 93.34%.

Mohammad Rabib Hossain and colleagues [12] have proposed a research paper centered around Bengali News Categorization, applying various machine learning and deep learning models such as SVM, NB, RF, LR, BiLSTM, and CNN for news classification. Of all the algorithms examined, the CNN model achieved the highest accuracy of 93.43%.

Shazia Usmani and her team [13] have introduced a research paper that employs NLP-based techniques to classify unlabeled news headlines from the Pakistani Stock Exchange.

Sharun Akter Khushbu and colleagues [14] have proposed a research paper that utilizes Neural Network Based Extraction to categorize news types from Bengali news headlines. Five machine learning models, including LR, NN, SVM, NB AND RF, were examined. Among these, the Neural Network (NN) algorithm outperformed others with a 90% accuracy rate.

Ruichao Wang and colleagues [15] have presented a research paper that recognises news headlines with a hybrid approach enhanced with machine learning. Various tests were conducted with systems like Hybrid, Trim, TFTrim, HybrideTrim, UTD, Topiary, and TF. The TFTrim system yielded the best results among all the tests.

Syeda Sumbul Hossain and her team [16] have proposed a research paper conducting a comparative analysis based on Machine Learning and Deep Learning Algorithms to analyze the sentiment of news headlines. Of the seven machine learning models and two deep learning models explored, Bernoulli NB in machine learning and CNN in deep learning displayed superior performance.

Paulo Santos and colleagues [17] have introduced a research paper focusing on sentiment analysis to classify Portuguese news headlines. The models Random Forest_1, Random Forest_2 classifier, and Sequential Minimal Optimization (SMO) were employed, achieving accuracy rates of 62.50%, 57.50%, and 61.00% without relation as a feature, and 62.70%,

In their research paper, Uchchhwas Saha and colleagues [18] have undertaken a sentiment analysis of Bengali comments employing a hybrid approach that combines firsttext and deep learning classifiers. This hybrid model includes models like "Adam" and "Glove," alongside BLSTM and CNN within the deep learning framework. Among all the algorithms examined, the hybrid model demonstrated the highest accuracy at 89.89%.

A.N.M. JuBaer and his team [19] have introduced a research paper dedicated to classifying toxic comments using a combination of Machine Learning and Deep Learning algorithms. They have explored various models, including MNB, SVM, MNB (from scikit-learn), GNB, Classifier Chain with MNB, Label Powerset with MNB, MLkNN, and BP-MLL NN. Among these algorithms, the BPMLL neural network produced the most promising results.

Mushfiqus Salehin and colleagues [20] have proposed a research paper centered around Sequence-to-sequence model based on attention mechanism for Bengali news headline categorization. Their work involves the use of their own Bengali dataset, yielding positive results that align with findings from other research papers.

2.3. Comparative Analysis and Summary

Table-1: Summary of Comparatives Analysis Paper

Paper No	Authors & Year	Used All Models	Height Accuracy Model	Height Model (%)
1	Adrita Barua (2021)	SVC,DT,RF,MNB,LR,TF-IDE	SVC	97.60%
2	Md. Majedul Islam (2019)	RF, NB, DT, KNN, SVM	SVM	75%
3	Ettilla Mohiuddin Eumi (2021)	With stop words in LibSVM, LibSVM devoid of stop terms, Sequential deep learning, BiGRU, NB, LR, RF, and Scikit Learn Librery	BiGRU	84%

4	Md. Rafiuzzaman Bhuiyan (2021)	LSTM	LSTM	91.22%.
5	Hoda Ahmed Galal Elsayed (2020)	SVM, RF, NB, KNN, Multilevel CNN, zeroR, and DT	Multilevel CNN	89.3%.
6	Ronald Tudu (2018)	LR classifier, SVM, MNB, SGD	SVM	87.5%.
7	Fatema Jahara (2022)	A multilayer perception (MLP) classifier system based on deep learning	MLP (news articles) & MLP (news headlines)	98.18% & 94.53%
8	Raghad Bogery (2019)	NB, MNB, KNN, SVM, and GB	Multinomial Naïve Bayes	90.12%
9	Md. Ferdouse Ahmed Foysal (2021)	LSTM	LSTM	84%
10	Ke Yahan (2018)	NN, SVC, RF, DT	NN	86.22%
11	Prakash Kumar Sing (2021)	HMM, CRF, SVM, and LSTM	LSTM	93.34%
12	Mohammad Rabib Hossain (2020)	ML- SVM, NB, RF, LR DL- BiLSTM, CNN	CNN	93.43%
13	Shazia Usmani (2020)	NLP-based methodology	-	-
14	Sharun Akter Khushbu (2020)	RF, SVM, NB, NN, LR	NN	90%
15	Ruichao Wang (2014)	TFTrim, HybrideTrim, Topiary, TF, Hybrid, Trim, UTD	TFTrim	Maximum Accuracy

16	Syeda Sumbul Hossain (2021)	ML- Stochastic gradient descent (SGD), SVC, Bernoulli NB, NB, multinomial NB, Nu support vector classifier, and LR. DL- LSTM, CNN.	ML- Bernoulli NB DL- CNN	ML- 82.68% DL- 70.33%
17	Paulo Santos (2015)	Ramdom Forest_2, SMO, and Random Forest_1 classifier	SMO (Without relations as features) Ramdom forest_2 (With relations as features)	62.50% 63.50%
18	Uchchhwas Saha (2022)	CNN, FastText, BiLSTM, and Hybrid.	Hybrid	89.89%
19	A.N.M. JuBaer (2019)	BP-MLL NN, Classifier Chain with Multinomial NB, Label Powerset with Multinomial NB, MLKNN, Gaussean NB, Multinomial NB, SVM, and Multinomial NB (from scikit-learn).	BPMLL Neural Network	60.00%
20	Mushfiqus Salehin (2019)	Model of sequence-to-sequence based on attention mechanism.	-	Best Accuracy

2.4. Scope of the Problem

As time moves on, different things happen all around the world at different periods. Again, they are available because of the internet thanks to social media sites like Among them are Google, YouTube, Facebook, Twitter, WhatsApp, Viber, Pinterest, Telegram, and Messenger. As a result, the volume of news keeps growing daily. The quantity of data or papers available online has increased along with the amount of news. But not to categorise the aforementioned information or documents—that is, not to specify which information belongs in which group. As a result, a significant amount of internet data is lost and is never recovered. To make all of these documents or data usable, a range of machine learning and deep learning techniques are therefore employed to categorise the data, including LgR, LnR, RF classifier, DT, NB, SVM, CNN, RNN, ANN, LSTM, and others. As a result, there are several uses for the papers or data that may be found online.

2.5. Challenges

The amount of news that is available online is growing daily in the modern world; within a few days, there will be more useless info than useful data. The inability to categorise the aforementioned data or documents—that is, to determine which data belongs in which category—is the cause of their unusability. As a result, a significant amount of internet data is lost and is never recovered. We will have a lot of issues later on if we don't make all of these data more usable. It will require a great deal of time and specialised personnel in the future to use data to make improvements. Both money and time will be wasted on this. Therefore, the data will be helpful and free from issues in the future if it is received today, i.e., whether information that is accessible via social media or other channels is classified.

CHAPTER 3
RESEARCH METHODOLOGY

3.1. Dataset Preprocessing

The information gathered via the internet is primarily unstructured. Therefore, processing the data is required once it has been collected. Due to the fact that data gathered from websites, blogs, and social media platforms sometimes contains null values, redundant data, and other errors. The dataset is cleaned, low length data is eliminated, unique data is eliminated, duplicate data is eliminated, and the dataset is converted into intelligible data. In addition, several dataset features have been extracted, such as word lists, sorted word lists, documents in each class, words in total, words that are unique in each class, and dataset splitting, etc. As a result, we must preprocess our dataset using data techniques. To use the pre-processing approach, take the actions listed below:

- Take out hashtags, screen names, and URLs.
- Eliminate all zero values, punctuation, symbols, emojis, and integers.
- Eliminate any superfluous symbols and retweets.
- Eliminate all short data.

3.2. Dataset Description

Three columns make up the primary dataset utilised in this study: Newspaper Name, Category, and Headline. Six groupings make up the categories: Politics, Sports, Entertainment, International, National, and IT. To enable the application of several machine learning and deep learning methods, we divided the dataset simply: we set aside 23,889 samples for validation, 13,272 samples for testing, and 95,552 samples for training.

Table-2: Description of Each Category

Source News	Category	Description	Label
	International	All forms of international news will be aired.	

Online News Headline	National	All forms of national news will be aired.	International, National, Sports, Amusement, Politics and IT.
	Sports	All news categories will include sports news.	
	Amusement	All news categories will have entertaining news.	
	Politics	All news categories will have political news.	
	IT	All news categories will have IT news shown.	

3.3. Statistical Analysis

- There are 136811 total data points in the dataset.
- The dataset maintains three columns.
- There are 95552 training data points in the dataset overall.
- There are 13272 testing data points in the dataset.
- There are 23889 validation data in all in the dataset.
- Six phases are used to categorise categories: international, national, sports, entertainment, politics, and IT.

3.4. Design Approach

We employed both supervised and unsupervised learning techniques as the data presents a multivariate categorization problem. Among the machine learning approaches are GB algorithm, LR, RF, SVM, multinomial naive bayes, and decision tree classifier. Deep learning techniques include BiLSTM, GRU, and Language Model (Uni-Gram). For every model, a confusion matrix, performance and accuracy forecasts, and an outcomes analysis were produced. Figure displays the overall system architecture diagram:

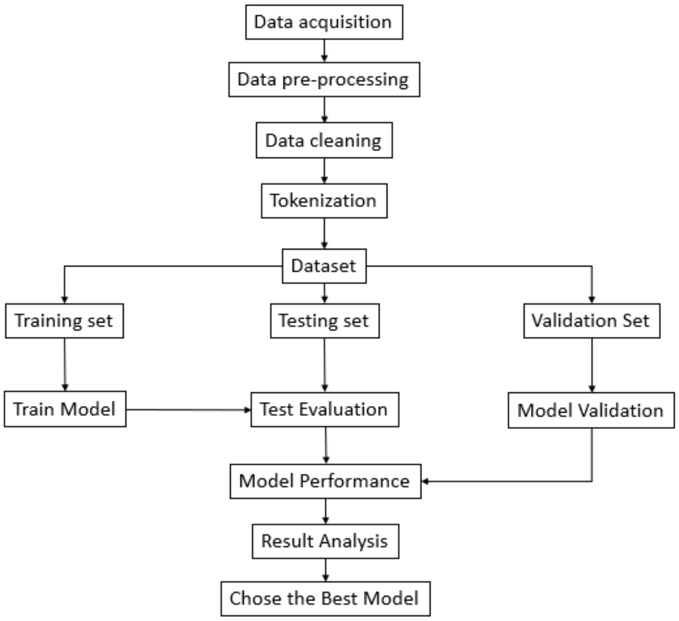


Figure-1: The structure of the work process

3.5. Proposed Methodology

3.5.1. Bidirectional Long Short-Term Memory (BiLSTM)

Distribute the dataset when it has been prepared. The collection has 136811 headlines in total. There are six categories in which the 136811 headlines are 15isualizati: international, national, athletic, entertainment, political, and IT. The most headlines are from the international category, followed by sports, national, and finally, IT-related headlines. Here is a data distribution graph:

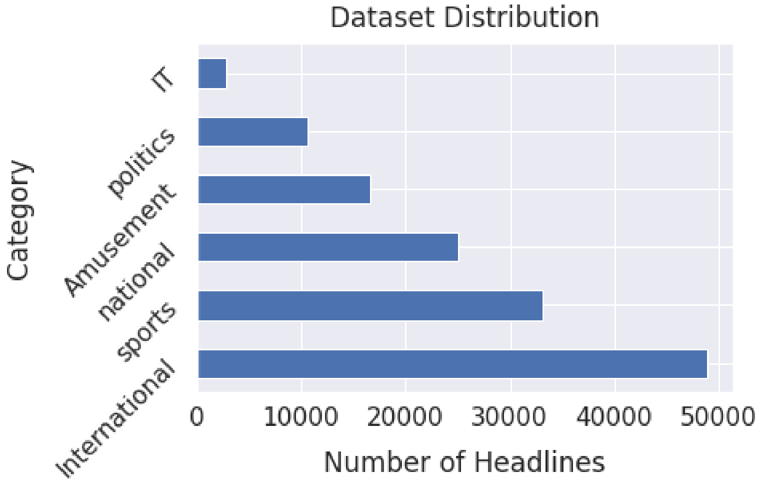


Figure-2: Distribution of Data for Bi-LSTM

Following data preparation, 4098 headlines were cleaned out of the dataset, bringing the total to 132713. Next, by examining the complete dataset, every distinct category is examined once again among all of the headlines, and the quantity of headlines, words, unique words, and most frequently occurring terms of each particular category are retrieved. Thus, the most common word (307354), unique word number (28710), headline number (47885), and word number (307354) of the international class have been retrieved through analysis. Through examination of the sports class, the most common word (30831), word number (152852), unique word number (18581), and headline number (30831) have been identified. The quantity of headlines, words, unique words, and most common terms related to IT, politics, entertainment, and the country have also been extracted. The dataset analysis graph is displayed below:

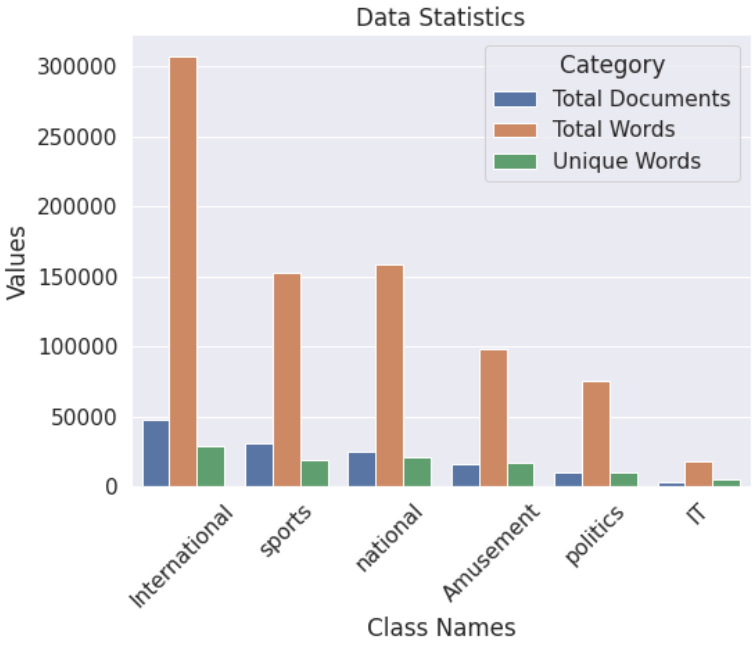


Figure-3: Statistics of All Category Data for Bi-LSTM

Following the 16isualization of the dataset, measurements were made of the headlines' length and frequency. Of them, the title's greatest length is 21, its minimum length is 3, and its average length is 6. The graph of the length frequency distribution is shown below:

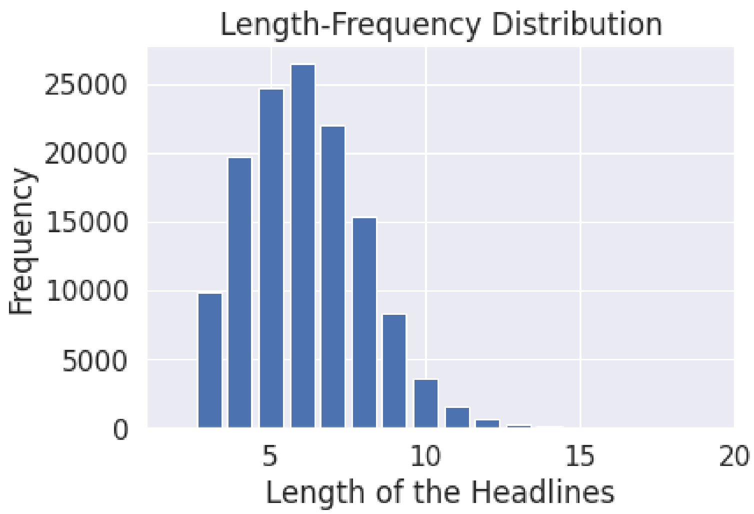


Figure-4: Bi-LSTM of Data Distribution Length Frequency

Label encoding is used in the encoding of this dataset. Label encoding is the process of translating the dataset's transform a category language into a numerical language, allowing the original may be translated into a machine language that the computer can understand. In essence, categorical language is converted to numerical language using the "LabelEncoder()" method.

The dataset is divided once all of the data has been converted to numerical form. The training data and the test data comprise the two halves of the dataset. Of the total 132713 data, 95552 are for training, of which 55055 are unique, and 13272 are for testing. There are 23889 pieces of data that are retained for validation checks.

3.5.1.1. Validation and Training accuracy

Table-3: Bi-LSTM of Validation Accuracy and Training Accuracy

Epoch No	Validation Accuracy	Training Accuracy
Epoch 1	0.81330	0.7239
Epoch 2	0.82992	0.8747
Epoch 3	0.83114	0.9203
Epoch 4	0.83114	0.9437

Epoch 5	0.83114	0.9571
Epoch 6	0.83114	0.9672
Epoch 7	0.83114	0.9741

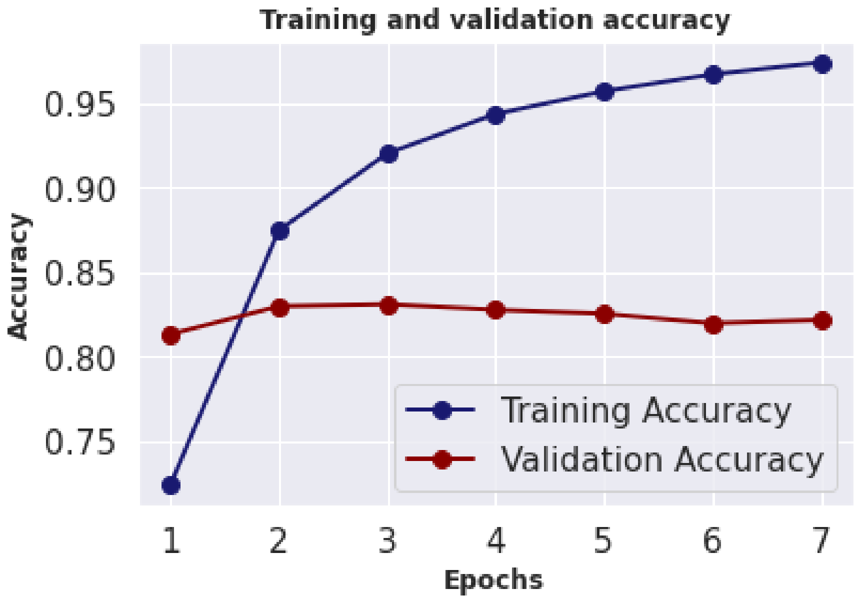


Figure-5: Bi-LSTM of Training and Validation Accuracy vs Epochs

3.5.1.2. Validation and Training loss

Table-4: Bi-LSTM of Validation Loss and Training Loss

Epoch No	Validation Loss	Training Loss
Epoch 1	0.5203	0.7419
Epoch 2	0.4690	0.3552
Epoch 3	0.5090	0.2279
Epoch 4	0.5495	0.1631

Epoch 5	0.6179	0.1212
Epoch 6	0.6767	0.0912
Epoch 7	0.7656	0.0707

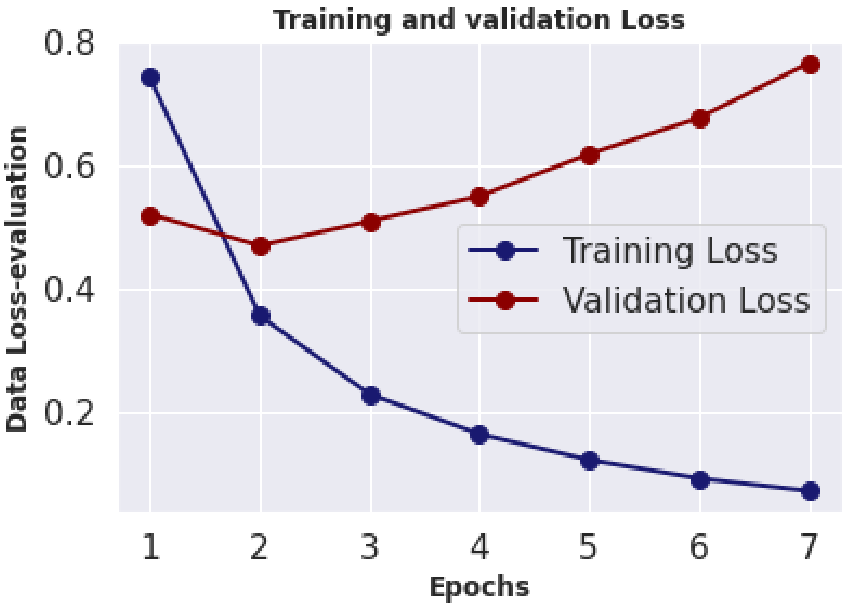


Figure-6: Bi-LSTM of Training and Validation Loss vs Epochs

3.5.1.3. Confusion matrix

The confusion matrix shows that the accuracy of the Bi-LSTM report is 83.42%.

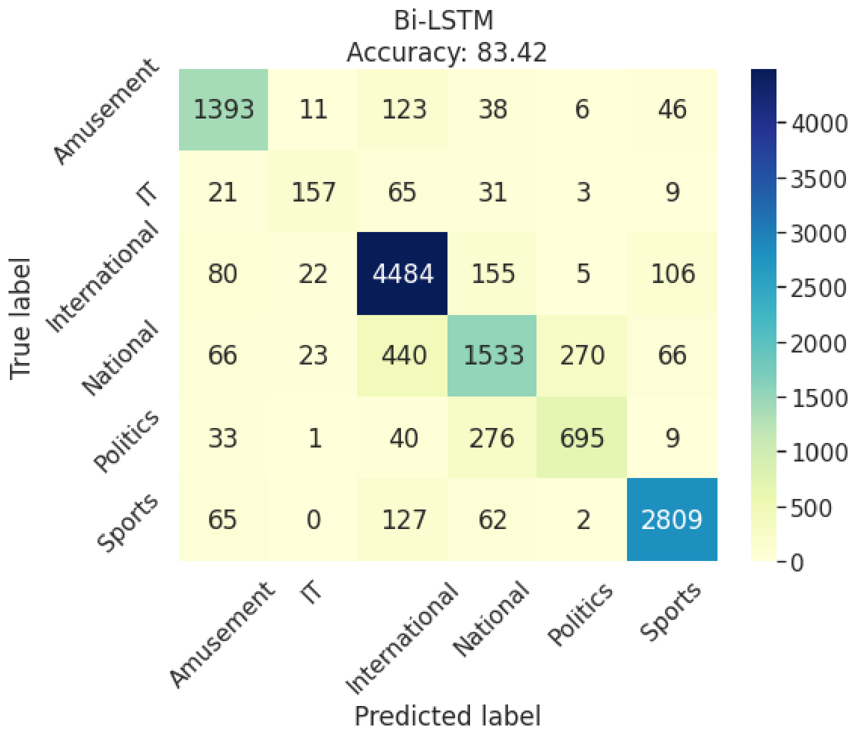


Figure-7: Bi-LSTM Algorithm of Confusion Matrix

3.5.1.4. Classification report

The B-LSTM was found to have an 83.42% assurance report by looking at the classes and the classification report, which comprises precision, recall, f1_score, and support. Headline News: Politics, Sports, Entertainment, IT, International, and National.

Table-5: Bi-LSTM Algorithm of Classification Report

News Type	Precision	Recall	F1_Score	Support
Amusement	84.02	86.15	85.07	1617.000000
IT	73.36	54.90	62.80	286.000000
International	84.94	92.42	88.52	4852.000000
National	73.17	63.93	68.24	2398.000000

Politics	70.85	65.94	68.30	1054.000000
Sports	92.25	91.65	91.95	3065.000000
Accuracy	83.42	83.42	83.42	0.834162
Macro Avg	79.77	75.83	77.48	13272.000000
Weighted Avg	83.02	83.42	83.07	13272.000000

For every type of news (amusement, IT, international, national, politics, sports), precision, recall, F1_Score, and support are obtained as can be seen in the categorization report. Accuracy, Macro Average, and Weighted Average values have also been determined. The BiLSTM algorithm has an accuracy of 83.42.

3.5.2. Gated recurrent units (GRU)

The dataset should first be distributed or prepared. The collection has 136811 headlines in total. There are six categories in which the 136811 headlines are categorised: international, national, athletic, entertainment, political, and IT. The most headlines are from the international category, followed by sports, national, and finally, IT-related headlines. The data set distribution graphs of GRU and BiLSTM demonstrate how close the two algorithms' data set distributions are to one another. Here is a data distribution graph:

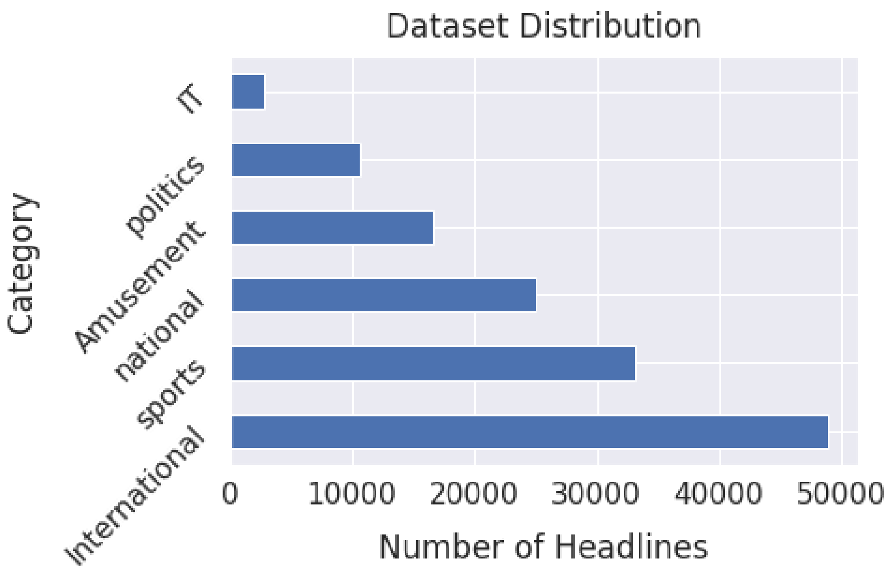


Figure-8: Distribution of Data for GRU

Following data preparation, 4098 headlines were cleaned out of the dataset, bringing the total to 132713. Next, by examining the complete dataset, every distinct category is examined once again among all of the headlines, and the quantity of headlines, words, unique words, and most frequently occurring terms of each particular category are retrieved. Thus, the most common word (307354), unique word number (28710), headline number (47885), and word number (307354) of the international class have been retrieved through analysis. Through examination of the sports class, the most common word (30831), word number (152852), unique word number (18581), and headline number (30831) have been identified. The quantity of headlines, words, unique words, and most common terms related to IT, politics, entertainment, and the country have also been extracted. The data collection yielded 57490 unique words in total. GRU and BiLSTM, two data statistics The data statistics graphs of these two methods are nearly identical, as seen by their respective graphs. The dataset analysis graph is displayed below:

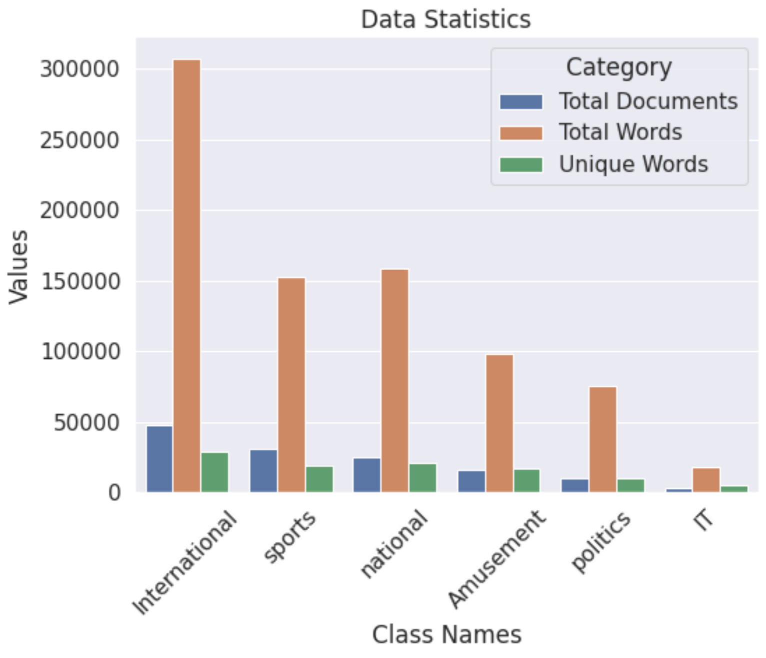


Figure-9: Statistics of All Category Data for GRU

Following the visualisation of the dataset, measurements were made of the headlines' length and frequency. Of them, the title's greatest length is 21, its minimum length is 3, and its average length is 6. The Length-Frequency Distribution graphs of GRU and BiLSTM demonstrate how similar the two algorithms' respective graphs are to one another. The graph of the length frequency distribution is shown below:

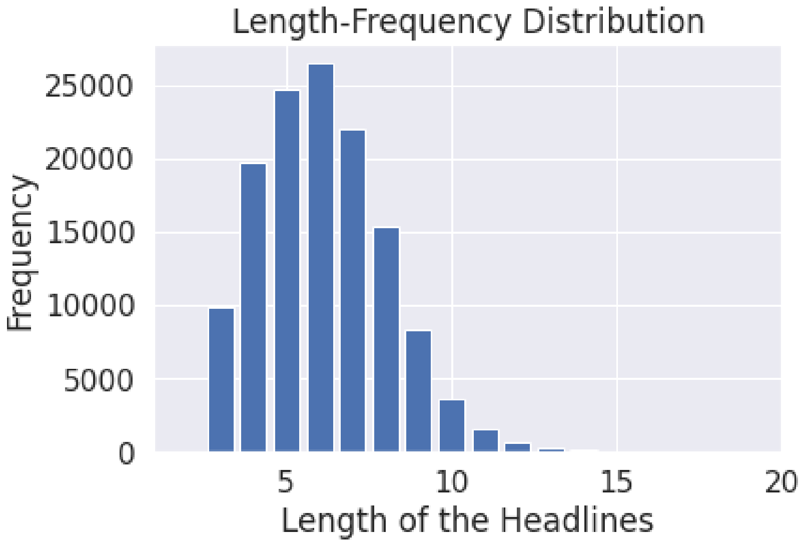


Figure-10: GRU of Data Distribution Length Frequency

Label encoding is used in the encoding of this dataset. Label encoding is the process of translating the dataset's transform a category language into a numerical language, allowing the original may be translated into a machine language that the computer can understand. In essence, categorical language is converted to numerical language using the "LabelEncoder()" method.

The dataset is divided once all of the data has been converted to numerical form. The training data and the test data comprise the two halves of the dataset. Of the total 132713 data, 95552 are for training, of which 55055 are unique, and 13272 are for testing. There are 23889 pieces of data that are retained for validation checks.

The update gate and reset gate are the two gates used by the recurrent neural network gating mechanism known as GRU. Machine learning challenges linked to memory and grouping are handled by GRU. The GRU model is primarily used in several research projects, including voice recognition, handwriting analysis, and the human genome.

3.5.2.1. Validation and Training accuracy

Table-6: GRU of Validation Accuracy and Training Accuracy

Epoch No	Validation Accuracy	Training Accuracy
Epoch 1	0.8205	0.7429
Epoch 2	0.8320	0.8749
Epoch 3	0.8365	0.9203
Epoch 4	0.8325	0.9426
Epoch 5	0.8278	0.9549
Epoch 6	0.8282	0.9644
Epoch 7	0.8230	0.9707

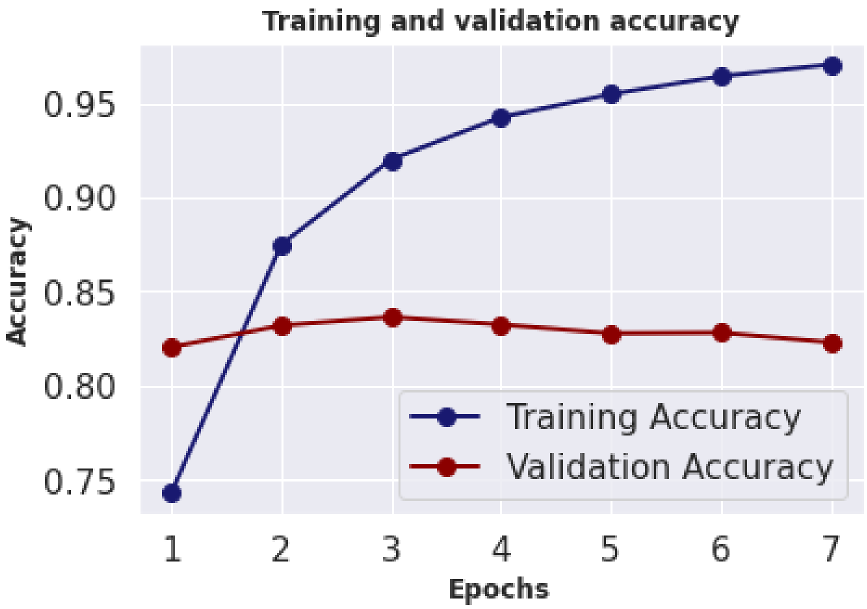


Figure-11: GRU of Training and Validation Accuracy vs Epochs

3.5.2.2. Validation and Training loss

Table-7: GRU of Validation Loss and Training Loss

Epoch No	Validation Loss	Training Loss
Epoch 1	0. 5015	0. 7035
Epoch 2	0. 4804	0. 3537
Epoch 3	0. 5007	0. 2279
Epoch 4	0. 5160	0. 1639
Epoch 5	0. 5953	0. 1262
Epoch 6	0. 6193	0. 1005
Epoch 7	0. 6870	0. 0819

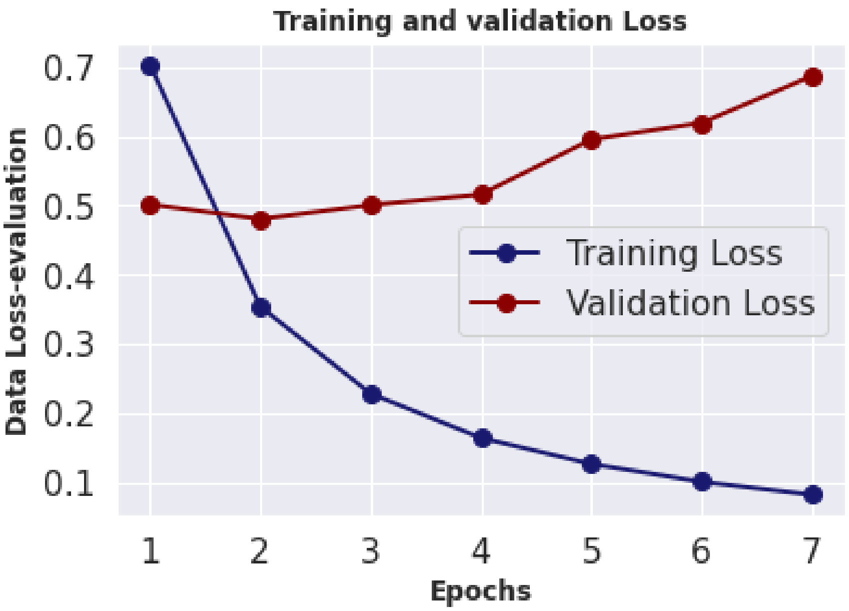
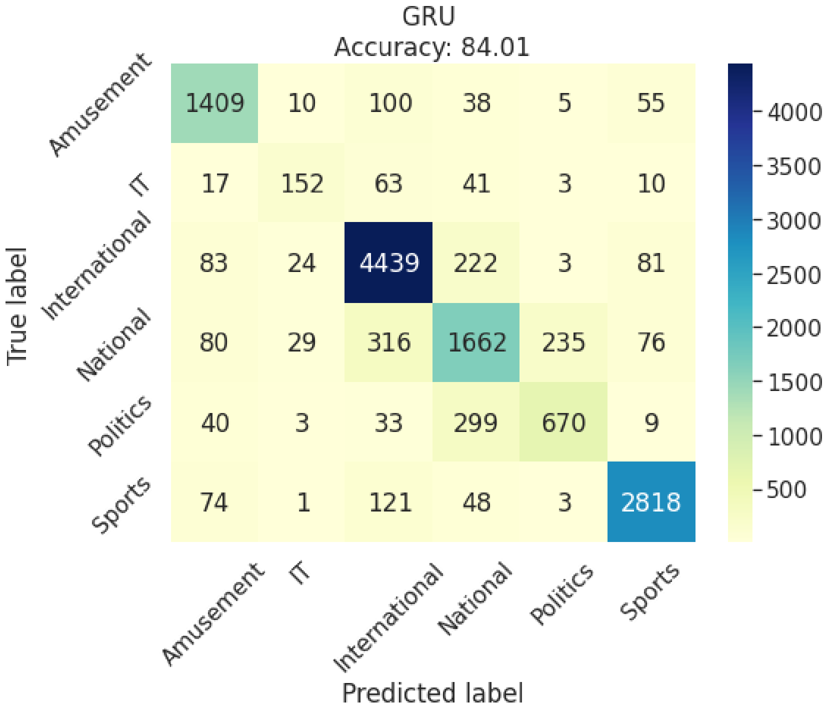


Figure-12: GRU of Training and Validation Loss vs Epochs

3.5.2.3. Confusion matrix

The confusion matrix indicates that the GRU report's accuracy is 84.01%.



	Amusement	IT	International	National	Politics	Sports
Amusement	1409	10	100	38	5	55
IT	17	152	63	41	3	10
International	83	24	4439	222	3	81
National	80	29	316	1662	235	76
Politics	40	3	33	299	670	9
Sports	74	1	121	48	3	2818

Figure-13: GRU Algorithm of Confusion Matrix

3.5.2.4. Classification report

It was concluded that GRU had an 84.01% assurance report after looking over the classes, the classification report (which includes precision, recall, f 1 score, and support), and the news headlines for national, international, sports, politics, entertainment, and IT.

Table-8: GRU Algorithm of Classification Report

News Type	Precision	Recall	F1_Score	Support
Amusement	82.74	87.14	84.88	1617.000000
IT	69.41	53.15	60.20	286.000000
International	87.52	91.49	89.46	4852.000000
National	71.95	69.31	70.60	2398.000000
Politics	72.91	63.57	67.92	1054.000000
Sports	92.42	91.94	92.18	3065.000000
Accuracy	84.01	84.01	84.01	0.840115
Macro Avg	79.49	76.10	77.54	13272.000000
Weighted Avg	83.71	84.01	83.78	13272.000000

The following metrics are obtained for all news categories: accuracy, recall, F1_Score, support, and entertainment (IT, international, national, politics, sports) as can be seen in the categorization report. Accuracy, Macro Average, and Weighted Average values have also been determined. The GRU algorithm's accuracy is 84.01.

3.5.3. Machine Learning (ML)

The dataset should first be distributed or prepared. The collection has 136811 headlines in total. There are six categories in which the 136811 headlines are categorised: foreign, national, athletic, entertainment, political, and IT. The category with the highest number of headlines is international, followed by sports, national, and lastly, IT-related stories. The data set distribution graphs of BiLSTM, GRU, and ML algorithms demonstrate how similar their respective data sets are distributed. Here is a data distribution graph:

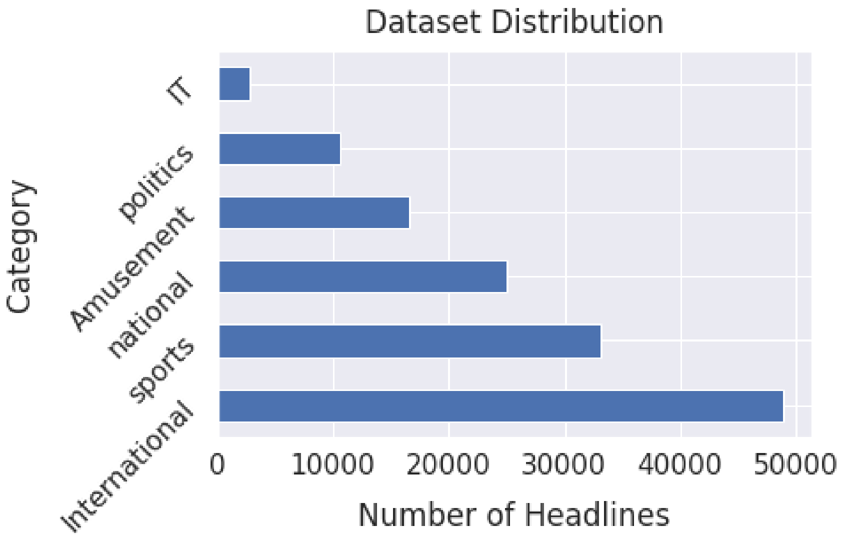


Figure-14: Distribution of Data for ML

Following data preparation, 4098 headlines were cleaned out of the dataset, bringing the total to 132713. Next, by examining the complete dataset, every distinct category is examined once again among all of the headlines, and the quantity of headlines, words, unique words, and most frequently occurring terms of each particular category are retrieved. Thus, the most common word (307354), unique word number (28710), headline number (47885), and word number (307354) of the international class have been retrieved through analysis. Through examination of the sports class, the most common word (30831), word number (152852), unique word number (18581), and headline number (30831) have been identified. The quantity of headlines, words, unique words, and most common terms related to IT, politics, entertainment, and the country have also been extracted. The data collection yielded 57490 unique words in total. The data statistics graphs of BiLSTM, GRU, and ML algorithms demonstrate a striking similarity in their respective data graphs. The dataset analysis graph is displayed below:

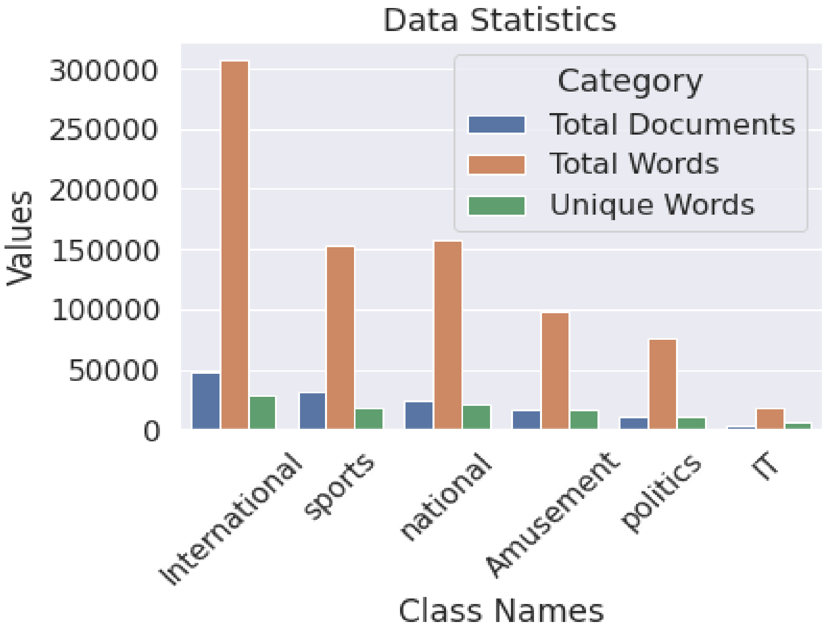


Figure-15: Statistics of All Category Data for ML

Following the visualisation of the dataset, measurements were made of the headlines' length and frequency. Of them, the title's greatest length is 21, its minimum length is 3, and its average length is 6. The Length-Frequency Distribution graphs of BiLSTM, GRU, and ML demonstrate how close these three algorithms' respective graphs are to one another. The graph of the length frequency distribution is shown below:

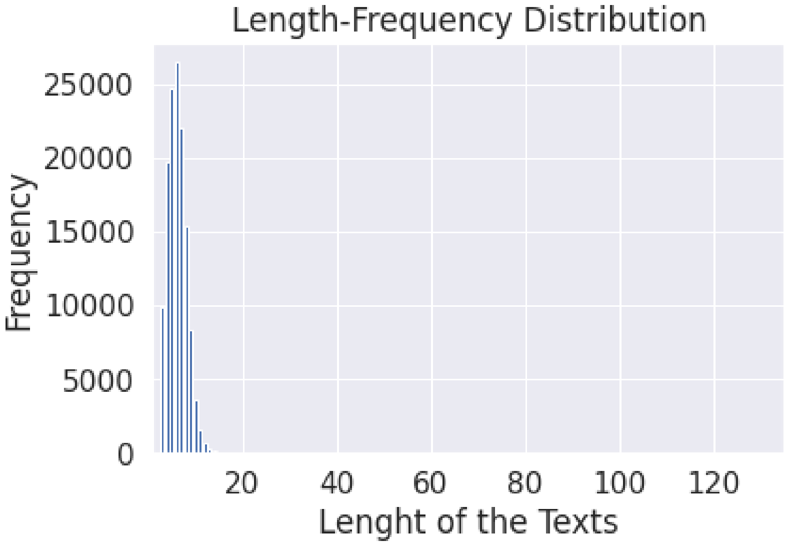


Figure-16: ML of Data Distribution Length Frequency

3.5.3.1. Logistic Regression

Supervised learning is what logistic regression is. Another name for logistic regression is a statistical analysis technique. The likelihood of essentially binary (0/1, yes/no) occurrences is predicted using this technique. In other words, it makes predictions by figuring out the likelihood of statistics and binary results, such as 0/1 or yes/no. The connection between the data, one or more nominal, ordinal, interval, or ratio-level independent variables, and dependent binary variables is typically described using logistic regression models. Big dataset classification is another use for logistic regression. In general, there are three different kinds of logistic regression models. As an illustration, consider binary, multinomial, and ordinal logistic regression.

The training data's accuracy score is 0.6645850993689366.

Table-9: Logistic Regression Algorithm of Classification Report

News Type	Precision	Recall	F1_Score	Support
Amusement	0.65	0.57	0.61	3214
IT	0.63	0.24	0.34	559
International	0.66	0.79	0.72	9577
National	0.56	0.46	0.50	4912
Politics	0.59	0.35	0.44	2115
Sports	0.67	0.74	0.70	6166
Accuracy			0.64	26543
Macro Avg	0.62	0.52	0.55	26543
Weighted Avg	0.64	0.64	0.63	26543

The following metrics are obtained for all news categories: accuracy, recall, F1_Score, support, and entertainment (IT, international, national, politics, sports) as can be seen in the categorization report. Accuracy, Macro Average, and Weighted Average values have also been determined. The approach for logistic regression has an accuracy of 0.64.

3.5.3.2. Multinomial Naive Bayes

In essence, the MNB Algorithm is a Bayesian learning technique. It is really well-liked in NLP. The primary instrument for deciphering textual input and resolving several class difficulties is this approach. The text-based classifier idea forms the foundation of this approach. Based on the Skit-Learn package, naïve bayes may be broadly classified into three categories. specifically Bernoulli, Polynomial, and Gaussian. The Naive Bayes approach relies on the following formula as its foundation: P(A|B) = P(A) * P(B|A)/P(B).

Table-10: Multinomial Naive Bayes of Classification Report

News Type	Precision	Recall	F1_Score	Support
Amusement	0.76	0.43	0.55	3214
IT	0.73	0.03	0.06	559
International	0.57	0.88	0.69	9577
National	0.54	0.42	0.47	4912
Politics	0.61	0.16	0.25	2115
Sports	0.70	0.64	0.67	6166
Accuracy			0.61	26543
Macro Avg	0.65	0.43	0.45	26543
Weighted Avg	0.63	0.61	0.58	26543

The following metrics are obtained for all news categories: accuracy, recall, F1_Score, support, and entertainment (IT, international, national, politics, sports) as can be seen in the categorization report. Accuracy, Macro Average, and Weighted Average values have also been determined. The Multinomial Naive Bayes method has an accuracy of 0.61.

3.5.3.3. Random Forest Classifier

Decision trees make form the supervised learning technique known as Random Forest. The primary applications of this technique are in the areas of regression and classification. By selecting several sample types, the algorithm builds a decision tree by averaging the regression and classifying the samples with the greatest vote. Furthermore, averages are used in dataset analysis to increase prediction accuracy.

Table-11: Random Forest Classifier of Classification Report

News Type	Precision	Recall	F1_Score	Support
Amusement	0.70	0.54	0.61	3214
IT	0.62	0.21	0.32	559
International	0.65	0.79	0.72	9577
National	0.59	0.46	0.51	4912
Politics	0.62	0.43	0.51	2115
Sports	0.67	0.75	0.70	6166
Accuracy			0.65	26543
Macro Avg	0.64	0.53	0.56	26543
Weighted Avg	0.65	0.65	0.64	26543

The following metrics are obtained for all news categories: accuracy, recall, F1_Score, support, and entertainment (IT, international, national, politics, sports) as can be seen in the categorization report. Accuracy, Macro Average, and Weighted Average values have also been determined. The Random Forest Classifier has an accuracy of 0.65.

3.5.3.4. Support Vector Machine (SVM)

Applications for SVM include handwriting recognition, web pages, email categorization, face identification, intrusion detection, and gene classification. One reason SVM is used in machine learning is for this reason. It is capable of handling regression and classification on both linear and non-linear data.

A supervised learning technique is Support Vector Machine (SVM). The majority of issues involving classification and regression are resolved using the SVM algorithm. But sophisticated algorithms perform best in categorization situations. It may be applied to both linear and non-linear problem types. The SVM approach aims to find a hyperplane in an n-dimensional space that clearly classifies all of the data points into a single class. Handwriting recognition, face identification, intrusion detection, email classification, gene classification, and web page classification are among the issues that this approach is utilised to resolve.

Table-12: SVM of Classification Report

News Type	Precision	Recall	F1_Score	Support
Amusement	0.70	0.54	0.61	3214
IT	0.62	0.21	0.32	559
International	0.65	0.79	0.72	9577
National	0.59	0.46	0.51	4912
Politics	0.62	0.43	0.51	2115
Sports	0.67	0.75	0.70	6166
Accuracy			0.65	26543
Macro Avg	0.64	0.53	0.56	26543
Weighted Avg	0.65	0.65	0.64	26543

The following metrics are obtained for all news categories: accuracy, recall, F1_Score, support, and entertainment (IT, international, national, politics, sports) as can be seen in the categorization report. Accuracy, Macro Average, and Weighted Average values have also been determined. The SVM algorithm has an accuracy of 0.65.

3.5.4. Traditional Approach Uni-gram

In its most basic form, a language model infers each word on its own, regardless of any training context. That is, every word in a phrase, no matter how many there are, is a unigram on its own. For example, in the sentence "I eat rice," the words "I," "eat," and "rice" are all unigrams. A unigram language model is one such type of model.

$$P_{uni}(t_1\ t_2\ t_3\ t_4) = P(t_1)\ P(t_2)\ P(t_3)\ P(t_4)$$

Our dataset is likewise subjected to the Unigram method. The unigram technique was used to break each headline statement up into separate words. Using the unigram technique, the original dataset shape counter and the resampled dataset shape counter are retrieved.

3.5.4.1. Decision Tree Classifier

A decision tree is a non-parametric supervised learning approach that uses a continuous data split. This method addresses both classification and regression problems. In essence, this decision tree is made up of internal nodes, leaf nodes, branches, root nodes, and internal nodes. This method predicts a problem's solution to create a tree. The decision tree therefore functions by creating the tree. Here, the 72% accurate Decision Tree Classifier was used to retrieve the following data: Precision, Recall, F1_Score, and Support.

Table-13: Decision Tree Classifier of Classification report

	Precision	Recall	F1_Score	Support
Accuracy			0.72	87215
Macro Avg	0.71	0.72	0.71	87215
Weighted Avg	0.72	0.72	0.72	87215

3.5.4.2. Gradient Bosting Algorithm

An algorithm for supervised learning is gradient boosting. Iterative functional gradient algorithm is another name for gradient boosting. It uses a method known as the boosting approach to merge decision trees. The gradient boosting approach was used to extract the following data: precision, recall, F1 score, and support. The algorithm's accuracy was 72%.

Table-14: Gradient Bosting Algorithm of Classification report

News Type	Precision	Recall	F1_Score	Support
0	0.69	0.78	0.73	15723
1	0.75	0.66	0.70	15829
Accuracy			0.72	31552
Macro Avg	0.72	0.72	0.72	31552
Weighted Avg	0.72	0.72	0.72	31552

The following metrics are obtained for all news categories: accuracy, recall, F1_Score, support, and entertainment (IT, international, national, politics, sports) as can be seen in the categorization report. Accuracy, Macro Average, and Weighted Average values have also been determined. The gradient bosting classifier's accuracy is 0.72.

3.5.4.3. Support Vector Machine (SVM)

One supervised machine learning approach is called SVM. Both regression and classification employ this approach. Regression is the best issue for classification, all things considered. This algorithm's primary goal is to locate a hyperplane in n-dimensional space and clearly identify the input points. Both linear and non-linear problems respond well to it. The support vector machine algorithm was used to extract the following data: precision, recall, F1 score, and support. The algorithm's accuracy was 87%.

Table-15: SVM of Classification report

News Type	Precision	Recall	F1_Score	Support
0	0.81	0.97	0.88	15723

1	0.96	0.77	0.86	15829
Accuracy			0.87	31552
Macro Avg	0.88	0.87	0.87	31552
Weighted Avg	0.88	0.87	0.87	31552

3.5.4.4. Logistic Regression

An approach for supervised classification is called logistic regression. Machine Learning Classification Algorithm is another name for it. which is employed to forecast the likelihood of specific classes based on points in the dependent variable. In a nutshell, this model calculates the total of the input parameters and the logistic of the result. The support vector machine algorithm was used to extract the following data: precision, recall, F1 score, and support. The algorithm's accuracy was 72%.

Table-16: Logistic Regression of Classification report

News Type	Precision	Recall	F1_Score	Support
0	0.72	0.74	0.73	15723
1	0.73	0.71	0.72	15829
Accuracy			0.72	31552
Macro Avg	0.72	0.72	0.72	31552
Weighted Avg	0.72	0.72	0.72	31552

3.5.4.5. Random Forest Classifier

An algorithm for supervised machine learning is called Random Forest. It is mostly applied to regression and classification issues. Saying it "will" or "won't" be played today is one example.

The support vector machine algorithm was used to extract the following data: precision, recall, F1 score, and support. The algorithm's accuracy was 89%.

Table-17: Random Forest Classifier of Classification report

News Type	Precision	Recall	F1_Score	Support
0	0.87	0.91	0.89	15723
12	0.91	0.87	0.89	15829
Accuracy			0.89	31552
Macro Avg	0. 89	0. 89	0. 89	31552
Weighted Avg	0. 89	0. 89	0. 89	31552

3.5.4.6. Summary of classifiers accuracy

RF Classifier, SVM, Gradient Boosting Algorithm, DT Classifier, and LR are just a few of the methods that Unigram utilises. The accuracy of DT is 72%, that of gradient boosting is 72%, that of SVM is 87%, that of logistic regression is 72%, and that of RF is 89%. Odd Forest turns out to have the best accuracy (89%).

Table-18: Accuracy Table of All Classifiers

No	Classifier	Accuracy %
1	Decision Tree Classifier	0.72
2	Gradient Bosting Algorithm	0.72
3	Support Vector Machine	0.87
4	Logistic Regression	0.72
5	Random Forest Classifier	0.89

3.5.4.7. Classification Report

Unigram has a 95% assurance report based on an analysis of the news headlines and classification reports including accuracy, recall, f 1_score, and support.

Table-19: Decision Tree Classifier of Classification report

News Type	Precision	Recall	F1_Score	Support
0	0.95	0.95	0.95	5258
1	0.95	0.95	0.95	5259
Accuracy			0.95	10517
Macro Avg	0. 95	0. 95	0. 95	10517
Weighted Avg	0. 95	0. 95	0. 95	10517

3.5.4.8. Receiver Operating Characteristic (ROC)

The receiver operating characteristic is known as ROC. A ROC curve is a depiction of test sensitivity. In order to anticipate two-sided outcomes, ROC curves evaluate sensitivity vs specificity throughout the whole range. The area under the ROC curve is a further indicator of test performance. For instance, based on a patient's state, medical diagnostic tests can distinguish between "diseased" and "non-diseased" individuals.

It is evident that the Random Forest Classifier has the best accuracy in this instance. Hence, the Random Forest Classifier is used to extract the ROC.

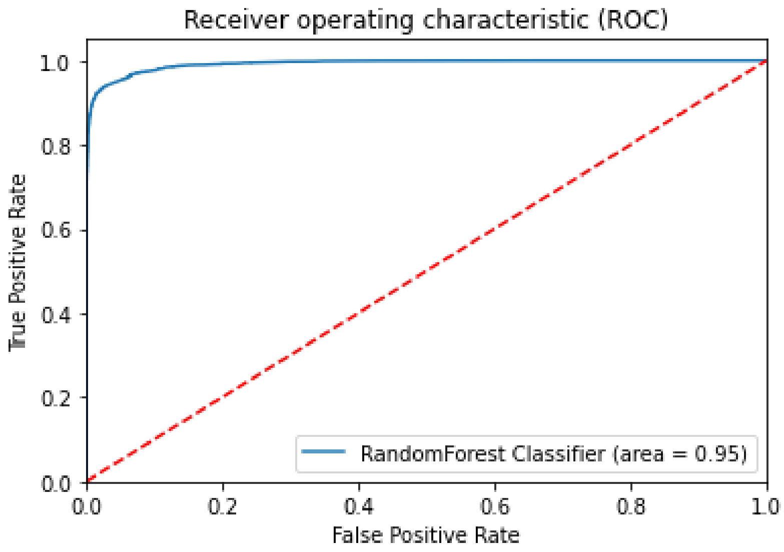


Figure-17: ROC Curve Graph

3.5.4.9. Classification and ROC Analysis

Since the Random Forest Classifier has the maximum accuracy, it has been used to derive the receiver operating characteristic. In order to determine whether anything is incorrect or whether everything is going according to plan, cross-validation is now carried out by extracting the random forest's ROC.

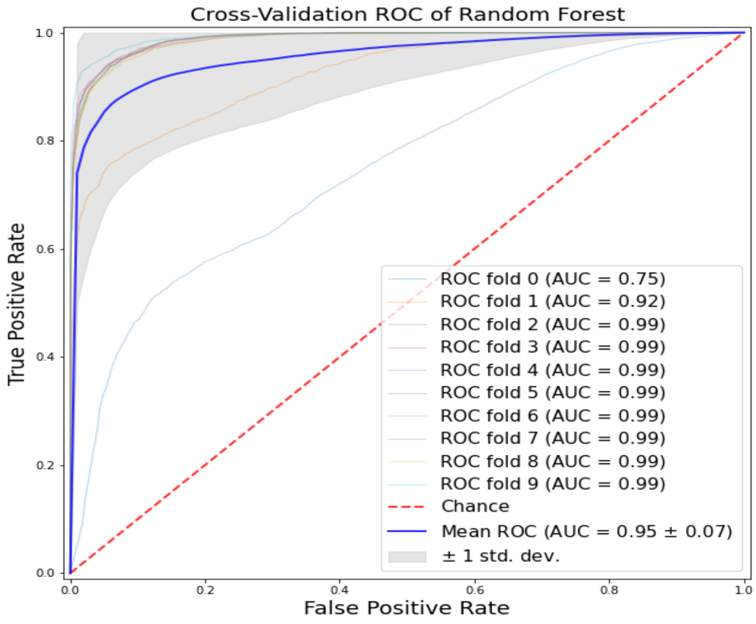


Figure-18: Cross-Validation ROC Curve Graph

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Discussion

Nowadays, Individuals are more drawn to reading news on internet news sites than from traditional newspapers since the internet has become so widely available. Facebook, Twitter, WhatsApp, Telegram, Instagram, and so on are examples of online news portals. Any organization's newspaper that prints news must have a specific quantity of space that cannot be scribbled on, and that space must be paid for. However, writing for websites, blogs, online news portals, and social media platforms like Instagram, Telegram, Facebook, Twitter, and WhatsApp is limitless. So, there is an opportunity to write any news online as big as you want. As a result, the amount of news on social media, blogs, websites, and online news portals is increasing day by day. The work will go a long way if the news from all these sectors can be refined, refined and classified. Data science through artificial intelligence has many algorithms such as machine learning algorithm, transfer learning algorithm, deep learning algorithm etc. which can be used to classify any text. Also, many other tools like ChatGPT, google birds etc. also take its help. However, algorithms and tools are used to classify news. That is, through sentiment analysis, people can understand any kind of news or any type of news by looking at the news headlines. If not, we will miss a lot of important news for the future. There are several areas into which news may be separated, including worldwide, national, sports, entertainment, politics, IT, entertainment, and education. Online news may therefore be digitally identified using artificial intelligence techniques, such as machine learning, transfer learning, deep learning, and other data science techniques, for a variety of significant future uses.

Several deep learning and machine learning methods are used in this research study. BiLSTM, GRU, Uni-gram, and Machine Learning (LR, MNB, RF, SVM) are the methods that are being used.

Table-20: Description Table of All Classifiers

Classifier		Description
Bi-LSTM		To provide an output that is more accurate, BILSTM integrates input sequences with data from the past and the future.
GRU		To sort nodes, GRUs are usually used with machine learning's in-memory clustering and clustering techniques. Update and reset gates are the two gates that are used.
Machine Learning	Logistic Regression	Predicting binary outcomes is a typical use of logistic regression. For instance, true or false, yes or no, and 0 or 1.
	Multinomial Naïve Bayes	In NLP generally, the Multinomial Naive Bayas algorithm is quite well-liked. The quickest and most straightforward method for doing sentiment analysis is this algorithm.
	Random Forest Classifier	In issues relating to regression and classification, the random forest classifier method is frequently employed. To build a better model, it builds a decision tree using the dataset, classes it, and averages it.
	SVM	SVM's job is to use the maximum marginal hyperplane to categorise the dataset's data points in an understandable manner.
Uni-Gram	Decision Tree	The decision tree algorithm creates decision nodes and leaves by continually dividing the dataset based on predetermined criteria.
	Gradient Boosting Algorithm	In contrast to weak prediction models, which are often connected to decision trees, the gradient boosting approach offers an additional updated prediction model.
	SVM	SVM's job is to use the maximum marginal hyperplane to categorise the dataset's data points in an understandable manner.
	Logistic Regression	Predicting binary outcomes is a typical use of logistic regression. For instance, true or false, yes or no, and 0 or 1.
	Random Forest Classifier	In issues relating to regression and classification, the random forest classifier method is frequently employed. To build a better model, it builds a decision tree using the dataset, classes it, and averages it.

4.2. Experimental Results and Analysis

This article uses two deep learning models: GRU and BiLSTM. Among them, BiLSTM demonstrated 83.42% accuracy, while GRU demonstrated 84.01% accuracy. The GRU model seems to have the greatest accuracy when deep learning is used. Some machine learning models—LR, MNB, RF, and SVM, for example—have been combined with declining. Here, the accuracy of LR is 64%, that of MNB is 61%, that of RF is 65%, and that of SVM is 65%. With an accuracy rate of 65%, it is evident that RF and SVM offered the highest accuracy of all the models. Again, the unigram language model is utilised in conjunction with the DT, gradient boosting technique, SVM, LR, and RF classifier algorithm. The DT has an accuracy rate of 72%, the gradient boosting technique has an accuracy rate of 72%, the SVM has an accuracy rate of 87%, the LR has an accuracy rate of 72%, and the RF classifier has an accuracy rate of 89%, according to the results. When applying the unigram approach in this case, the RF classifier produces the highest accuracy.

Table-21: Accuracy, Recall and Precision Score of All Classifiers

Algorithm techniques		Accuracy %	Recall %	Precision %
Bi-LSTM		83.42	83.42	83.42
GRU		84.01	84.01	84.01
ML	Logistic Regression	64	64	64
	Multinomial Naïve Bayes	61	61	63
	Random Forest Classifier	65	65	65
	SVM	65	65	65
Uni-Gram	Decision Tree	72	72	72
	Gradient Bosting Algorithm	72	72	72
	SVM	87	87	88
	Logistic Regression	72	72	72
	Random Forest Classifier	89	89	89

CHAPTER 5
IMPECT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1. Impact on Society

The internet has given modern living a new dimension through virtual entertainment. In rural areas, people no longer read newspapers; instead, they use cellphones to browse and read news websites. Using virtual entertainment platforms such as Instagram, LinkedIn, YouTube, Messenger, Pinterest, Viber, WhatsApp, Facebook, Trambler, Snapchat, Twitter, Google, and others, one may access all that is happening both domestically and internationally. Virtual entertainment timelines and newsfeeds are becoming overflowing with both important and irrelevant news. Social media is used by 70% of internet users globally. In children, the percentage is about 90%. According to a survey, Facebook is used by almost 80% of Bangladeshi online users. The extent of social communication has increased due to the use of the internet compared to earlier times. Individuals exchange information, thoughts, images, videos, and other media through various data and document formats thanks to technological advancement. Since none of these types of information and documents are characterised, it is difficult to find a specific spot while searching for it. This makes it a significant disadvantage. Data categorization is completed to address this issue. That is, it will be very useful for finding news if it is organised into several categories such as Politics, IT, International, National, Sports, and Entertainment, among others. Therefore, if you search for news by creating any kind of keyword, the likelihood of finding new information on that particular type of catchphrase will increase consistently. In this sense, it's important to first categorise the news into different groups before reading any kind of material. It will be much easier for people from different backgrounds to consume virtual pleasure or read news on the internet if it is divided into different classes.

5.2. Impact on Environment

Since ancient times, people have been interested in information. News was read by people from newspapers back in the day. However, people are becoming less interested in reading news articles in newspapers as time passes and innovation advances. These days, social media and the internet provide access to a wide variety of information. We can quickly get any type of information about anything happening anywhere on the earth. We receive online reports about the climate from the Meteorological Division in a timely manner. We may find out all kinds of information on the internet, such as the location of cyclones, floods, and twisters on Earth, as well as the environmental conditions in a certain area of the globe. We have instant access to information on the weather in any location or nation. We do not look at the weather or the environment to determine what job should be feasible in the near future. It is possible to survive on the residue of many significant accidents or catastrophes in this way. As a result, information contributes significantly to the climate.

5.3. Ethical Aspects

Every news outlet has a distinct publication approach. The media can adopt strategies on its own. Typically, the media provides false information to publications on social media. For example, providing false information about a person online in relation to financial, political, income tax, or other matters is dishonest. In any event, on social media, individuals propagate unfavourable information about others in various places. Despite this, there is less false information available online now than there ever was. That suggests that everyone should practise morality and behave decently when using social media or the internet.

5.4. Sustainability

We must separate the data or reports into different classifications in order to arrange our informational gathering into different groups. Thus, reports and news may be identified in different classifications. with the intention that the data can be used going forward and for any worthwhile purpose. For this reason, a lengthy setup is necessary so that different methods may be used to pre-process and divide any type of data into different groups. For this reason, it is important to have an aircraft for long-term planning.

CHAPTER 6
SUMMARY, CONCLUSION, RECOMENDATION AND IMPLEMENTATION FOR FUTURE RESEARCH

6.1. Summary of the Study

These days, more individuals choose to read news articles from internet news sources than from traditional newspapers since the internet has become so widely available. Facebook, Twitter, WhatsApp, Telegram, Instagram, and so on are examples of online news portals. The number of readers on the online news site is growing daily in tandem with the volume of content it contains. It is not enough to get news; the news has to be absolutely refined and classified. Because only refined and classified news can be used. Purifying and classifying news in this modern age is not a difficult task. There are many algorithms of data science through artificial intelligence such as machine learning algorithm, transfer learning algorithm, deep learning algorithm etc. which can be used to classify any text. Besides, many other types of tools also take its help such as ChatGPT, Google Birds etc. However, algorithms and tools are used to categorize news. That instance, by using sentiment analysis, readers may determine the nature of the news simply by glancing at the title. We have divided the news into six categories for this research study. There are six distinct categories: Politics, IT, International, National, Sports, and Entertainment. These categories have been classified using various methods, including BLSTM, GRU, Uni-Gram, as well as traditional Machine Learning approaches such as LR, MNB, RF Classifier, and SVM. Among the deep learning models, GRU achieves an accuracy rate of 84.01%, while BiLSTM achieves 83.42%. In the realm of machine learning, LR has an accuracy of 64%, MNB scores 61%, RF Classifier performs at 65%, and SVM also reaches 65% accuracy. The proliferation of e-news and online news portals, both in Bangladesh and globally, continues to grow steadily. These e-news and online news portals are available on various social sites. That means providing regular breaking and updated news on our favorite sites and the sites that the team spends the most time on. This is because the news of the site is refined and classified, which is very useful and easy to understand for the readers.

6.2. Conclusions

A closer look at current events will reveal that while the amount of e-news is growing daily, not all of it is classified. Many news pieces are becoming useless as a result. People will gain in several ways if the pointless news can eventually be turned into something helpful and practical. In other words, a lot of news will become helpful if it can be categorised. Any text may be classified using one of the various data science methods made possible by artificial intelligence, including the machine learning, transfer learning, deep learning, and others. In addition, a lot of other tools use it as well, such Google Birds and ChatGPT. We have divided the news into six categories for this research study. Politics, IT, International, National, Sports, and Entertainment are the six main areas. Numerous approaches, such as BLSTM, GRU, and Uni-Gram, as well as machine learning techniques including LR, MNB, RF, SVM, have been used to classify these categories. BiLSTM has an accuracy rate of 83.42% among deep learning models, however GRU has a better accuracy rate of 84.01%. In terms of machine learning, the following algorithms do well: Random Forest Classifier, which achieves 65% accuracy, Multinomial Naive Bayes, which scores 61%, and Support Vector Machine, which achieves 65% accuracy.

6.3. Implication for Further Study

This paper's major goal was to demonstrate how to use artificial intelligence—specifically, machine learning and deep learning techniques in data science—to categorise a huge number of news items into several categories.

- If artificial intelligence data science can be used to categorize news of all types i.e., social media, online platforms in the future.
- Since our work is NLP related work, apart from the applied algorithm, there are many other types of algorithms that may get better results if applied, such as machine learning algorithm, deep learning algorithm, transfer learning, etc.
- Our paper shows that 4 algorithms for machine learning, 5 algorithms for Uni-Gram and 2 algorithms for deep learning are used to classify the data. However, more numerical algorithms will be used for this type of work in the future.
- This article has six categories for news or data., which will be divided into more categories in the future.

- As far as accuracy is currently found in this paper, these algorithms will become more accurate in the future, properly predicting news stories.

REFERENCE

[1] Barua, Adrita, et al. “Multi-Class Sports News Categorization Using Machine Learning Techniques: Resource Creation and Evaluation.” *Procedia Computer Science*, vol. 193, 2021, pp. 112–121., https://doi.org/10.1016/j.procs.2021.11.002.

[2] Islam, Md. Majedul, et al. “Performance Measurement of Multiple Supervised Learning Algorithms for Bengali News Headline Sentiment Classification.” *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, https://doi.org/10.1109/smart46866.2019.9117477.

[3] Mohiuddin, Ettilla, and Abdul Matin. “Multilevel Categorization of Bengali News Headlines Using Bidirectional Gated Recurrent Unit.” *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, 2021, https://doi.org/10.1109/acmi53878.2021.9528006.

[4] Bhuiyan, Md. Rafiuzzaman, et al. “An Approach for Bengali News Headline Classification Using LSTM.” *Advances in Intelligent Systems and Computing*, 2021, pp. 299–308., https://doi.org/10.1007/978-981-15-9927-9_30.

[5] Galal Elsayed, Hoda Ahmed, et al. “A Two-Level Deep Learning Approach for Emotion Recognition in Arabic News Headlines.” *International Journal of Computers and Applications*, vol. 44, no. 7, 2020, pp. 604–613., https://doi.org/10.1080/1206212x.2020.1851501.

[6] Tudu, Ronald, et al. “Performance Analysis of Supervised Machine Learning Approaches for Bengali Text Categorization.” *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, 2018, https://doi.org/10.1109/apwconcse.2018.00043.

[7] Jahara, Fatima, et al. “Automatic Categorization of News Articles and Headlines Using Multi-Layer Perceptron.” *Intelligent Computing & Optimization*, 2022, pp. 155–166., https://doi.org/10.1007/978-3-030-93247-3_16.

[8] Bogery, Raghad, et al. “Automatic Semantic Categorization of News Headlines Using Ensemble Machine Learning: A Comparative Study.” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 11, 2019, https://doi.org/10.14569/ijacsa.2019.0101190.

[9] Ahmed Foysal, Md. Ferdouse, et al. “Bengali News Classification Using Long Short-Term Memory.” *Advances in Intelligent Systems and Computing*, 2021, pp. 329–338., https://doi.org/10.1007/978-981-33-4367-2_32.

[10] Ke Yahan, R. Qu, Lu Xiaoxia. January 2022. "Classification Of Fake News Headline Based On Neural Networks".

[11] Singh, Prakash Kumar, and Sanchita Paul. “Deep Learning Approach for Negation Handling in Sentiment Analysis.” *IEEE Access*, vol. 9, 2021, pp. 102579–102592., https://doi.org/10.1109/access.2021.3095412.

[12] Rabib, Mohammad, et al. “Different Machine Learning Based Approaches of Baseline and Deep Learning Models for Bengali News Categorization.” *International Journal of Computer Applications*, vol. 176, no. 18, 2020, pp. 10–16., https://doi.org/10.5120/ijca2020920107.

[13] Usmani, Shazia, and Jawwad A. Shamsi. “News Headlines Categorization Scheme for Unlabelled Data.” *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, 2020, https://doi.org/10.1109/icetst49965.2020.9080726.

[14] Khushbu, Sharun Akter, et al. “Neural Network Based Bengali News Headline Multi Classification System: Selection of Features Describes Comparative Performance.” *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, https://doi.org/10.1109/icccnt49239.2020.9225611.

[15] Ruichao W., John D., Joe C. May 2014. “Machine Learning Approach To Augmenting News Headline Generation.”

[16] Hossain, Syeda Sumbul, et al. “Context-Based News Headlines Analysis: A Comparative Study of Machine Learning and Deep Learning Algorithms.” *Vietnam Journal of Computer Science*, vol. 08, no. 04, 2021, pp. 513–527., https://doi.org/10.1142/s2196888822500014.

[17] Santos, António Paulo, et al. “Sentiment Classification of Portuguese News Headlines.” *International Journal of Software Engineering and Its Applications*, vol. 9, no. 9, 2015, pp. 9–18., https://doi.org/10.14257/ijseia.2015.9.9.02.

[18] Saha, Uchchhwas, et al. “Sentiment Classification in Bengali News Comments Using a Hybrid Approach with Glove.” *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2022, https://doi.org/10.1109/icoei53556.2022.9777096.

[19] Jubaer, A.N.M., et al. “Bangla Toxic Comment Classification (Machine Learning and Deep Learning Approach).” *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, https://doi.org/10.1109/smart46866.2019.9117286.

[20] Salehin, Mushfiqus, et al. “Generating Bengali News Headlines: An Attentive Approach with Sequence-to-Sequence Networks.” *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, https://doi.org/10.1109/smart46866.2019.9117554.