

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

BY

**Labony Akter
ID: 191-15-12534
AND**

**Md Shahriar Zaman
ID: 191-15-12570**

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

Supervised By

Dr. Moushumi Zaman Bonny
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Sazzadur Ahamed
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

24 JANUARY 2023

APPROVAL

This Project/internship titled “**A Machine Learning and Deep Learning Approach for Bengali News Headline Categorization**”, submitted by Labony Akter, ID No: 191-15-12534 & Md Shahriar Zaman, ID No: 191-15-12570 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 *24 January 2023*.

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

Abdus Sattar

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

Fatema Tuj Johra

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



External Examiner

Dr. Dewan Md Farid

Professor


Department of Computer Science and Engineering

United International University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of Dr. Moushumi Zaman Bonny, Assistant Professor, 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:

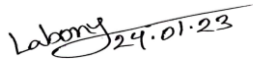
 23.01.2023

Dr. Moushumi Zaman Bonny
Assistant Professor
Department of CSE
Daffodil International University


Co-Supervised by:

Md. Sazzadur Ahamed
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

 24.01.23

Labony Akter
ID: 191-15-12534
Department of CSE
Daffodil International University

 23.01.2023

Md Shahriar Zaman
ID: 191-15-12570
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 successfully.

We really grateful and wish our profound our indebtedness to Supervisor **Dr. Moushumi Zaman Bonny, Assistant Professor**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of Machine Learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Moushumi Zaman Bonny, Assistant Professor, Md. Sazzadur Ahamed, Assistant Professor & Dr. Touhid Bhuiyan, Professor & Head**, Department of CSE, for their 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

Internet is a thing through which a huge amount of information and data is available. As the amount of online news is increasing drastically due to the availability of internet in all parts of the world, people are also interested in reading news from online news portals due to the availability of internet. The online news portals are- Facebook, Twitter, WhatsApp, Telegram, Instagram, Blog etc. As the amount of news is increasing in the news portals, the number of readers is also increasing. As the amount of digital data is increasing in the world, the need for data classification for that digital data is also increasing. There are several methods of data classification, such as machine learning, deep learning, etc., as well as other data mining algorithms. Data is categorized using these algorithms, so that people read the news headlines before reading the news to easily understand the main theme of the news. Natural language processing approaches are used to classify data in any language for such problems. In this research paper, Bengali news has been classified into 7 categories using machine learning and deep learning. The categories are International, National, Sports, Amusement, Politics and IT. BiLSTM, GRU, Uni-gram, Machine Learning (Logistics regression, Multinational naïve bayes, Random Forest classifier, Support vector machine) have been used to classify these categories. While the accuracy of BiLSTM is 83.42%, the accuracy of GRU is 80.01%. Among machine learning, the accuracy of Logistics regression is 64%, the accuracy of Multinational naïve bayes is 61%, the accuracy of Random Forest classifier is 65% and the accuracy of Support vector machine is 65%.

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 | 2 |
| 1.3 Relational of the Study | 3 |
| 1.4 Research Questions | 3 |
| 1.5 Expected Outcome | 3 |
| 1.6 Report Layout | 3 |
| | |
| CHAPTER 2: BACKGROUND | 4-9 |
| 2.1 Terminology | 4 |
| 2.2 Related work | 4-6 |
| 2.3 Comparative Analysis and Summary | 6-9 |
| 2.4 Scope of the Problem | 9 |
| 2.5 Challenges | 9 |

| | |
|---|--------------|
| CHAPTER 3: RESEARCH METHODOLOGY | 10-35 |
| 3.1 Dataset Description | 10 |
| 3.2 Dataset Pre-Processing | 10 |
| 3.3 Statical Analysis | 11 |
| 3.4 Design Approach | 11 |
| 3.5 Proposed Methodology | 12 |
| 3.5.1 Bidirectional Long Short-Term Memory (BiLSTM) | 12 -14 |
| 3.5.1.1 Validation and Training accuracy | 14 |
| 3.5.1.2 Validation and Training loss | 15 |
| 3.5.1.3 Confusion matrix | 17 |
| 3.5.1.4 Classification report | 17 |
| 3.5.2 Gated recurrent units (GRU) | 18 |
| 3.5.2.1 Validation and Training accuracy | 20 -21 |
| 3.5.2.2 Validation and Training loss | 22 |
| 3.5.2.3 Confusion matrix | 23 |
| 3.5.2.4 Classification report | 23 24 |
| 3.5.3 Machine Learning (ML) | 24 -26 |
| 3.5.3.1 Logistic Regression | 26-27 |
| 3.5.3.2 Multinomial Naive Bayes | 27-28 |
| 3.5.3.3 Random Forest Classifier | 28-29 |
| 3.5.3.4 Support Vector Machine (SVM) | 29-30 |
| 3.5.4 Traditional Approach Uni-gram | 30-31 |

| | | |
|---|---|--------------|
| 3.5.4.1 | Decision Tree Classifier | 31 |
| 3.5.4.2 | Gradient Bosting Algorithm | 31 -32 |
| 3.5.4.3 | Support Vector Machine (SVM) | 32 |
| 3.5.4.4 | Logistic Regression | 32 -33 |
| 3.5.4.5 | Random Forest Classifier | 33 |
| 3.5.4.6 | Summary of classifiers accuracy | 33-34 |
| 3.5.4.7 | Classification Report | 354 |
| 3.5.4.8 | Receiver Operating Characteristic (ROC) | 35 |
| 3.5.4.9 | Classification and ROC Analysis | 35 |
| CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION | | 36-38 |
| 4.1 | Discussion | 36-37 |
| 4.2 | Experimental Results and Analysis | 38 |
| CHAPTER 6: SUMMARY, CONCLUSION, RECOMENDATION AND IMPLEMENTATION FOR FUTURE RESEARCH | | 41-42 |
| 6.1 | Summary of the Study | 41 |
| 6.2 | Conclusions | 41 |
| 6.3 | Implication for Further Study | 42 |
| REFERENCE | | 43-44 |
| APPENDICES | | 44 |

LIST OF FIGURES

| FIGURES | PAGE NO |
|---|----------------|
| Figure 1: Architecture of Working Process | 12 |
| Figure 2: Dataset Distribution for Bi-LSTM | 13 |
| Figure 3: Dataset Statistics for Bi-LSTM | 13 |
| Figure 4: Length-frequency distribution for Bi-LSTM | 14 |
| Figure 5: Epochs vs validation and training accuracy plot for Bi-LSTM | 15 |
| Figure 6: Epochs vs validation and training loss plot for Bi-LSTM | 16 |
| Figure 7: Confusion Matrix of Bi-LSTM Algorithm | 17 |
| Figure 8: Dataset Distribution for GRU | 19 |
| Figure 9: Dataset Statistics for GRU | 19 |
| Figure 10: Length-Frequency Distribution for GRU | 20 |
| Figure 11: Epochs vs validation and training accuracy plot for GRU | 21 |
| Figure 12: Epochs vs validation and training loss plot for GRU | 22 |
| Figure 13: Confusion matrix for GRU | 23 |
| Figure 14: Dataset Distribution for ML | 25 |
| Figure 15: Data Statistics for ML | 25 |
| Figure 16: Length-Frequency Distribution for ML | 26 |
| Figure 17: Receiver Operating Characteristic Graph | 35 |
| Figure 18: Check Validation Receiver Operating Characteristic | 35 |

LIST OF TABLES

| TABLES | PAGE NO |
|--|----------------|
| Table 1: Summary of Related Works | 9 |
| Table 2: Category name or Description | 10 |
| Table 3: Validation and Training accuracy table for Bi-LSTM | 15 |
| Table 4: Validation and Training loss for Bi-LSTM | 16 |
| Table 5: Classification report of Bi-LSTM Algorithm | 18 |
| Table 6: Validation and Training accuracy table for GRU | 21 |
| Table 7: Validation and Training loss table for GRU | 22 |
| Table 8: Classification report for GRU | 24 |
| Table 9: Classification report for Logistic Regression | 27 |
| Table 10: Classification report for Multinomial Naive Bayes | 28 |
| Table 11: Classification report for Random Forest Classifier | 29 |
| Table 12: Classification report for SVM | 30 |
| Table 13: Classification report for Decision Tree Classifier | 31 |
| Table 14: Classification report for Gradient Bosting Algorithm | 31 |
| Table 15: Classification report for SVM | 32 |
| Table 16: Classification report for Logistic Regression | 33 |
| Table 17: Classification report for Random Forest Classifier | 33 |
| Table 18: Summary of Classifiers accuracy | 34 |
| Table 19: Classification report for Decision Tree | 34 |
| Table 20: Classifiers Description | 37 |
| Table 21: Classifiers accuracy, recall and precision | 38 |

CHAPTER 1

INTRODUCTION

1.1 Introduction

The Internet is the main source of information and an integral part of people's lives. Online news sources are growing at a great rate in the world and due to the availability of internet, people are interested in reading daily news portals. Thousands of portals are providing hourly news updates and headlines in Bengali. In the digital age now most of the people read news through internet instead of reading news in newspapers. Nowadays online portals, Twitter, Facebook, blogs etc. are being used in a large amount through applications, due to which the use of internet is expanding and a large amount of information is available on websites. That is, the people of the world have become heavily dependent on Internet news. Also, in combination with high-speed Internet and handheld multimedia devices, users are creating and accessing large amounts of information every day. According to a report, there are about 80.83 million internet users at the end of January 2018, out of which about 30 million are social media users in Bangladesh. In recent years, the amount of online news production and access has increased day by day, with a focus on the Internet as the paper boom has been reduced. Many news organizations appear to be creating and uploading news online instead of releasing it. The news that is available from the Internet and the news that is published on the Internet is called e-news. This news readership is increasing day by day as a result of internet user scholarship. Due to which a large number of different news are being registered in the database of the website. News is a very important domain in developing countries like India, Bangladesh, Pakistan as news spreads knowledge as well as increases the level of public awareness about the news of neighboring countries.

As a citizen of Bangladesh, Bengali language is our mother tongue and in 1952 many language martyrs gave their lives for the mother tongue Bengali. Bengali language is popular in several parts of India as well as Bangladesh. Currently, 228 million people speak Bengali in the world every day and 37 million foreigners speak Bengali. A count of speakers around the world shows that Bengali is the seventh most spoken language. The last few decades have shown that both positive and negative effects of news reach the public very quickly. Many studies show that the negative impact of news has a negative effect on readers. Most of the videos seem to focus on releasing bad news rather than good news. These events or effects result in both physical and mental changes in individuals. Such psychological effects, such as negative thoughts or nervousness, depression, loss of concentration, stress, initiative, fear etc.

Effective information retrieval is a principle of information technology. A higher generalization of text content is news headlines. The internet caters to various types of news such as sports, computers, Hollywood, Bollywood, music, politics and social sciences. Users can locate and view any type of news from the internet. Through news headlines, users can easily search and view news according to their needs.

1.2 Motivation

Automatic text classification is an emerging topic. As the amount of digital data in the world is increasing at a massive rate, the need for text classification for this growing digital data is also increasing. Machine learning methods are used to classify data along with other data mining algorithms to classify text. Text classification is also used in some applications such as content tagging, spam filtering and business intelligence. Bengali language websites have huge amount of data from which it is quite difficult to find data. On the other hand, if you want to post on a forum, categorizing readers and hash tagging keywords becomes a very important task. Text categorization has not been done before in any application platform which is a big problem, thus status categorization has become very important for Bengali language. Again, the current need of text mining is increasing day by day due to which Bangladeshi and Indian researchers are focusing on creating applications using text mining. There are numerous works on text mining as well as sentiment analysis in Bengali which show results with good elimination. Most of the people read the news headlines before reading the news so that they can easily understand the sentiment of the news and what is wanted in the news. Any type of language problem in classification is said to be solved by NLP. As these problems express human language problem concepts and attempt to output them, machine learning algorithms are the most useful algorithms for understanding NLP problems. There are several categories of machine learning such as supervised, unsupervised and semi-supervised learning. Supervised, unsupervised and semi-supervised learning in supervised learning. Supervised learning needs to provide inputs and outputs as well as leveled data. Unsupervised learning requires providing input and output as well as levelfree data. Semi-supervised learning is basically made up of supervised and unsupervised learning where labeled and unleveled data are combined. However, classifying so much data or news is time consuming, challenging and difficult. Machine learning algorithms are used to overcome these time consuming challenging and difficult algorithms. Text classification has been widely applied over the last few years with the development of machine learning. News headline classification is a type of text classification that can generally be divided into three parts, namely feature extraction, classifier extraction and evaluation. Another aspect of e-news in Bangladesh is that the readers prefer those sites which give regular breaking news and update news. The five popular sites are Pratham Alo, Bangladesh Pratidin, Nayadigant, Jugantar, Samakal etc. Analyzing from Google trend data, it can be seen that the number of online readers of “Daily News” is decreasing day by day.

1.3 Relational of the Study

Our research paper is divided into 6 parts to find out which category a news item falls into. Our research has divided the news into 6 categories like International, National, Sports, Entertainment, Politics and IT. We found several such works; They divided the news into different categories like us and extracted the security using different algorithms like DL & ML.

1.4 Research Questions

We may have many types of questions in this study. For example-

1. What is the purpose of this research?
2. Why is this research being done?
3. Can we benefit from this research?
4. In which direction can we benefit from this research?
5. What are the results of this study?
6. What is the outcome of this study?

1.5 Expected Outcome

News is a type of information, through which futures can be used and any kind of work can be done. Since our research is to classify news headlines, once our research is complete, we can test with any type of data which category the news falls into. Besides, a lot of data can be saved from being lost, as there is a lot of unclassified data which is of no use, unclassified data can be classified and converted into usable data.

1.6 Report Layout

However, due to improper use of data obtained from online, social media, i.e., not classifying, sorting and analyzing the data in a proper way, the news portals are failing to use a lot of potential data. If the data can be classified automatically using machine learning (ML), deep learning (DL) and NLP in a faster, more agile, cheaper and reliable way, this classification can be a huge potential solution. The remainder of the paper is divided into four sections: Section 2- background, Section 3- methodology, Section 4- experimental result and discussion, Section 5- impact on society, environment and sustainability. Finally, Section 6 will conclude the paper.

CHAPTER 2

BACKGROUND

2.1 Terminology

As the world is constantly expanding, events are happening everywhere. And the said incidents are spreading in online social media in no time thanks to internet. In this way, the amount of news is increasing as time goes by. But not to categorize the news, that is, not to define which news falls under which category. Due to this, a lot of online data is lost which cannot be used in future. So, to make all the data or documents usable, the data is classified using different algorithms of machine learning. So that today's data becomes usable for tomorrow.

2.2 Related Works

Adrita Barua et al. [1] propose a research paper classifies NLP related language detection through machine learning. Here 6 machine learning algorithms such as Logistic Regression (LR), Support Vector Classifier (SVC), Decision Tree (DT), Multinomial Naive Bayes (MNB), Random Forest (RF), Term Frequency Inverse Document Frequency (TF-IDF) are discussed in 4 categories (cricket, football, tennis and athletics). The highest accuracy is 97.60% obtained with the SVC algorithm. Md. Majedul Islam et al. [2] propose a research paper, which is detect Bengali word title sentiment using supervised algorithm. 5 Machine Learning Algorithms such as KNN, Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF) are discussed here. Highest accuracy is 75% obtained with SVM algorithm. Ettilla Mohiuddin Eumi et al. [3] propose a research paper focuses on classifying Bengali news headlines. Here the accuracy is calculated using some algorithms like LibSVM with stop words, LibSVM without stop words, Scikit Learn Library, SVM, NB, LR, RF, Sequential deep learning, BiGRU Model. Headlines are divided into 6 different categories viz. The highest accuracy was 84% obtained using the BiGRU model algorithm. Md. Rafiuzzaman Bhuiyan et al. [4] propose a research paper focuses on categorizing news headlines. Here LSTM algorithm model is created and accuracy is calculated with 4580 trained data. Headlines are divided into 4 different categories like Rational, International, Science, Sports. Accuracy has come from 91.22% using LSTM model. Hoda Ahmed Galal Elsayed et al. [5] propose a research paper, the psychological influence of news headlines on readers is extracted through machine learning algorithm. Emotions are divided into 7 categories like anger, disgust, fear, happiness, neutral, sadness, and surprise. 6 machine learning classifiers like zeroR, KNN, CNN, decision trees, naïve Bayes, random forest and SVM have been used for this. Accuracy is achieved by using multilevel CNN 89.3%. Ronald Tudu et al. [6] propose a research paper classifies headlines with the help of machine learning. Headlines are divided into

10 different categories such as Crime (CR) Economics (EC) International (IN) Sports (SP) Accident (AC) Environment (EV) Science and Technology (ST) Entertainment (EN) Politics (PO) Education (ED). Here the accuracy is measured using SVM, MNB, SGD, LR classifier. Accuracy is 87.5% achieved by using SVM algorithm. Fatema Jahara et al. [7] propose a research paper, Deep Learning Based Framework Multilayer Perception (MLP) classifier is used to classify newspaper headlines. Headlines are divided into 4 different categories such as accident, crime, entertainment, sports. Using the MLP algorithm, the highest accuracy was found for 98.18% (news articles) and 94.53% (news headlines). Raghad Bogery et al. [8] propose a research paper discusses some machine learning algorithms including NLP for classifying a large number of news headlines. 5 machine learning classifiers like KNN, SVM, NB, MNB and Gradient boosting have been used for this. The best performance was achieved using the Multinomial Naïve Bayes algorithm, with accuracy 90.12% and recall 90%. Headlines are divided into 3 different categories like travel, style & beauty, parenting. Md. Ferdouse Ahmed Foysal et al. [9] propose a research paper using LSTM algorithm of machine learning to classify news headlines achieved good accuracy. Accuracy is brought to 84% using LSTM algorithm. Headlines are divided into 5 different categories such as entertainment, national, sports, city state news. Ke Yahan et al. [10] propose a research paper takes a dataset containing 18 years of news and classifies it using machine learning NLP. 4 machine learning classifiers like Decision tree, Random Forest, SVC, NN have been used for this. The best performance was achieved using the NN algorithm, which has an accuracy of 0.8622. Prakash Kumar Sing et al. [11] propose a research paper which deep neural network model based on LSTM, SVM, HMM, CRF is used in this research paper. BiLSTM provided better accuracy 93.34% than other models. Mohammad Rabib Hossain et al. [12] propose a research paper different methods are used to classify news using machine learning baseline like SVM, Naive-Bayes, Random Forrest, Logistic Regression and deep learning model like BiLSTM, CNN of all these models. CNN provided the best accuracy 93.43%. Usmani, Shazia et al. [13] propose a research paper, NLP based technique is used to classify Pakistani stock exchange news. Sharun Akter Khushbu et al. [14] propose a research paper is to extract the type of news from the headline of the news. For this, 5 machine learning classifiers such as SVM, NB, Logistic Regression, Neural Network, Random Forest have been used. Using the NN algorithm performed best, with an accuracy of 90%. Ruichao Wang et al. [15] propose a research paper, news headlines are detected by a hybrid system. These tests have been done through some systems like TFTrim, HybrideTrim, Topiary, TF, Hybrid, Trim, UTD etc. Through which TFTrim system has got the best results. Syeda Sumbul Hossain et al. [16] propose a research paper which 3383 news headlines were examined in this research paper. 7 machine learning models like Naïve Bayes 80.57%, Multinomial Naïve Bayes (MNB) 76.51%, Bernoulli Naïve Bayes 82.68%, Logistic regression 76.05%, Stochastic gradient descent (SGD) 74.55%, Linear support vector classifier (SVC)

75.90%, Nu support vector classifier 75.75% and two planning models such as Long short-term memory (LSTM) 68.82%, Convolutional neural network (CNN) 70.33% are used for sentiment analysis of news headlines. In machine learning Bernoulli Naïve Bayes and in deep learning Convolutional Neural Network (CNN) performed well. Paulo Santos et al. [17] propose a research paper, news headlines are used to categorize. For this experiment SMO, Random Forest_1, Random Forest_2 classifier was used. In this experiment 62.50%, 57.50%, 61.00% accuracy was found in without relations as a feature and 62.70%, 59.00%, 63.50% accuracy was found respectively in with relations as a feature. Uchchhwas Saha et al. [18] propose a research paper analyzes the sentiment of Bengali comments using a hybrid approach, FirstText, Deep Learning classifier. “Adam”, “Glove” is used in hybrid model and BiLSTM, CNN in deep learning. The hybrid (89.89%) model obtained higher accuracy than the FastText (62.25%) model. A.N.M. JuBaer et al. [19] propose a research paper presents several ways to categorize toxic comments. A few models like MultinomialNB 52.30%, SVM 30.76%, MultinomialNB (from scikit-learn) 52.30%, GausseanNB 49.23%, Classifier Chain with MultinomialNB 52.30%, Label Powerset with MultinomialNB 58.46%, MLkNN 58.46%, BP-MLL Neural Networks 60.00% have been used. Among all the models, the BPMLL Neural Network performed well. Mushfiqus Salehin et al. [20] propose a research paper attention mechanismbased sequence to sequence model is proposed. Worked here with own Bengali dataset and achieved good results from other research papers.

2.3 Comparative Analysis and Summary

Table-1: Summary of Related Works

| Paper No | Authors & Year | Used All Models | Height Accuracy Model | Height Model (%) |
|----------|---------------------|--|-----------------------|------------------|
| 1 | Adrita Barua (2021) | Logistic Regression (LR), Support Vector Classifier (SVC), Decision Tree (DT), Multinomial Naive Bayes (MNB), Random Forest (RF), Term Frequency-Inverse Document Frequency (TF-IDF) | SVC | 97.60% |

| | | | | |
|----|----------------------------------|---|--|-----------------|
| 2 | Md. Majedul Islam (2019) | KNN, Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF) | SVM | 75% |
| 3 | Ettilla Mohiuddin Eumi (2021) | LibSVM with stop words, LibSVM without stop words, Scikit Learn Librery, SVM, NB, LR, RF, Sequential deep learning, BiGRU | BiGRU | 84% |
| 4 | Md. Rafiuzzaman Bhuiyan (2021) | LSTM | LSTM | 91.22%. |
| 5 | Hoda Ahmed Galal Elsayed (2020) | zeroR, KNN, Multilevel CNN, decision trees, Naïve Bayes, random forest and SVM | Multilevel CNN | 89.3%. |
| 6 | Ronald Tudu (2018) | SVM, MNB, SGD, LR classifier | SVM | 87.5%. |
| 7 | Fatema Jahara (2022) | deep learning-based framework multilayer perception (MLP) classifier | MLP (news articles) & MLP (news headlines) | 98.18% & 94.53% |
| 8 | Raghad Bogery (2019) | KNN, SVM, NB, MNB and Gradient boosting | Multinomial Naïve Bayes | 90.12% |
| 9 | Md. Ferdouse Ahmed Foyzal (2021) | LSTM | LSTM | 84% |
| 10 | Ke Yahan (2018) | Decision tree, Random Forest, SVC, NN | NN | 86.22% |

| | | | | |
|----|-------------------------------|---|---|--------------------------|
| 11 | Prakash Kumar Sing (2021) | LSTM, SVM, HMM, CRF | LSTM | 93.34% |
| 12 | Mohammad Rabib Hossain (2020) | ML- SVM, Naive-Bayes, Random Forrest, Logistic Regression DL- BiLSTM, CNN | CNN | 93.43% |
| 13 | Shazia Usmani (2020) | NLP based technique | - | - |
| 14 | Sharun Akter Khushbu (2020) | SVM, NB, Logistic Regression, Neural Network, Random Forest | NN | 90% |
| 15 | Ruichao Wang (2014) | TFTrim, HybrideTrim, Topiary, TF, Hybrid, Trim, UTD | TFTrim | Maximum Accuracy |
| 16 | Syeda Sumbul Hossain (2021) | ML- NB, Multinomial NB, Bernoulli NB, Logistic regression, Stochastic gradient descent (SGD), Linear support vector classifier (SVC), Nu support vector classifier. DL- LSTM, CNN. | ML- Bernoulli NB DL- CNN | ML- 82.68% DL- 70.33% |
| 17 | Paulo Santos (2015) | SMO, Random Forest_1, Random forest_2 classifier | SMO (Without relations as features) Random forest_2 (With relations as features) | 62.50% 63.50% |
| 18 | Uchchhwas Saha (2022) | BiLSTM, CNN, FastText, Hybrid. | Hybrid | 89.89% |

| | | | | |
|----|--------------------------|--|----------------------|---------------|
| 19 | A.N.M. JuBaer (2019) | Multinomial NB, SVM, Multinomial NB (from scikit-learn), Gaussean NB, Classifier Chain with Multinomial NB, Label Powerset with Multinomial NB, MLKNN, BP-MLL Neural Networks. | BPMLL Neural Network | 60.00% |
| 20 | Mushfiqus Salehin (2019) | attention mechanismbased sequence-tosequence model. | - | Best Accuracy |

2.4 Scope of the Problem

As time passes, different things are happening in the world at different times. Again, thanks to online they are available on social media, Facebook, twitter, WhatsApp, viber, Pinterest, telegram, messenger, google, YouTube etc. Thus, the amount of news is increasing day by day. The increase in the amount of news is the increase in the amount of data or documents online. But not to classify the said data or documents, that is, not to define which data falls under which category. Due to this, a lot of online data is lost which cannot be used in future. Therefore, to make all these data or documents usable, various algorithms of machine learning and deep learning such as logistic regression, linear regression, random forest classifier, decision tree, naive biyas, SVM, CNN, RNN, ANN, LSTM etc. the data is classified. As a result, the documents or data obtained online can be used for various purposes.

2.5 Challenges

In today's world the amount of news is increasing day by day, in a few days there will be more unusable data than usable data online. The reason for the un-usability of the said data or documents is not to classify, that is, not to define which data falls under which category. Due to this, a lot of online data is lost which cannot be used in future. If we do not change the usability of all these data now, we will face many problems in the future. Using data in the future to make changes will require a lot of time and skilled manpower. This will waste both time and money. So, if the data received now i.e., the data available on social media or other platforms is classified then the data will be useful and will not face problems in the future.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Dataset Description

The main dataset used in this paper consists of 3 columns namely Headline, Category and Newspaper Name. The category is divided into 6 categories, namely- International, National, Sports, Amusement, Politics and IT. To run the Machine Learning and Deep Learning algorithms, we used a fundamental partitioning of the data set consisting of 95552 samples for training, 13272 samples for testing, and 23889 samples for validation.

Table-2: Category name or Description

| Source News | Category | Description | Label |
|-------------------------|---------------|--|--|
| Online News Headline | International | International news will be shown from all types of news. | International, National, Sports, Amusement, Politics and IT. |
| | National | National news will be shown from all types of news. | |
| | Sports | Sports news will be shown from all types of news. | |
| | Amusement | Amusement news will be shown from all types of news. | |
| | Politics | Politics news will be shown from all types of news. | |
| | IT | IT news will be shown from all types of news. | |

3.2 Dataset Pre-Processing

The data taken from online is basically unstructured data. So, after collecting the data it is necessary to process the data. Because the data obtained from online, blogs, social media contains duplicate data, unintelligible data, null values etc. The dataset is processed to remove low length data, remove unique data, remove duplicate data, dataset cleaning and transform it into understandable data. Besides, word list, sort the dictionary of word list, documents per class, total word per class, unique words per class, dataset

splitting etc. have been extracted from the dataset. Consequently, we need to apply data pre-processing methods to our dataset. To apply pre-processing method, we follow below steps-

- Remove URLs, screen names and hashtags.
- Remove zero values, emojis, symbols, punctuation and numbers.
- Remove all retweets and unnecessary symbols.
- Remove all low length data.

3.3 Statistical Analysis

- The dataset contains a total of 136811 amount of data.
- The dataset keeps 3 columns.
- The dataset contains a total of 95552 training data.
- The dataset contains a total of 13272 testing data.
- The dataset contains a total of 23889 validation data.
- Category are classified into 6 steps (International, National, Sports, Amusement, Politics and IT).

3.4 Design Approach

Since the data is a multivariate classification problem, we used supervised and unsupervised learning algorithms. Deep learning such as BiLSTM, GRU, Language Model (Uni-Gram), Machine Learning such as Logistic Regression, Random Forest Classifier, Multinomial Naive Bayes, SVM, Decision Tree Classifier, Gradient Boosting Algorithm etc. Their accuracy and performance predictions, results analysis, and a confusion matrix were generated for each model. The architecture diagram for the entire system is shown in figure:

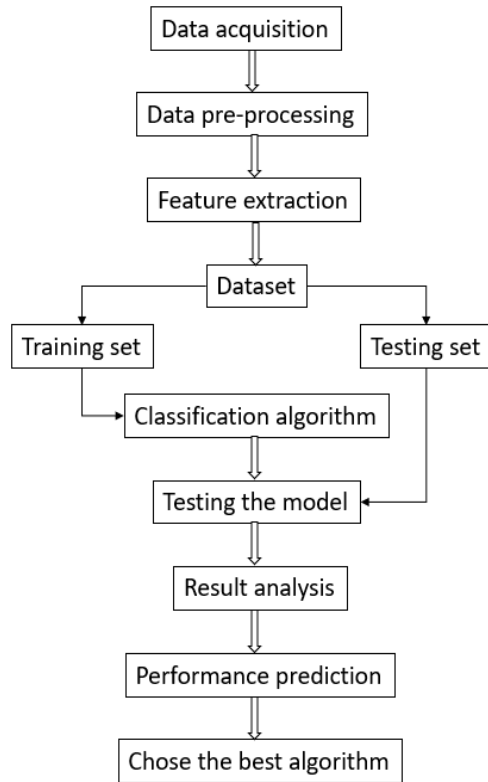


Figure-1: Architecture of Working Process

3.5 Proposed Methodology

3.5.1 Bidirectional Long Short-Term Memory (BiLSTM)

BiLSTM means Bidirectional LSTM, it is a sequence processing model. This sequence processing model consists of two LSTM, a) one with no backward input, b) the other with no forward input. This algorithm is a widely used algorithm in NLP. This model is used in many applications in NLP such as speech recognition, Handwritten recognition, Protein structure prediction, etc. Can demonstrate improved performance for sequential classification problems.

First prepare the dataset and distribute the dataset. A total of 136811 headlines appears in the dataset. The 136811 headlines are divided into 6 categories (international, sports, national, amusement, politics, IT), among which international headlines are the most, followed by sports as the second highest, then national as the third highest, then respectively amusement, politics and finally IT related headlines come the least. Below is a graph of data distribution:

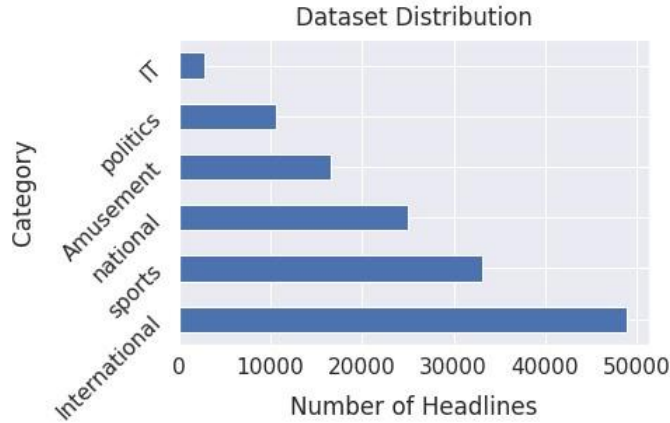


Figure-2: Dataset Distribution for Bi-LSTM

After data preparation, the total headlines were 132713 and 4098 were removed from the dataset by cleaning the data. Then by analyzing the entire dataset, each separate category is analyzed again among all the headlines and the number of headlines, number of words, number of unique words and most frequent word of that individual category have been extracted. Thus, by analyzing the international class, its headline number (47885), word number (307354), unique word number (28710) and most frequent word have been extracted. By analyzing the sports class, its headline number (30831), word number (152852), unique word number (18581) and most frequent word have been extracted. Similarly, the number of headlines, number of words, number of unique words and most frequent word of national, amusement, politics and IT have been extracted. Below is the graph of dataset analysis:

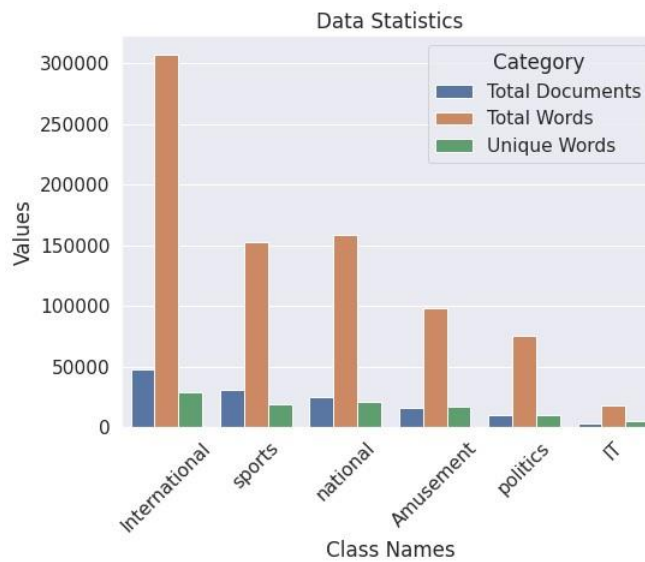


Figure-3: Dataset Statistics for Bi-LSTM

After the dataset visualization, the length and frequency of the headlines of the dataset were measured. Among which the maximum length of the title is 21, the minimum length of the title is 3 and the average length of the title is 6. Below is the visualized graph of length frequency distribution:

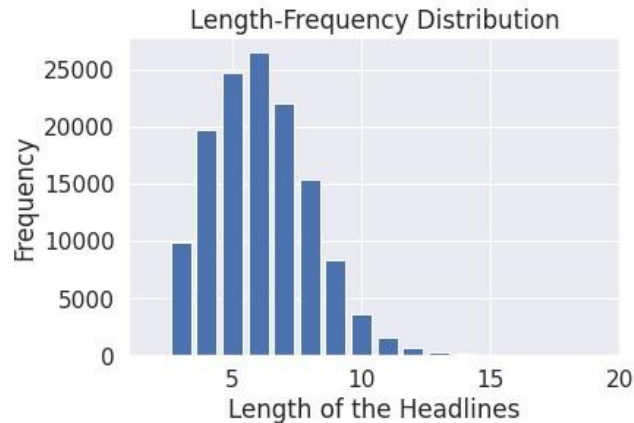


Figure-4: Length-frequency distribution for Bi-LSTM

This dataset is encoded using label encoding. Label encoding means converting the categorical language of the dataset into a numerical language, so that the natural language can be converted into a machine language that can be understood by the machine. Basically "LabelEncoder" function is used to convert categorical language to numerical language.

After converting all the data into numerical language, the dataset is split. The dataset is divided into 2 parts, a) training data, b) test data. Out of total 132713 data, training data is 95552 of which unique data is 55055 and testing data is 13272. Some data is kept for validation checking, there are 23889 data for validation checking.

A model discussed in BiLSTM is called sequential model. Input or output data using this model is known as sequence model. The sequence model basically works with Audio clips, Text streams, time-series data, video clips. Total parameters (4246240), trainable parameters (4246240) and non-trainable parameters (0) are fitted to the dataset using this model.

3.5.1.1 Validation and Training accuracy

Validation accuracy and training accuracy are calculated with 7 epochs in the models. In validation accuracy, the accuracy in eps-1 is improved from infinity to 0.81330, the accuracy in eps-2 is improved from 0.81330 to 0.82992, the accuracy in eps-3 is improved from 0.82992 to 0.83114. It has been done, the accuracy in App-4 is from 0.83114 to 0.83114 i.e., there is no improvement, the validation accuracy in App-5,6,7 is coming with the same value. That is, the value of validation accuracy is upward.

On the other hand, in training accuracy, the training accuracy in Eps-1 is 0.7239, in Eps-2 the training accuracy is 0.8747, in Eps-3 the training accuracy is 0.9203, in Eps-4 the training accuracy is 0.9437, Training assurance in EPS-5 is 0.9571, training assurance in EPS-6 is 0.9672, training assurance in EPS-7 is 0.9741. That is, the value of training insurance is parallel, but the value of a small number is upward, that is, the difference is very less.

Table-3: Validation and Training accuracy table for Bi-LSTM

| 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 |

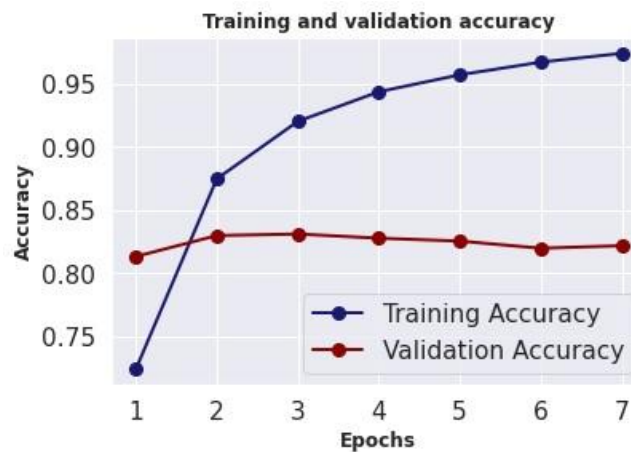


Figure-5: Epochs vs validation and training accuracy plot for Bi-LSTM

3.5.2.2 Validation and Training loss

Validation and training loss are calculated with 7 epochs in the models. In validation loss, validation loss in epoch-1 is 0.5203, validation loss in epoch-2 is 0.4690, validation loss in epoch-3 is 0.5090, validation loss in epoch-4 is 0.5495, validation loss at epoch-5 is 0.6179, validation loss at epoch-6 is 0.6767, validation loss at epoch-7 is 0.7656. That is, the value of validation loss is upward.

On the other hand, training loss in epoch-1 is 0.7419, training loss in epoch-2 is 0.3552, training loss in epoch-3 is 0.2279, training loss is 0.1631 in epoch-4, training loss in epoch-5 is 0.1212, training loss in epoch-6 is 0.0912, training loss in epoch-7 is 0.0707. That is, the value of training loss is downward.

Table-4: Validation and Training loss for Bi-LSTM

| 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 |

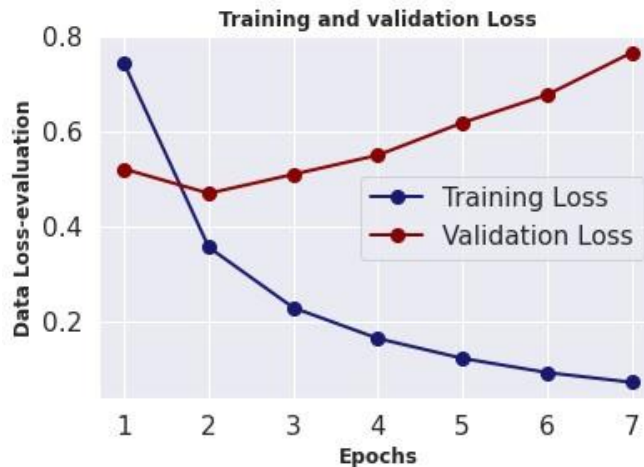


Figure-6: Epochs vs validation and training loss plot for Bi-LSTM

3.5.2.3 Confusion matrix

We can see from the confusion matrix that the Bi-LSTM report accuracy is 83.42%.

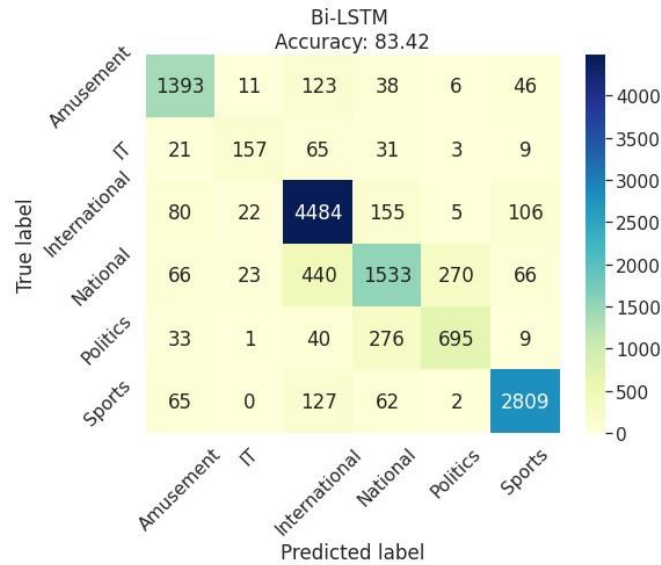


Figure-7: Confusion Matrix of Bi-LSTM Algorithm

3.5.2.4 Classification report

News Headline- International, Sports, Politics, National, Amusement and IT, by analyzing the classes, classification report i.e., precision, recall, f1_score and support showed that B-LSTM has 83.42% assurance report.

Table-5: Classification report of Bi-LSTM Algorithm

| 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 |
|--------|-------|-------|-------|-------------|

| | Precision | Recall | F1_Score | Support |
|--------------|-----------|--------|----------|--------------|
| 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 |

In the classification report we can see that Precision, Recall, F1_Score and Support are extracted for all news types (Amusement, IT, International, National, Politics, Sports). Along with finding out Accuracy, Macro Avg, Weighted Avg value has also been found out. Accuracy of BiLSTM algorithm is 83.42.

3.5.2 Gated recurrent units (GRU)

Gated recurrent units (GRU) model is a gating mechanism system of a recurrent neural network, this mechanism system was introduced in 2014 by Kyunghyun Cho. The GRU model basically uses sequential connections of nodes to perform machine learning tasks related to memory and clustering. It is basically a part of recurrent neural network model. GRU is similar to LSTM, but GRU has one fewer parameter than LSTM. Because GRU has no output gate. The two gates of GRU are reset and update. On the other hand, the three gates of LSTM are input, output and forget. This means that GRU is less complex than LSTM because the number of gates in GRU is less.

How GRU works - There is a GRU cell that works basically like an LSTM or RNN cell. At each timestamp t , it takes an input X_t and the hidden state H_{t-1} from the previous timestamp $t-1$. Later it outputs a new hidden state H_t which again passed to the next timestamp.

First prepare the dataset or distribute the dataset. A total of 136811 headlines appears in the dataset. The 136811 headlines are divided into 6 categories (international, sports, national, amusement, politics, IT), among which international headlines are the most, followed by sports as the second highest, then national as the third highest, then respectively amusement, politics and finally IT related headlines come the least. BiLSTM and GRU, the data set distribution graphs of these two algorithms show that the data set distributions of the two algorithms are almost similar. Below is a graph of data distribution:

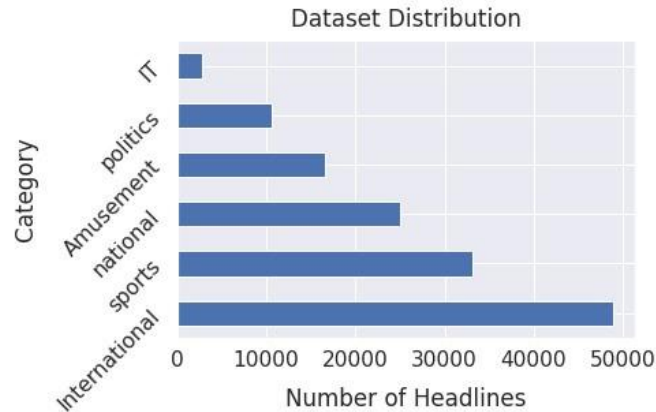


Figure-8: Dataset Distribution for GRU

After data preparation, the total headlines were 132713 and 4098 were removed from the dataset by cleaning the data. Then by analyzing the entire dataset, each separate category is analyzed again among all the headlines and the number of headlines, number of words, number of unique words and most frequent word of that individual category have been extracted. Thus, by analyzing the international class, its headline number (47885), word number (307354), unique word number (28710) and most frequent word have been extracted. By analyzing the sports class, its headline number (30831), word number (152852), unique word number (18581) and most frequent word have been extracted. Similarly, the number of headlines, number of words, number of unique words and most frequent word of national, amusement, politics and IT have been extracted. A total of 57490 unique words were found from the data set. BiLSTM and GRU, the data statistics graphs of these two algorithms show that the data statistics graphs of the two algorithms are almost similar. Below is the graph of dataset analysis:

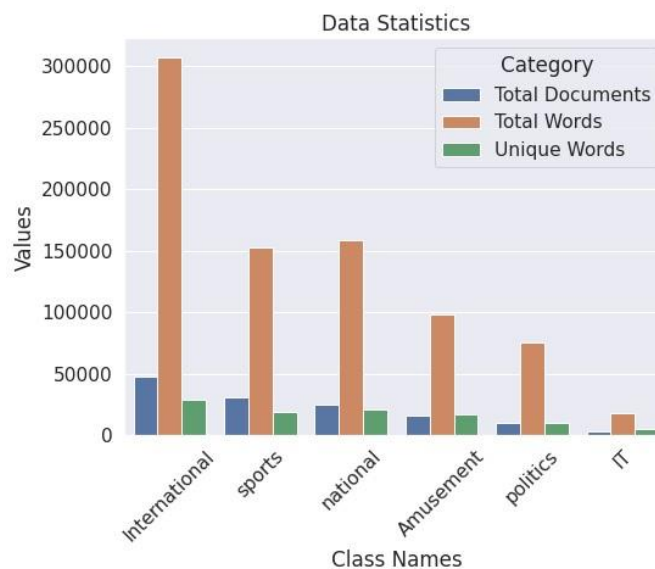


Figure-9: Dataset Statistics for GRU

After the dataset visualization, the length and frequency of the headlines of the dataset were measured. Among which the maximum length of the title is 21, the minimum length of the title is 3 and the average length of the title is 6. BiLSTM and GRU, the Length-Frequency Distribution graphs of these two algorithms show that the Length-Frequency Distribution graphs of the two algorithms are almost similar. Below is the visualized graph of length frequency distribution:

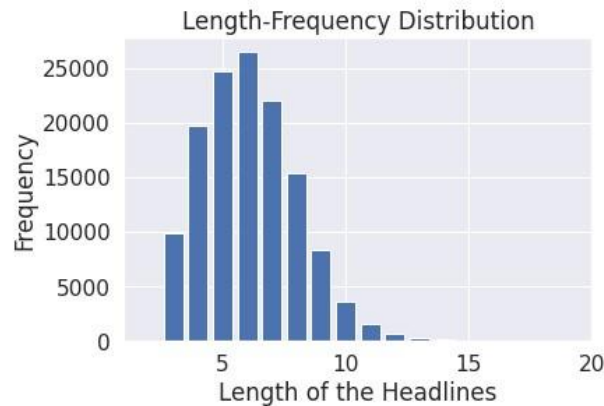


Figure-10: Length-Frequency Distribution for GRU

This dataset is encoded using label encoding. Label encoding means converting the categorical language of the dataset into a numerical language, so that the natural language can be converted into a machine language that can be understood by the machine. Basically "LabelEncoder()" function is used to convert categorical language to numerical language.

After converting all the data into numerical language, the dataset is split. The dataset is divided into 2 parts, a) training data, b) test data. Out of total 132713 data, training data is 95552 of which unique data is 55055 and testing data is 13272. Some data is kept for validation checking, there are 23889 data for validation checking.

GRU is a gating mechanism in recurrent neural network that uses two gates: update gate and reset gate. GRU handles the clustering and memory related tasks of machine learning. The GRU model is mainly applied in speech recognition, human genome, handwriting analysis and many other researches. Total parameters (3701166), trainable parameters (3701166) and non-trainable parameters (0) are fitted to the dataset using this model.

3.5.2.1 Validation and Training accuracy

Validation and training accuracy are calculated with 7 epochs in the models. In validation accuracy, validation accuracy in epoch-1 is 0.8205, validation accuracy in epoch-2 is 0.8320, validation accuracy in

epoch-3 is 0.8365, validation accuracy in epoch-4 is 0.8325, validation accuracy at epoch-5 is 0.8278, validation accuracy at epoch-6 is 0.8282, validation accuracy at epoch-7 is 0.8230.

On the other hand, training accuracy in epoch-1 is 0.7429, training accuracy in epoch-2 is 0.8749, training accuracy in epoch-3 is 0.9203, training accuracy is 0.9426 in epoch-4, training accuracy in epoch-5 is 0.9549, training accuracy in epoch-6 is 0.9644, training accuracy in epoch-7 is 0.9707. That is, the value of training accuracy is upward.

Table-6: Validation and Training accuracy table for GRU

| 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 |

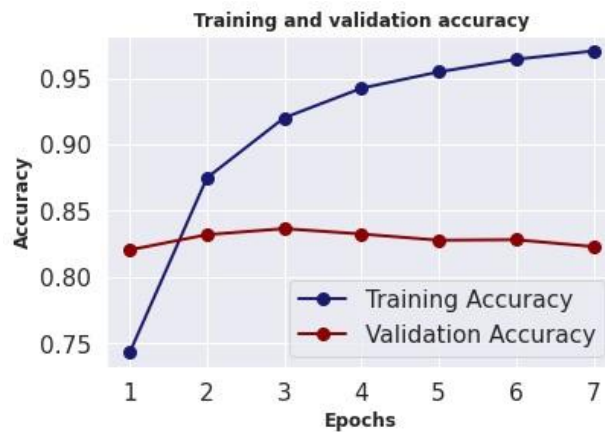


Figure-11: Epochs vs validation and training accuracy plot for GRU

3.5.2.2 Validation and Training loss

Validation and training loss are calculated with 7 epochs in the models. In validation loss, validation loss in epoch-1 is 0. 5015, validation loss in epoch-2 is 0. 4804, validation loss in epoch-3 is 0. 5007, validation loss in epoch-4 is 0. 5160, validation loss at epoch-5 is 0. 5953, validation loss at epoch-6 is 0. 6193, validation loss at epoch-7 is 0. 6870. That is, the value of validation loss is upward.

On the other hand, training loss in epoch-1 is 0. 7035, training loss in epoch-2 is 0. 3537, training loss in epoch-3 is 0. 2279, training loss is 0. 1639 in epoch-4, training loss in epoch-5 is 0. 1262, training loss in epoch-6 is 0. 1005, training loss in epoch-7 is 0. 0819. That is, the value of training loss is downward.

Table-7: Validation and Training loss table for GRU

| 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 |

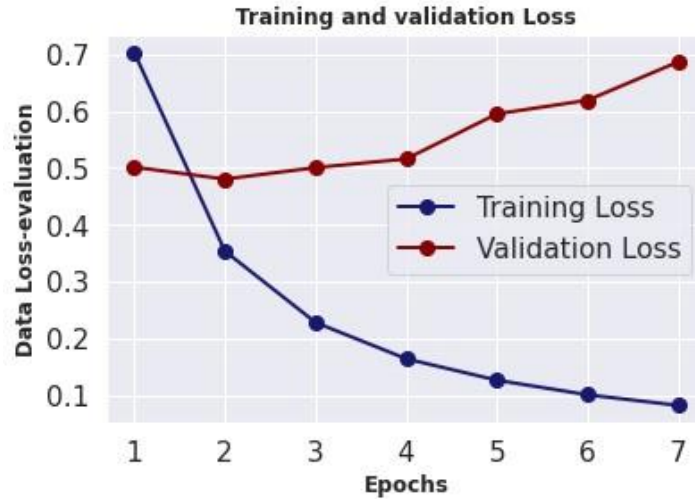


Figure-12: Epochs vs validation and training loss plot for GRU

3.5.2.3 Confusion matrix

We can see from the confusion matrix that the GRU report accuracy is 84.01%.

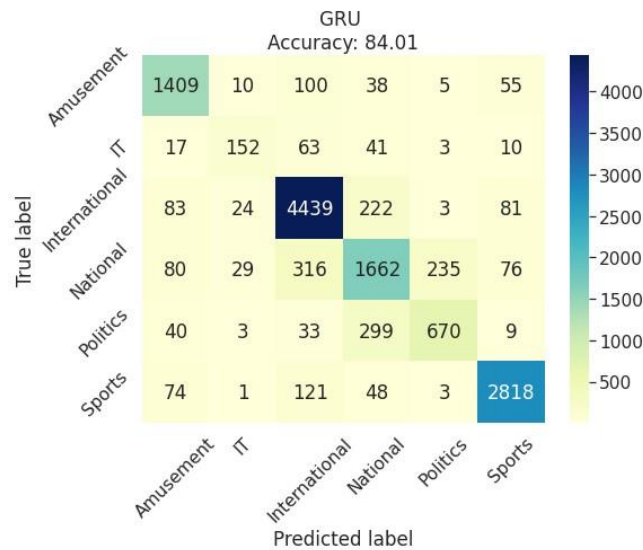


Figure-13: Confusion matrix for GRU

3.5.2.4 Classification report

News Headline- International, Sports, Politics, National, Amusement and IT, by analyzing the classes, classification report i.e., precision, recall, f1_score and support showed that GRU has 84.01% assurance report.

Table-8: Classification report for GRU

| 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 |

| | Precision | Recall | F1_Score | Support |
|--------------|-----------|--------|----------|--------------|
| 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 |

In the classification report we can see that Precision, Recall, F1_Score and Support are extracted for all news types (Amusement, IT, International, National, Politics, Sports). Along with finding out Accuracy, Macro Avg, Weighted Avg value has also been found out. Accuracy of GRU algorithm is 84.01.

3.5.3 Machine Learning (ML)

Computer science and artificial intelligence are fields that look at how humans learn and observe, using data and algorithms to gradually improve their results and accuracy. And machine learning is a branch of artificial intelligence (AI) and computer science. Machine learning algorithms are basically of four types. Namely – supervised, semi-supervised, unsupervised and reinforcement.

There are many types of models in machine learning algorithms. Like-Linear regression, Logistic regression, Decision tree, SVM algorithm, Naive Bayes algorithm, KNN algorithm, K-means, Random Forest algorithm, Dimensionality reduction algorithms, Gradient boosting algorithm and AdaBoosting algorithm.

First prepare the dataset or distribute the dataset. A total of 136811 headlines appears in the dataset. The 136811 headlines are divided into 6 categories (international, sports, national, amusement, politics, IT), among which international headlines are the most, followed by sports as the second highest, then national as the third highest, then respectively amusement, politics and finally IT related headlines come the least. BiLSTM, GRU and ML, the data set distribution graphs of these three algorithms show that the data set distributions of the three algorithms are almost similar. Below is a graph of data distribution:

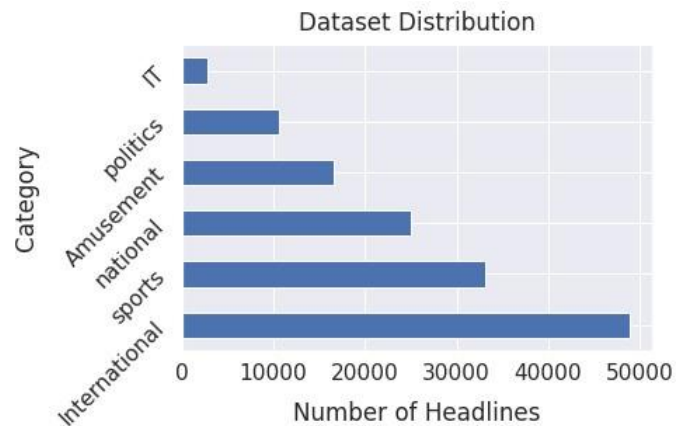


Figure-14: Dataset Distribution for ML

After data preparation, the total headlines were 132713 and 4098 were removed from the dataset by cleaning the data. Then by analyzing the entire dataset, each separate category is analyzed again among all the headlines and the number of headlines, number of words, number of unique words and most frequent word of that individual category have been extracted. Thus, by analyzing the international class, its headline number (47885), word number (307354), unique word number (28710) and most frequent word have been extracted. By analyzing the sports class, its headline number (30831), word number (152852), unique word number (18581) and most frequent word have been extracted. Similarly, the number of headlines, number of words, number of unique words and most frequent word of national, amusement, politics and IT have been extracted. A total of 57490 unique words were found from the data set. BiLSTM, GRU and ML, the data statistics graphs of these three algorithms show that the data statistics graphs of the three algorithms are almost similar. Below is the graph of dataset analysis:

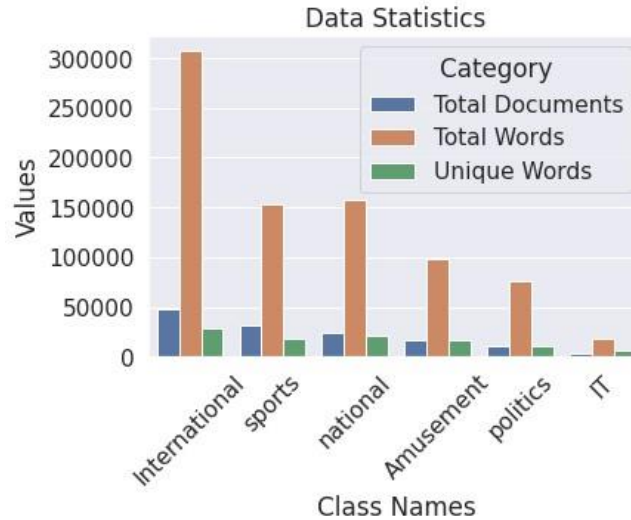


Figure-15: Data Statistics for ML

After the dataset visualization, the length and frequency of the headlines of the dataset were measured. Among which the maximum length of the title is 21, the minimum length of the title is 3 and the average length of the title is 6. BiLSTM, GRU and ML, the Length-Frequency Distribution graphs of these three algorithms show that the Length-Frequency Distribution graphs of the three algorithms are almost similar. Below is the visualized graph of length frequency distribution:

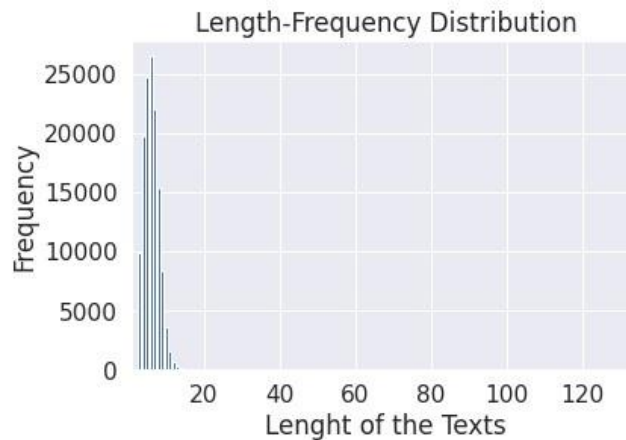


Figure-16: Length-Frequency Distribution for ML

3.5.3.1 Logistic Regression

Logistic regression is supervised learning. Logistic regression is also called a statistical analysis method. This algorithm is used to predict the probability of essentially binary (0/1, yes/no) events. That is, it predicts by calculating the probability of statistics and binary outcomes like (0/1, yes/no). Logistic regression models are primarily used to describe the relationship between data, dependent binary variables, and one or more

nominal, ordinal, interval, or ratio-level independent variables. Logistic regression is also used to classify large datasets. There are basically three types of logistic regression models. Eg – Binary Logistic Regression, Multinomial Logistic Regression, Ordinal Logistic Regression.

Accuracy score of the training data: 0.6645850993689366

Table-9: Classification report for Logistic Regression

| 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 |

| | Precision | Recall | F1_Score | Support |
|--------------|-----------|--------|----------|---------|
| Accuracy | | | 0.64 | 26543 |
| Macro Avg | 0.62 | 0.52 | 0.55 | 26543 |
| Weighted Avg | 0.64 | 0.64 | 0.63 | 26543 |

In the classification report we can see that Precision, Recall, F1_Score and Support are extracted for all news types (Amusement, IT, International, National, Politics, Sports). Along with finding out Accuracy, Macro Avg, Weighted Avg value has also been found out. Accuracy of Logistic Regression algorithm is 0.64.

3.5.3.2 Multinomial Naive Bayes

Multinomial Naive Bayes Algorithm is basically Bayesian learning method. Which is very popular in NLP. This method is the main tool for analyzing textual input and solving many class problems. This method is mainly based on the principle of text-based classifier. There are basically three types of naive bayes based

on the Skit-Learn library. namely Gaussian, Polynomial and Bernoulli. Naive Bayes method is basically based on the following formula: $P(A|B) = P(A) * P(B|A)/P(B)$.

Table-10: Classification report for Multinomial Naive Bayes

| 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 |

| | Precision | Recall | F1_Score | Support |
|--------------|-----------|--------|----------|---------|
| Accuracy | | | 0.61 | 26543 |
| Macro Avg | 0.65 | 0.43 | 0.45 | 26543 |
| Weighted Avg | 0.63 | 0.61 | 0.58 | 26543 |

In the classification report we can see that Precision, Recall, F1_Score and Support are extracted for all news types (Amusement, IT, International, National, Politics, Sports). Along with finding out Accuracy, Macro Avg, Weighted Avg value has also been found out. Accuracy of Multinomial Naive Bayes algorithm is 0.61.

3.5.3.3 Random Forest Classifier

Random Forest is a supervised learning method composed of decision trees. This algorithm is mainly used to solve classification and regression problems. The algorithm constructs a decision tree by deciding on many types of samples and takes the highest vote of classification and average in regression. Additionally, dataset analysis takes averages to improve predictive accuracy.

Table-11: Classification report for Random Forest Classifier

| 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 |

| | Precision | Recall | F1_Score | Support |
|--------------|-----------|--------|----------|---------|
| Accuracy | | | 0.65 | 26543 |
| Macro Avg | 0.64 | 0.53 | 0.56 | 26543 |
| Weighted Avg | 0.65 | 0.65 | 0.64 | 26543 |

In the classification report we can see that Precision, Recall, F1_Score and Support are extracted for all news types (Amusement, IT, International, National, Politics, Sports). Along with finding out Accuracy, Macro Avg, Weighted Avg value has also been found out. Accuracy of Random Forest Classifier is 0.65.

3.5.3.4 Support Vector Machine (SVM)

SVMs are used in applications such as handwriting recognition, intrusion detection, face detection, email classification, gene classification, and web pages. This is one reason why we use SVM in machine learning. It can handle both classification and regression on linear and non-linear data.

Support Vector Machine (SVM) is a supervised learning. The SVM algorithm is mainly used to solve both classification and regression types of problems. But in classification problems civilized algorithms work best. It is used to solve both linear and non-linear types of problems. The purpose of the SVM algorithm is to find a hyperplane in an n-dimensional space that neatly groups all data points into one class. This algorithm is used to solve problems like handwriting recognition, intrusion detection, face detection, email classification, gene classification, web pages.

Table-12: Classification report for SVM

| 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 |

| | Precision | Recall | F1_Score | Support |
|--------------|-----------|--------|----------|---------|
| Accuracy | | | 0.65 | 26543 |
| Macro Avg | 0.64 | 0.53 | 0.56 | 26543 |
| Weighted Avg | 0.65 | 0.65 | 0.64 | 26543 |

In the classification report we can see that Precision, Recall, F1_Score and Support are extracted for all news types (Amusement, IT, International, National, Politics, Sports). Along with finding out Accuracy, Macro Avg, Weighted Avg value has also been found out. Accuracy of SVM algorithm is 0.65.

3.5.4 Traditional Approach Uni-gram

The simplest form of language model simply strips away all conditioning context and infers each word independently and separately. That is, as many words as there are in a sentence, each of them is a unigram separately. E.g. - "I eat rice", in this sentence "I" is a unigram, "eat" is a unigram, "rice" is a unigram. Such a model is called a unigram language model.

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

Unigram approach is also applied to our dataset. Each headline sentence was converted into individual words by putting it through the unigram approach. Original dataset shape Counter and Resampled dataset shape Counter are extracted by applying unigram approach.

The dataset is split into 2 parts, a) training data, b) test data. Out of total 132713 data, training data is 70% and testing data is 30%. Data mining algorithm is applied after dividing the dataset into training and testing. Data mining algorithms such as Decision Tree Classifier, Random Forest Classifier, KNN Algo, Multinomial Naive Bias, Gradient Boosting, SVM, Logistic Regression have been applied.

3.5.4.1 Decision Tree Classifier

A decision tree is a non-parametric supervised learning algorithm, where the data is split continuously. This algorithm is used for both classification and regression problems. This decision tree basically consists of hierarchical tree, root nodes, branches, internal nodes and leaf nodes. This algorithm generates a tree by predicting the solution to a problem. Thus, the decision tree works by generating the tree. Here, Precision, Recall, F1_Score, Support have been extracted through Decision Tree Classifier, the accuracy of which is 72%.

Table-13: Classification report for Decision Tree Classifier

| | 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

Gradient boosting is a supervised learning algorithm. Gradient boosting is also called iterative functional gradient algorithm. It combines decision trees through a technique called boosting technique. Here, Precision, Recall, F1_Score, Support have been extracted through Gradient boosting algorithm, the accuracy of which is 72%.

Table-14: Classification report for Gradient Bosting Algorithm

| 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 |

In the classification report we can see that Precision, Recall, F1_Score and Support are extracted for all news types (Amusement, IT, International, National, Politics, Sports). Along with finding out Accuracy, Macro Avg, Weighted Avg value has also been found out. Accuracy of Gradient Bosting Classifier is 0.72.

3.5.4.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm. This algorithm is used in both classification and regression. All things considered; regression is the most suitable problem for classification. The main objective of this algorithm is to find a hyperplane in n-dimensional space and classify the data points unambiguously. It works well for both linear and non-linear problems. Here, Precision, Recall, F1_Score, Support have been extracted through support vector machine algorithm, the accuracy of which is 87%.

Table-15: Classification report for SVM

| 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

Logistic regression is a supervised classification algorithm. It is also called Machine Learning Classification Algorithm. which is used to predict the probability of certain classes from some dependent variable points. Briefly, this model computes the sum of input characteristics and the logistic of the outcome. Here, Precision, Recall, F1_Score, Support have been extracted through support vector machine algorithm, the accuracy of which is 72%.

Table-16: Classification report for Logistic Regression

| 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

Random Forest is a supervised machine learning algorithm. It is mainly used in classification and regression problems. For example, classifying whether it "will" or "won't" be played today. Here, Precision, Recall, F1_Score, Support have been extracted through support vector machine algorithm, the accuracy of which is 89%.

Table-17: Classification report for Random Forest Classifier

| 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

Algorithms such as Decision Tree Classifier, Gradient Boosting Algorithm, Support Vector Machine, Logistic Regression, Random Forest Classifier etc. have been applied in Unigram. Among them, the

accuracy of Decision Tree is 72%, the accuracy of Gradient Boosting is 72%, the accuracy of SVM is 87%, the accuracy of Logistic Regression is 72%, the accuracy of Random Forest is 89%. It turns out that Random Forest's accuracy is the highest at 89%.

Table-18: Summary of Classifiers accuracy

| 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

Analyzing the type of news headlines, classification reports such as precision, recall, f1_score and support showed that Unigram has a 95% assurance report.

Table-19: Classification report for Decision Tree

| 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)

ROC is the receiver operating characteristic. ROC curve refers to the test sensitivity plot. ROC curves compare sensitivity versus specificity over the entire range for predicting two-sided outcomes. Another measure of test performance is the area under the ROC curve. For example, medical diagnostic tests can discriminate between "diseased" and "non-diseased" patients by status.

Here it can be seen that the accuracy of Random Forest Classifier is found to be the highest. So, ROC is extracted with Random Forest Classifier.

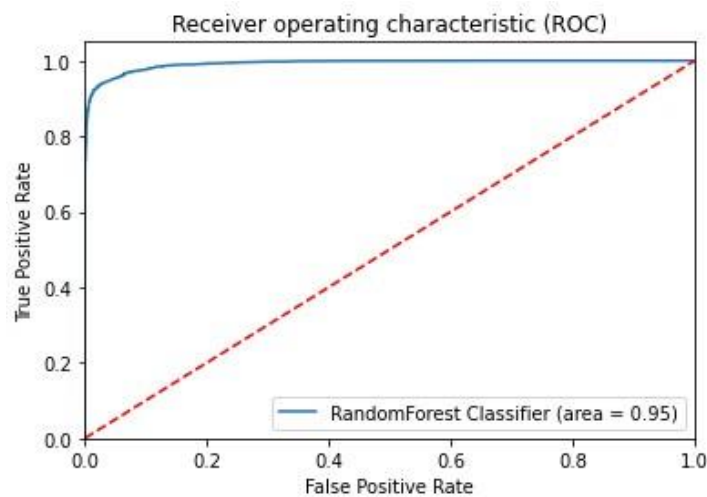


Figure-17: Receiver Operating Characteristic Graph

3.5.4.9 Classification and ROC Analysis

Since the accuracy of Random Forest Classifier is the highest, the receiver operating characteristic has been extracted with it. Now cross-validation is done by extracting the ROC of the random forest to see if there is anything wrong or if everything is coming right.

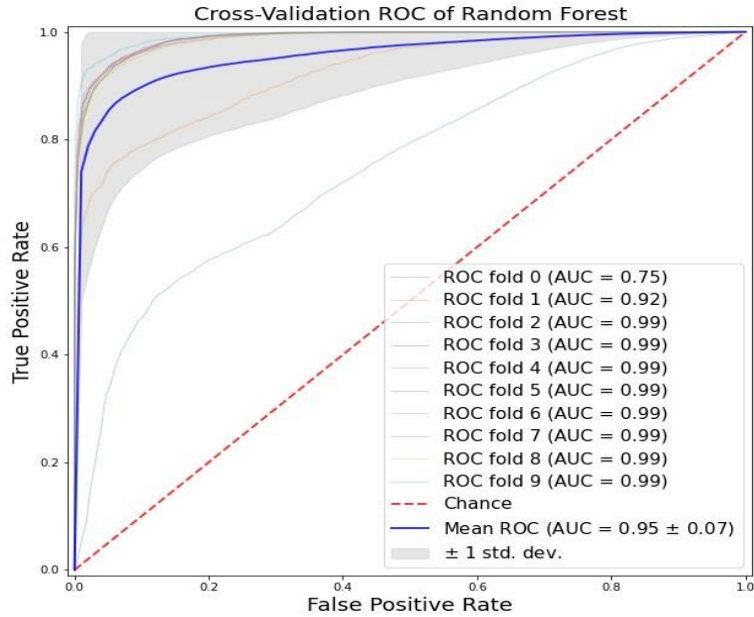


Figure-18: Check Validation Receiver Operating Characteristic

CHAPTER 4 EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Discussion

In today's world, the amount of news is increasing rapidly. As the day goes by, different types of news are being created about different things. News is not only published in newspapers, but also on various online news portals, social media, blogs and various types of news-oriented websites. There is a certain measure for writing newspaper news that cannot be overwritten. But in online based news portals, websites, blogs or social media there is no limit to writing so much. Therefore, the online news is written in a very detailed and descriptive manner. As a result, the amount of news has increased tremendously and the amount of news is constantly increasing. If this cumulative news can be categorized in some way, then this news can be used in future. As this news are not categorized in this way, many important news is missed by us. The news can be divided into International, National, Sports, Amusement, Politics, IT, Entertainment, Education and many more categories. Therefore, if the online news is classified in a digitized way using machine learning or deep learning, this news can be used for many important purposes in the future.

In this research paper, some algorithms of machine learning and deep learning are applied. The applied algorithms are BiLSTM, GRU, Uni-gram, Machine Learning (Logistics regression, Multinational naïve bayes, Random Forest classifier, Support vector machine).

Table-20: Classifiers Description

| Classifier | | Description |
|------------------|-----------------------------|---|
| Bi-LSTM | | BILSTM combines information from both the past and the future with input sequences to produce a more accurate output. |
| GRU | | GRUs are typically used in combination with the in-memory and clustering of machine learning to sort nodes. It uses two gates: update gate and reset gate. |
| Machine Learning | Logistic Regression | Logistic regression is commonly used to predict binary outcomes. For example: yes or no, true or false, 0 or 1. |
| | Multinomial Naïve Bayes | Multinomial Naive Bayes algorithm is very popular in NLP in general. This algorithm is the fastest and easiest way to successfully perform sentiment analysis. |
| | Random Forest Classifier | The random forest classifier algorithm is commonly used in classification and regression related problems. It constructs a decision tree on the dataset, classifies and averages it to create a better model. |
| | SVM | The task of SVM is to clearly classify the data points of the dataset through the maximum marginal hyperplane. |
| UniGram | Decision Tree | The decision tree algorithm divides the dataset continuously according to certain parameters to obtain two entities namely decision nodes and leaves. |
| | Gradient Boosting Algorithm | The gradient boosting algorithm provides another updated prediction model against weak prediction models, which are usually associated with decision trees. |
| | SVM | The task of SVM is to clearly classify the data points of the dataset through the maximum marginal hyperplane. |
| | Logistic Regression | Logistic regression is commonly used to predict binary outcomes. For example: yes or no, true or false, 0 or 1. |
| | Random Forest Classifier | The random forest classifier algorithm is commonly used in classification and regression related problems. It constructs a decision tree on the dataset, classifies and averages it to create a better model. |

4.2 Experimental Results and Analysis

Two deep learning models are used in the paper namely BiLSTM and GRU. BiLSTM and GRU have 83.42% accuracy in BiLSTM and 84.01% accuracy in GRU. It appears that using deep learning has yielded the highest accuracy in the GRU model. Along with declining, some machine learning models such as LR, MNB, RF and SVM have been applied. Here LR's accuracy is 64%, MNB's accuracy is 61%, RF's accuracy is 65%, and SVM's accuracy is 65%. It can be seen that RF and SVM provided the highest accuracy out of all the models, both with an accuracy value of 65%. Again, unigram language model is used where decision tree, gradient boosting algorithm, SVM, logistic regression and random forest classifier algorithm are applied. It is seen that the accuracy of decision tree is 72%, the accuracy of gradient boosting algorithm is 72%, the accuracy of SVM is 87%, the accuracy of logistic regression is 72%, and the accuracy of random forest classifier is 89%. Following the unigram approach here, the highest accuracy is obtained with the random forest classifier.

Table-21: Classifiers accuracy, recall and precision

| 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

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Because of the web, virtual entertainment has now added a new dimension to present day life. People in country regions never again newspapers, they search and read news online through cellphones. All that is occurring in the country and abroad is accessible online through YouTube, Messenger, Instagram, LinkedIn, Pinterest, Trambler, Snapchat, Viber, WhatsApp, Facebook, Twitter, Google and other virtual entertainment. Virtual entertainment timelines or newsfeeds are getting loaded up with a wide range of essential and pointless news. 70% of internet users worldwide are connected through social media. Among youngsters the rate is around 90%. A report shows that around 80% of web clients in Bangladesh use Facebook. The utilization of web has expanded the degree of social communication commonly than previously. Information, opinions, pictures, videos and so forth of any individual are traded through different kinds of data and documents through innovation. None of these kinds of data and documents are characterized, so in the event that we look for a particular point, it doesn't come up well. Thus, this is a major drawback. Data classification is finished to overcome this trouble. That is, assuming that the news is arranged into various classifications like International, National, Sports, Amusement, Politics and IT and so on, it will be exceptionally helpful in tracking down news. Thus, assuming you look for news by composing any sort of watchword, the possibilities of fresh insight about that catchphrase type will increment commonly. In this way, to peruse any sort of information it is vital to initially characterize the news into various classifications. Classifying it into various classes will make it a lot more straightforward for individuals from various different backgrounds to read news on the web or through virtual entertainment.

5.2 Impact on Environment

People have been interested on information since old times. People used to read news from papers since old times. Yet, as the day goes by the approach of innovation individuals don't read news in papers any longer. These days all types of information are accessible through internet and all types of information are accessible through social media. Any kinds of information about what's going on anyplace on the planet comes to us in a second. Any report about the climate through the Meteorological Division comes to us online inside a brief timeframe. Like where on earth there is a twister, where there is a flood, where there is a cyclone and what is the environment of a region of the world, all kind of information come to us through on the web. We can know immediately what the weather conditions resembles in any space or any country.

We don't take a gander at the climate or the meteorological forecast to conclude what work should be possible sooner or later. In this manner it is feasible to make due from the substance of numerous huge mishaps or misfortunes. Therefore, the contribution of information in the climate is quite a large number.

5.3 Ethical Aspects

Each news media has a particular strategy or publication strategy. Media can embrace approaches autonomously. Ordinarily, the media gives wrong data to newspapers on the social media. E.g.: It is deceptive to give wrong data online about any individual connected with political, financial, income tax and different issues. In any case, people in many spots spread negative news against people on the social media. Although today there is less phony information via online than ever before. That implies we all should take care of morals and utilize on the social media or online morally.

5.4 Sustainability

To put our informational collection into various classes we need to divide the information or reports into various classifications. So, news or reports can be recognized in various classes. With the goal that the information can be utilized from now on and the information can be utilized for any valuable reason. That is the reason there ought to be a long arrangement, so that any kind of information can be pre-processed and partitioned into various classes by applying various algorithms. That is the reason having a plane for long-range thinking is significant.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMENDATION AND IMPLEMENTATION FOR FUTURE RESEARCH

6.1 Summary of the Study

Nowadays, due to the availability of internet, successful people of the world are now interested in getting news from online news portals. As the amount of news on online news portals like Facebook, twitter, WhatsApp, telegram, Instagram etc. is increasing day by day so is the number of readers. But all the news or data will be useful only when the news or data can be classified. So, the need to categorize news is increasing day by day. Algorithms in data science can be classified in various ways such as machine learning algorithms, deep learning algorithms etc. These are used to categorize news so that people can understand what kind of news it is by looking at the headline. In this research paper of ours also the relationship is divided into eight categories such as international, national, sports, entertainment, politics and IT. BiLSTM, GRU, Uni-Gram, Machine Learning (Logistic Regression, Multinational Naive Bayes, Random Forest Classifier, Support Vector Machine) have been used to classify these categories. BiLSTM's accuracy is 83.42%, GRU's accuracy is 80.01%. Among machine learning, logistic regression has an accuracy of 64%, Multinational Naive Base has an accuracy of 61%, Random Forest Classifier has an accuracy of 65% and Support Vector Machine has an accuracy of 65%. In Bangladesh, the amount of enews is increasing day by day. These e-news are available on different types of sites, among which the sites which regularly provide breaking news and updated news are preferred by the readers. The five popular sites are Prothom Alo, Bangladesh Pratidin, Nayadigant, Jugantar, Samakal etc. The news on these sites is usually classified, which makes these sites very popular among readers.

6.2 Conclusions

At the same rate as the amount of e-news is growing in the world, news is not being categorized. As a result, all the news cannot be used. If a large number of news can be classified then all the news can be used. Data science can be classified in various ways by applying algorithms such as machine learning algorithms, deep learning algorithms etc. BLSTM, GRU, Uni-Gram, Machine Learning (Logistic Regression, Multinational Naive Bayes, Random Forest Classifier, Support Vector Machine) algorithms are used in this research paper. The news or data is divided into eight categories such as International, National, Sports, Entertainment, Politics and IT. Among the algorithms used in the paper, BiLSTM has an accuracy of

83.42%, GRU has an accuracy of 80.01%. Among machine learning, logistic regression has an accuracy of 64%, Multinomial Naive Bayes has an accuracy of 61%, Random Forest Classifier has an accuracy of 65% and Support Vector Machine has an accuracy of 65%.

6.3 Implication for Further Study

In this paper we show how to classify a large amount of news into different categories. If a large number of news can be classified in this way, then all the news can be used in the future. Different types of algorithms in data science can be classified by applying them in different ways, such as machine learning algorithms, deep learning algorithms, etc. In the future, more such new algorithms will be used to classify data. This paper shows that only five algorithms are used, but more algorithms will be used to classify such data in the future. And in this paper, only news or data is divided into eight categories, but more categories will be divided in the future. In the future, the accuracy of the algorithms used in this paper will be increased and the news headlines can be accurately predicted.

References

- [1] A. Barua, "Multi-class Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation," *Procedia Computer Science*, vol. 193, no. 1, pp. 112-121, January 01, 2021.
- [2] M. M. Islam, "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)*, no. 1, pp. 235-239, November 2019.
- [3] E. Mohiuddin, "Multilevel Categorization of Bengali News Headlines using Bidirectional Gated Recurrent Unit.," *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, no. 1, pp. 1-6, July 2021.
- [4] M. R. Bhuiyan, "An Approach for Bengali News Headline Classification Using LSTM.," *Emerging Technologies in Data Mining and Information Security*, no. 1, pp. 299-308, 2021.
- [5] H. A. Galal Elsayed, "A Two-Level Deep Learning Approach for Emotion Recognition in Arabic News Headlines.," *International Journal of Computers and Applications*, vol. 44, no. 1, pp. 1-10, 2020.
- [6] R. Tudu, "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)*, no. 1, pp. 221-226, December 2018.
- [7] F. Jahara, "Automatic Categorization of News Articles and Headlines Using Multi-layer Perceptron.," *International Conference on Intelligent Computing & Optimization*, no. 1, pp. 155-166, December 2021.
- [8] R. Bogery, "Automatic Semantic Categorization of News Headlines using Ensemble Machine Learning: A Comparative Study.," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, 2019.
- [9] M. F. Ahmed Foysal, "Bengali News Classification Using Long Short-Term Memory.," *Emerging Technologies in Data Mining and Information Security*, no. 1, pp. 329-338, 2021.
- [10] K. R. Q. Yahan, "Classification Of Fake News Headline Based On Neural Networks.," no. 1, January 2022.
- [11] P. K. Singh, "Deep Learning Approach for Negation Handling in Sentiment Analysis.," vol. 9, no. 1, pp. 102579-102592, July 2021.
- [12] M. R. Hossain, "Different Machine Learning based Approaches of Baseline and Deep Learning Models for Bengali News Categorization.," *International Journal of Computer Applications*, vol. 176, no. 1, pp. 10-16, 2020.

- [13] S. Usmani, "News headlines categorization scheme for unlabelled data.," *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, no. 1, pp. 1-6, March 2020.
- [14] S. A. Khushbu, "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)*, no. 1, pp. 1-6, July 2020.
- [15] R. Wang, "Machine learning approach to augmenting news headline generation," *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts.*, no. 1, May 2014.
- [16] S. S. Hossain, "Context-based news headlines analysis: A comparative study of machine learning and deep learning algorithms.," *Vietnam Journal of Computer Science*, vol. 08, no. 1, pp. 513-527, November 2021.
- [17] A. P. Santos, "Sentiment classification of Portuguese news headlines.," *International Journal of Software Engineering and Its Applications*, vol. 9, no. 1, pp. 9-18, September 2015.
- [18] U. Saha, "Sentiment Classification in Bengali News Comments using a hybrid approach with Glove.," *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, no. 1, pp. 01-08, April 2022.
- [19] A. N. M. Jubaer, "Bangla toxic comment classification (machine learning and deep learning approach).," *2019 8th international conference system modeling and advancement in research trends (SMART)*, no. 1, pp. 62-66, November 2019.
- [20] M. Salehin, "Generating Bengali News Headlines: An Attentive Approach with Sequence-toSequence Networks.," *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, no. 1, pp. 256-261, November 2019.

APPENDICES

BiLSTM = Bidirectional Long Short-Term Memory

SVM = Support Vector Machine

GRU = Gated Recurrent Unit

ROC = Receiver Operating Characteristics

ML= Machine Learning

labony

ORIGINALITY REPORT

24%

SIMILARITY INDEX

19%

INTERNET SOURCES

18%

PUBLICATIONS

%

STUDENT PAPERS

PRIMARY SOURCES

| | | |
|---|--|----|
| 1 | dspace.daffodilvarsity.edu.bd:8080 Internet Source | 5% |
| 2 | Ettilla Mohiuddin, 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 Publication | 1% |
| 3 | dif7uuh3zqcps.cloudfront.net Internet Source | 1% |
| 4 | link.springer.com Internet Source | 1% |
| 5 | www.arxiv-vanity.com Internet Source | 1% |
| 6 | www.researchgate.net Internet Source | 1% |
| 7 | htmtpdf.herokuapp.com Internet Source | 1% |