

**DAFFODIL INTERNATIONAL UNIVERSITY**  
**DHAKA, BANGLADESH**



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

**BENGALI NEWS HEADLINE CATEGORIZATION USING ML APPROACH**

**BY**

**MD. SHAJADUR RAHMAN**

**ID:182-15-11723**

Supervised By

**Mr. Md. Sadekur Rahman**

Assistant Professor

Department of Computer Science and Engineering

Daffodil International University

Co-Supervised By

**Md. Abbas Ali Khan**

Assistant Professor

Department of CSE

Daffodil International University

**January 2023**

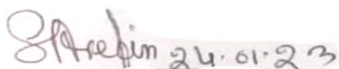
## **APPROVAL**

This Project titled “**Bengali News Headline Categorization Using ML Approach**”, submitted by Md. Shajadur Rahman to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment 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 2023.

### **BOARD OF EXAMINERS**

---

**Professor Dr. Touhid Bhuiyan**  
**Chairman**  
**Professor and Head**  
Department of CSE  
Faculty of Science & Information Technology  
Daffodil International University

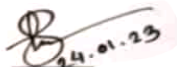


Signature: Touhid Bhuiyan, 24.01.23

---

**Dr. Mohammad Shamsul Arefin**  
**Professor**  
Department of CSE  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**

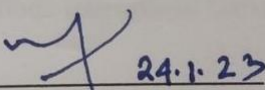


Signature: Shamsul Arefin, 24.01.23

---

**Md. Sabab Zulfiker**  
**Senior Lecturer**  
Department of CSE  
Faculty of Science & Information Technology  
Daffodil International University

**Internal Examiner**



Signature: Sabab Zulfiker, 24.1.23

---

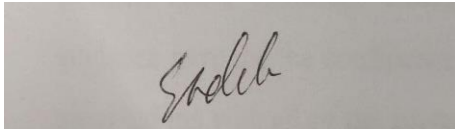
**Dr. Ahmed Wasif Reza**  
**Associate Professor**  
Department of Computer Science and Engineering  
East West University

**External Examiner**

## DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mr. Md. Sadekur Rahman, Assistant Professor, Dept. of CSE**, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

### Supervised by:

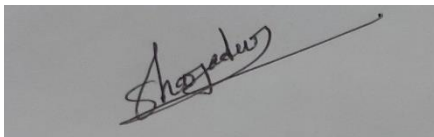


### **Mr. Md. Sadekur Rahman**

Assistant Professor

Department of Computer Science and Engineering  
Daffodil International University

### Submitted by:



### **Md. Shajadur Rahman**

ID: 182-15-11723

Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First and foremost, I appreciate God for guiding me down the path to earning an honorable B.Sc. in Computer Science & Engineering from Daffodil International University. Then I'll always be grateful to my parents for their love and support.

I'd like to express my gratitude to **Supervisor Mr. Md.Sakedur Rahman, Assistant Professor**, Department of CSE Daffodil International University, Dhaka for providing me with the necessary guidance and advice in order to accomplish this fantastic research project on “ **Bengali News Headline Categorization Using ML Approach**“. His encouragement and guidance gave me the confidence I needed to accomplish my research assignment correctly. He provided me with all of the necessary materials and knowledge to begin this investigation from beginning. I'd want to express my gratitude to my coworkers for their assistance in shaping the dataset and other relevant duties.

I am also grateful to my **Co-Supervisor Md. Abbas Ali Khan, Assistant Professor**, Department of CSE Daffodil International University, Dhaka whose advice was a huge help in accomplishing the project's aim.

## **ABSTRACT**

One of the most well-liked applications of natural language processing is text classification. Bengali is becoming more and more popular in this subject, much as many other languages, and the most well-known effort here is the categorization of various unlabeled news items. categories, such as national, international, IT and so on. Bengali news portals are becoming more and more prevalent today. The ease of access to web has made browsing news online a common activity.

The news site features a variety of news categories. This article presents a technique for categorizing news headlines from websites or news portals. An algorithm for machine learning makes predictions. Many of the gathered data were tested then trained. As which was before activities like tokenization, number removal, exclamation mark withdrawal, sign removal, and stop-word elimination are completed. Additionally, a list of stop phrases is manually prepared. Effective stop words improve performance. Stop words elimination is the most important factor in feature choice. Instead of analyzing news items from various online publications, this study focuses on categorizing Bengali News Headlines. There are eight different types of news. This work is being considered, and the news headlines are being utilized to categorize it. The model is used to model the input data. The overall model was attained the best performance by the GRU method. The height of the accuracy consisted of in case 84%.

## TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	i
Declaration	ii
Acknowledgments	iii
Abstract	iv
<b>CHAPTER</b>	
<b>CHAPTER 1: INTRODUCTION</b>	<b>1-3</b>
1.1 Introduction	1-2
1.2 Motivation	2
1.4 Research questions	2
1.5 Expected output	3
1.6 Report layout	3
<b>CHAPTER 2: BACKGROUND STUDIES</b>	<b>4-9</b>
2.1 Introduction	4
2.2 Related work	5-8
2.3 Research summary	8
2.4 Scope of the problem	7
2.5 Challenges	8-9
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>9-18</b>
3.1 Introduction	9-11
3.2 Classifier Algorithm	12-13
3.3 Research Subject and Instrumentation	14-15
3.4 Preprocessing	15-17
3.5 Stop Word Remove	18
3.6 Tokenization	18

3.7 Implementation requirements	18
<b>CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION</b>	<b>18-25</b>
4.1 Introduction	18
4.2 Model Performance	19-25
4.3 Summary	25
<b>CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY</b>	<b>25-27</b>
5.1 Impact on Society	25
5.2 Impact on Environment	25-26
5.3 Ethical Aspects	26
5.4 Sustainability	27
<b>CHAPTER 6: CONCLUSION AND FUTURE WORK</b>	<b>27-28</b>
6.1 Conclusion	27
6.2 Recommendations	28
6.3 Implication for further study	28
<b>REFERENCES</b>	<b>29-31</b>

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 1: Model Flow Chart	10
Figure 2: LSTM cell structure	11
Figure 3: Data Preprocessing Method	15
Figure 4: Data Distributions	16
Figure 5: Dataset Statistics	17
Figure 6: ROC Curve model performance	19
Figure 7: Random Forest Roc Curve	22
Figure 8: ROC curves of ML Classifiers	23
Figure 9: Training and Validation Accuracy & Loss Of RNN	23



## LIST OF TABLES

<b>TABLES</b>	<b>PAGE NO</b>
Table 1.0: Highest accuracy evaluation	20
Table 1.1: The comparison between classifiers	20
Table 1.2: The comparison between Deep learning model	21
Table 1.3: Comparative Analysis (CA)	23-24

## **LIST OF ABBREVIATION**

NLP: Natural Language Process

ML: Machine Learning

RF: Random Forest

SVM: Support vector Machine

DT: Decision Tree

XGB: XGBoost

GBC: Gradient Boost Classifier

ABC: Ada Boost Classifier

RNN: Recurrent Neural Network

LSTM: Long-short Term Memory

BI-LSTM: Bidirectional Long-short Term Memory

GRU: Gated Recurrent Unit

BERT: Bidirectional Encoder Representations from Transformers

# Chapter 1

## Introduction

### 1. Introduction

Various techniques aid the NLP system in comprehending text and symbols. The practice of classifying a text into a set of terms is known as text clarification [1]. Text categorization, often known as classification, is a method of categorizing articles into one or more preset groups [2]. It is the challenge of categorizing free-text texts into predetermined categories. It can give theoretical perspectives on document collections and has practical applications [3]. It enables users to find information more quickly by allowing them to search solely inside the categories rather than the complete information space. When the information is too large in terms of volume, the need of categorizing text becomes even more obvious. There are several studies on news headline classification systems for various languages. There are, however, a few pieces for the Bangla newspaper. As a result, we created a technique for categorizing news in Bangla newspapers. This research will aid in the development of an autonomous system by introducing machine learning-based categorization algorithms. Classifiers are created (or trained) with a collection of training documents in these approaches. Following that, the trained classifiers are used to assign documents to the appropriate categories. We picked the domain of online news from the large amount of information available on the internet since we saw that existing news websites lack effective search capability based on particular categories and do not enable any sort of visualization to evaluate or comprehend data and trends. The fact that news data is constantly published and cited makes the issue much more pressing. This prompted us to design a system that caters to two categories of users: the newsreader who is interested in pursuing news other works have become much more popular and helpful towards the evolution of the computer science.

We offer related works in Section 2, datasets in Section 3, methods in Section 4, and data visualization in Section 5. The complete result analysis is reported in section 6, and the online interface of our system is demonstrated in section 7. The final section of the report also includes our conclusion and future plans for this project.

## **1.2 Motivation**

Scientists have always been interested by the potential to construct and work on languages in order to understand and predict human behavior when it comes to Narrative study. This necessity can only be comprehended via the use of words. That is why I choose this study genre to focus on for future progress. Sentiment determination towards various words is a very natural and continuous method in terms of sentiment detection presently. By utilizing machine learning to distinguish between negative and positive evaluations, we can foresee a liberal approach for those terms. That is our primary motivation for performing this job.

## **1.3 Research Questions**

- What precisely is virtual communities??
- What is headlines Categorizations?

## **1.4 Expected Output**

The model, which comprises of several separate algorithms that must be trained and then assessed utilizing linguistic Bengali text data acquired from many social networks, was projected to be able to detect which attitudes are unfavorable and which are good. When working with text data, algorithms are frequently utilized. As we all know, dealing with text data in Bengali may be

difficult for actual machine learning models; in this instance, the dataset was properly preprocessed so that the output does not vary depending on the criteria, which are the trash values.

## **1.5 Report Layout**

Six sections make the report. Each chapter details many facets of it - the "**News Headline Categorization**". Every chapter has different parts described in detail.

### **Chapter 1: Introduction**

The inspiration is clarified and the proposition objective and introduction are presented.

### **Chapter 2: Background Studies**

The applicable work is talked about and significant popular techniques are introduced corresponding related work.

### **Chapter 3: Research Methodology**

Presents the information assortment, information pre-handling, and the element determination methodology.

### **Chapter 4: Design Specification**

the philosophies for assessment grouping are clarified and the result discussed.

### **Chapter 5: Implementation**

The 3-assessment plan, the precision assessment, and the investigation are introduced.

### **Chapter 6: Conclusion and Future Scope**

The end is drawn and my commitments are portrayed.

## **CHAPTER 2**

### **Background Studies**

#### **2.1 Introduction**

In the subject of data mining, "Text Mining" has recently received attention due to the widespread involvement of various researchers conducting research in this particular sector. The technique of searching through large amounts of digital text to uncover relevant and accurate information is called text mining. The use of this topic in several fields demonstrates how important it is; for example, it has become a significant component of machine learning, which makes use of knowledge discovery techniques to generate logical rules for categorizing text. So now we will develop a model that can categorized news headlines. Means it will classify the class of that news headline. Researchers that work with actual data benefit from classification techniques. At a period when technological means were few, researchers conducted some of the most daring research ever. Some researchers have had success with machine learning classifiers, while others have received RNN access. This section discusses relevant work that has a high level of accuracy on the classifiers we've employed, as a source of inspiration.

#### **2.2 Related Work**

Pranshengit Dhar and Md. Zainal Abedin applied the best machine learning concepts [6]. As machine learning classifiers, they employed SVM, Naive Bayes, and Adaboost. They were able to attain an accuracy rate of roughly 81%. Sheikh Abu jar proposed a neural network-based Bengali news multi-classification system with comparable performance [7]. They prepare over 86 thousand news headlines. As machine learning techniques, they employed SVM, NB, Random Forest, Logistic Regression, and Neural Network. They were able to reach an accuracy of around 90% using Neural Network methods. Bjorn Gamback focused on text categorization for hate speech [8]. The convolutional neural network is something he wants to emulate. With the help of CNN, they were able to attain a score of 86.68 percent. They use a different method of word embedding that raises this number by 7.3 percent when using the softmax function and

max pooling. Even so, the values are raised immediately. Word embedding is required to prepare the data for analysis. According to Roger Alan Stein, word embedding minimizes the system's worst performance [9]. Amin Omidvar employed clickbait web data from the media in their research [10], which they subsequently analyzed with such a machine learning classifier and a neural network. The main objective of emotion research is to divide the task into positive or negative amplitude in order to separate parental attitudes or details. The purpose of this study is to improve customer penetration, revenue, and branding. Techniques, as well as various fields including finance, economy, and some spam detecting stock exchange, purchasing and selling goods, as well as several other businesses. Since they can react quickly and provide people the opportunity to profit from the required behavior or decision-making, effective intuition analyses might have a significant impact in many fields, including policy, governance or organization, campaigns, and enterprises. Cost-effectively may be acquired neural networks. many emails, comments, and assessments totaling thousands. Text classification techniques should be broadened to include all sizes of businesses. There are several critical situations that organizations must be aware of and act upon as promptly and effectively as feasible. To be able to swiftly identify crucial characteristics, computer information retrieval should often and in real-time mimic the designer labeling. The idea of text classification is not new in the field of natural language processing. The Bangla text is still being worked on, nevertheless. Online news is categorized in this industry in a wide variety of ways. In the era of online news sources, people depend on this issue. The suggested study, which is based on the Bangla language, has as its objective this categorization. Certain Bangla datasets make use of some of the study materials that are included in our literature survey section. Compared to other machine learning approaches, our hybrid modeling approach is more successful developed by Zellers et al. to assess the accuracy of media from the news text. To extract the content-based characteristics from news writings, additional publications [11, 12] take into account lexicons, bag-of-words, TFIDF, and latent themes. An LSTM network-based false news detection method called DEFEND is proposed by Cui et al. [13]. The DEFEND takes into account comment threads to determine whether certain news is true or not. In order to learn the models of social situations, Nguyen et al [14] proposed false news detection technique FANG makes use of the graph learning architecture. The techniques covered above are recognized as gold standards in the study of false news. Instead of early false information identification, the country predominantly focuses on based on deep learning approaches. A few publications [15, 16] suggest early false news identification. Rubin et al methodology's [17] was suggested to identify British news

assessment. 360 satirical news articles from primarily 4 categories were examined and evaluated: civics, science, business, and delicate news. On the basis of their parody news research, they created an SVM classification model. The 2016 US surveys are the most noteworthy illustration of how fake news has spread swiftly during the previous 10 years [28]. The increasing spreading of false information online has caused many problems, not just in politics but also in a wide range of other industries such sports, medicine, and academics [19]. In order to extract themes from the a media platforms corpus in 2016, a useful theme controller is designed on LDA was created .

In order to extract themes from the a media platforms corpus in 2016, a useful theme controller is designed on LDA was created [20]. A collection of 90,527 textual documents with an airline and airport management focus was employed in this investigation. Online platforms have surpassed print media as a major contributor of current affairs for the general public [21, 22]. Through these venues, people may now openly share their thoughts and leave comments. Just Facebook is the source of news for almost 35% of Americans [23]. Numerous individuals, groups, and sites on Facebook offer their opinions about current global events. The risk of creating fake news exists because of this freedom of speech. People have been exposed to false information since there is no verification of the veracity of news stories posted on Facebook, that has an effect on society [24]. The Twitter and Facebook also take the correct actions to combat false news in addition to machine learning techniques. Twitter and Facebook regularly delete users, webpages, and organizations that spread inaccurate info [25]. People can indeed be useful in identifying any types of misleading info they come across on social networking sites. Every posting, tweet, or remark which might seem deceptive to them can be reported. Tools and extensions have also been created to recognize and gather bogus news [26, 27]. Vosoughi et al. [28] take a novel technique to investigate the characteristics of news diffusion on social media; specifically, they explain how news (rumors) travel on Twitter and examine how false news varies from credible news in term of its Twitter dissemination. The paper explores the spread of disinformation online using a variety of analysis techniques, including depth, size, highest broadness, systemic popularity, actually imply broadening of correct and incorrect rumor spirals at different depths, quantity Twitter accounts attained at any deep, and time in mins for correct and incorrect rumor feedback loops to reach detail and amount of Twitter users. To conduct experiments, Iskandar et al. [29] gathered information from Fb, Instagram, and several blogging websites. They demonstrated how Nave Bayes is the ideal algorithm for their task since it deals with probability



by critically analyzing several methods [30]. According to Johnston et al. [31], a machine-learning algorithm that can categorize Sunni-related misinformation. To identify bogus news in Bangla, Hossain et al. [32] create a dataset of 50 k occurrences. They have done a thorough analysis of both language and computer having to learn aspects. Utilizing SVM with such a linear kernel, a method proves how to find threatening and offensive Bengali terms in social media [33]. The model ran experiments on 5644 txt files, and it was able to predict words with a high precision of 78%. To use a classification and boolean classification, Dinakar et al. [34] created a library of Comment sections for the purpose of identifying textual cyberbullying. Utilizing SVM, lexical, word and tf-idf characteristics, a new method for identifying hateful speech in Indonesian has been described [35]. a technique to identify offensive material and cyber abuse on Chinese social media. Their model, which combined LSTM with user-specific behavioral and character variables, had a 95 percent accuracy rate. Hammer [36] proposed a technique for identifying threats and acts of violence directed against minority groups in internet debates. This study took into account manually labeled sentences that included bigram properties of important words.

## **2.3 Research Summary**

The study's information was acquired at randomly from multiple websites and platforms on the online. Extraneous data, numeric values, and special characters were removed from the dataset before to detection to produce a full and accurate detection result. The dataset included multiple repeat occurrences of various numbers and phrases, which were thoroughly evaluated and deleted for performance reasons. And we were hopeful to achieve the expected variable and outcome, which was very much likely to be helpful because of our research study.

## **2.4 Scope of the problem**

Information, stories, or forgeries that are intended to mislead viewers in order to profit from their agreement or viewpoint on economic, social, religious, or other matters are referred to as fake news. These articles are frequently written to change people's opinions, forward a political goal, or place people in perilous circumstances, and they may bring in money for internet news outlets. On occasion, websites that have the same names and web links as reputable news sources may

spread false information. In order to mislead readers and propagate false information through social media and the internet, deepfakes and channels employ their fake news material.

## **2.5 Challenges**

We examined the slant assessment relying on voyager input in regards to carrier organizations in this investigation. According to our proposed method, both element determination and over-inspection procedures are equally relevant in terms of increasing our outcomes. Using highlight selection methods, we were able to recover the best subset of highlights while reducing the number of calculations necessary to generate our classifiers. It has, however, decreased the lopsided appropriation of classes found in a large portion of our smaller datasets without causing overfitting. Our findings show that the proposed model has great grouping precision in anticipating events structure the two groups positive, negative data. Organizing Bengali text and processing it for model training was also a significant problem.

## CHAPTER 3

### Research Methodology

#### 3.1 Introduction

Here's how to do it: Data is gathered from a number of Bangla newspapers. For scraping news from the website, we utilized the Python module BeautifulSoup. We eliminate superfluous symbols from datasets after gathering data and summarize the datasets. This section contains information on the number of words, documents, and unique words per class. From the clean datasets, we calculate the length frequency distribution. The datasets for the model must then be prepared. We trained using 80% of the data and tested with 20% of the data. Then, using an encoded sequence, label the data. With 10 epochs and 64 batch size, I trained the model. As a result, our model's data is ready. To forecast news headlines and compare the outcomes, we deployed two deep learning systems. I used 10 epochs and batch size 64 to train the model. As a result, data for our model has been prepared. Long short-term memory (LSTM) and Gated recurrent unit are two deep learning algorithms used to forecast news headlines and compare the outcomes (GRU). We discovered accuracy, Precision, recall, and F1 score are all obtained from these models. Following that, the outcomes will be compared.

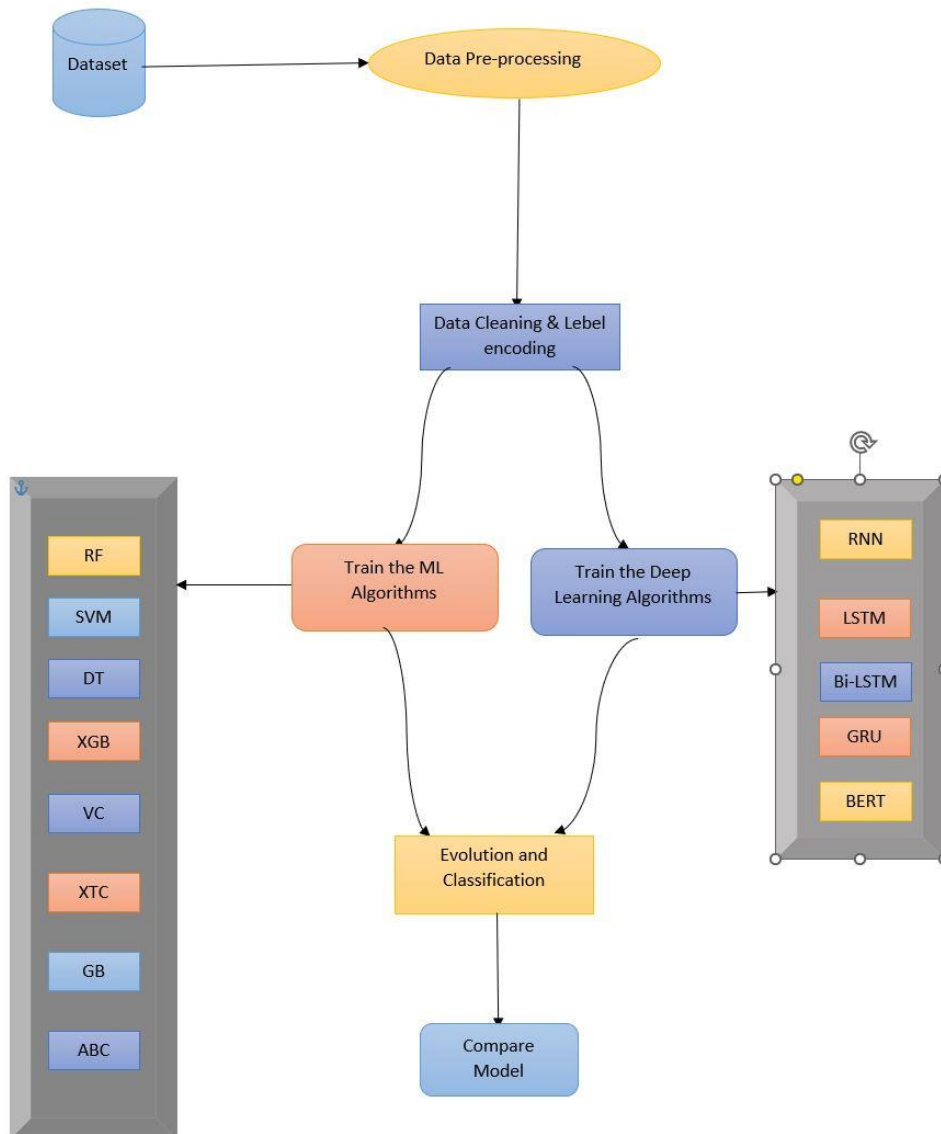


Figure 1. Model Architecture

And if we look at the deep learning approach, the identification of fake news is a binary classifier issue. By analyzing the bias that is inherent in authored news stories and examining the relationship between both the title with body of the item, the suggested approach for spotting fake news evaluates the authenticity of the findings in the piece.

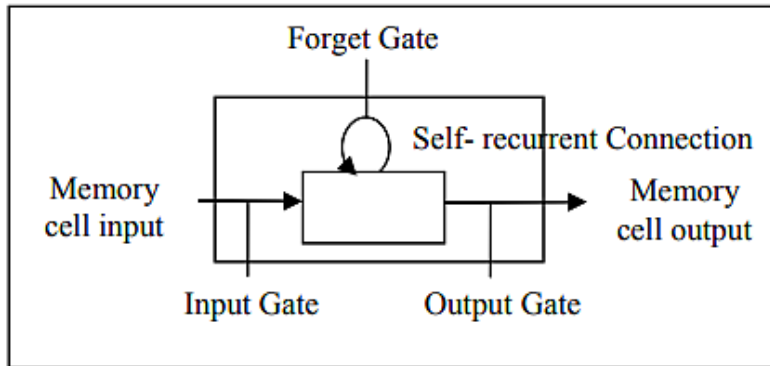


Fig 2. LSTM cell structure

The model is given the word embedding output. The ml model used in this application is a combination of iterative and incremental with a first layer that embeds values for vocabulary size, number of features, and sentence length. The next step is LSTM, which has 100 neurons per layer, followed by a dense layer with sigmoid function as we only need one output in the end. In order to prevent overfitting, we have employed binary cross variance to calculate loss, Adam algorithm for adaptive estimation, and ultimately adding a fall out layer in between. The model was then trained and tested. For both of the pre-assembled testing data sets, the result is predicted. If the predicted value is larger than 0.5 and less than 0.5, it is classified as 1 and as less as 0.  $(TP+TN)/Total$  is a measure of accuracy. The following terminology was employed: True Negative (TN), which signifies that the test cases and prediction were in fact unfavorable; True Positive (TP), which denotes that the forecast and test instances were both accurate; False Negative (FN), which happens when a prediction is made but the results of the testing process are positive; When a prognosis is produced but the testing process turn out be negative, it is known as a false positive (FP). The architecture employed in LSTM is as follows: input consists of input shape, embedding vector features, and vocabulary size. Dropout value 0.2 to prevent overfitting, SpatialDropout1D layers using parameter 0.4, 256 LSTM units, a thick layer containing two neurons with soft max activation. And in case of Bi-LSTM the Dropout value was same, Bi directional LSTM units were 64. The RNN was comprehended on the similar consequences and was tuned according to the datasets preference.

### 3.2 The Classifier Algorithms

A preceding classification technique for literature and data gathering is called the Decision Tree Classifier (DTC). In several disciplines, DTC is a powerful tool for categorization. Making a tree which supported a attribute for classified data points is the main idea. The major problem with a DTC is that some features or attributes may be more appropriate for children than for parents. To remedy this drawback, a quantitative modeling was employed to feature selection inside the tree. For a training data set that consists of  $n$  negative and  $p$  positive values.

$$H\left(\frac{p}{n+p}, \frac{n}{n+p}\right) = -\frac{p}{n+p} \log_2 \frac{p}{n+p} - \frac{n}{n+p} \log_2 \frac{n}{n+p}$$

The training set  $E$  splits into the prefixes of "E1," "E2," "...," and "Ek," and you select  $K$  in the characteristic with the unique value. The anticipated entropy (EH) will continue to stay after the effort inside the attribute (branches  $I = 1, 2, \dots, k$ ), including

$$EH(A) = \sum_{i=1}^K \frac{p_i + n_i}{p + n} H\left(\frac{p_i}{n_i + p_i}, \frac{n_i}{n_i + p_i}\right)$$

MNB is a variant of NB classifiers, which are refers generally built on the Bayes theorem. It takes into account how many times a team has practiced using the training dataset. For text categorization, MNB outperforms Bernoulli Naive Bayes (BNB) [49].

In the period of instruction, RF generates several random decision trees. It gathers the data out of each tree throughout projection and then delivers the outcome that most of the trees have correctly anticipated. The RF classifiers utilized in this writer's studies have a forest of 10 trees.

Famous machine learning (ML) algorithm Support Vector Machine produces an ideal hyperplane to classify fresh examples from testing phase. Various kernels are used by SVM classifiers to create the best hyperplane. Kernels are processes that look for patterns. The studies in this paper employ one of the SVM kernels.

There are two distinct node kinds in the decision tree: interior and exterior. While internal nodes hold the characteristics necessary for classification, outer nodes indicate the decision class. The clustering algorithm was assessed using a top-down method that divided homogenous data into subgroups. The uniformity of samples is defined by its entropy, which is computed using the equation.

$$E(S) = \sum_{i=1}^n p_i \log_2 p_i$$

Here,  $p_i$  is the likelihood that a sample will belong to the training course, while  $E(S)$  stands for the graph's entropy. Entropy was utilized to assess the split's quality. Every characteristic was taken into account while deciding on the appropriate split for each node. Possible combination of the characteristics is controlled by randomized state 0.

The decision trees that make up the Random Forest (RF) each function independently. The most likely conclusion line is identified using the "Gini index" from each branch. This index was determined using equation.

$$\text{Gini} = 1 - \sum_{i=1}^c (p_i)^2$$

Here,  $c$  stands for the overall number of classes, while  $p_i$  denotes the likelihood of each class. Inside the forest, where the 'Gini' scale is used to assess the level of split, we utilized 100 trees. If there are at least two inner nodes as well as all 's extremely taken into account within every component, the network nodes are separated.

### 3.3 Research Subject and Instrumentation

The title I've picked is "A machine Learning And Deep Learning Approach for Bengali News Headline Categorization." This is a critical topic of research in Natural Language Processing. So far, I have investigated the approach for conducting estimation research in Bangla using a specific and theoretical strategy. A superb learning model requires a powerful computer and a plethora of equipment. An example of a concept analyzer is shown below the main instrument for this model.

#### Hardware and Software:

- 2GB RAM and Intel core i7 7<sup>th</sup> generation 2.4ghz.
- 500 GB Hard Disk.

#### Tools:

- Windows 11
- Python 3.10
- Jupyter Notebook
- NLTK
- Pandas
- Numpy

### 3.4 Preprocessing

We wanted to refresh the entire dataset by deleting the trash character in order to prepare the data for the algorithms' train instance. What more characters are there except special characters ("!", "@", "#", "\$", "%", "&", "'", "\*"), number characters (1, 2, 3, 4, 5, 6, 7, 8, 9, 0), white space, and duplicate characters? Because the character a particular was reconciled numerous times when adopting the data collection, the duplicated character may be kept. In order for a computer to discern between classes, the data must be raw. As shown in the fig the dataset statistics was determined for processing the models capacitation.



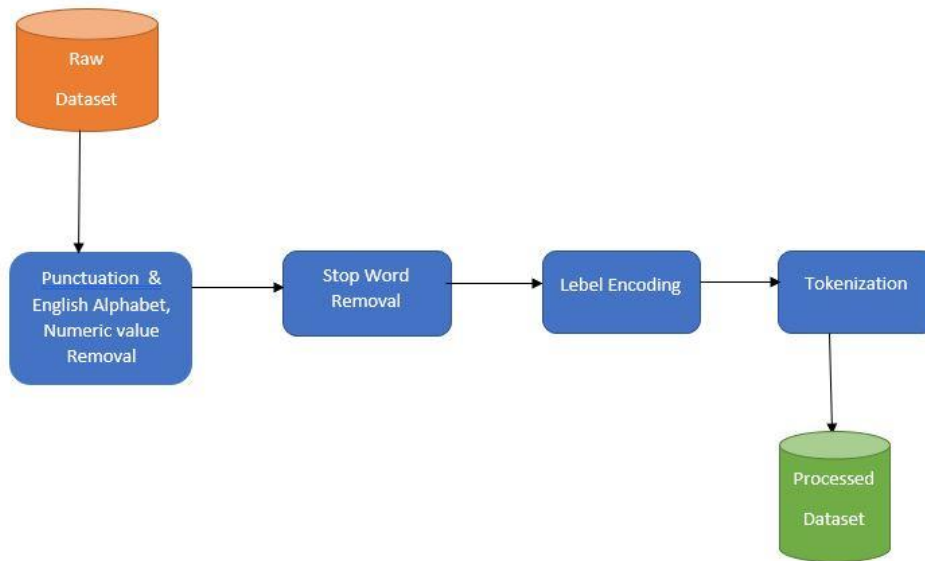


Fig 3. Data Preprocessing Method

### 3.5 Data collecting

Scraping was used to acquire data from numerous Bangla newspapers. Our collection contains over a million records. We gathered information from publications such as Bangladesh pratinidin [17], dainik jugantor [18], daily inqilab [19], kalerkantho, and others. These are Bangladesh's most widely read newspapers. We gathered data from these newspapers, which aids in determining which kinds of data are most frequently accessed by readers. For collecting data from websites, we utilized Chrome Web Scrapper and Python tools. In our dataset, there are three columns. These are the headlines, the category, and the name of the newspaper. The data is open to the public. 1 The following graph depicts the headline dispersion of each category. This set of data is unbalanced.

The dataset is depicted in the diagram below:

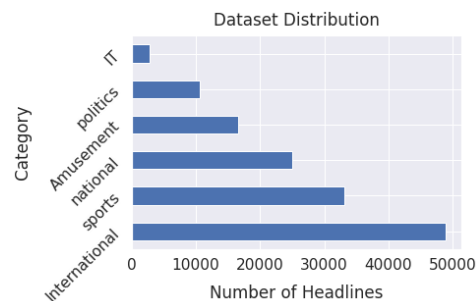


Fig-4: Dataset Distribution

### 3.6 Cleaning of data

Because the headlines are short in length, it is not necessary to eliminate the stopwords [20]. Regular expressions are used to eliminate redundant data from our sample. The data sample would look after just cleaning.

Original: ক্ষমা চেয়েও মুক্তি পেলেন না পরিচালক গাজী মাহবুব

Cleaned: ক্ষমা চেয়েও মুক্তি পেলেন না পরিচালক গাজী মাহবুব

Original: ব্র্যান্ডউইথের ব্যবহার ৮০০ জিবিপিএস ছাড়িয়ে

Cleaned: ব্র্যান্ডউইথের ব্যবহার ৮০০ জিবিপিএস ছাড়িয়ে

Original: জামিনে মুক্তি পেলেন ছাত্রদল সভাপতি

Cleaned: জামিনে মুক্তি পেলেন ছাত্রদল সভাপতি

Original: দ. কোরিয়ায় ১০০টি খালি কফিন পাঠিয়েছে যুক্তরাষ্ট্র

Cleaned: দ কোরিয়ায় ১০০টি খালি কফিন পাঠিয়েছে যুক্তরাষ্ট্র

Original: ফ্লোরিডায় হামলাকারী ‘মানসিকভাবে অসুস্থ’: ট্রাম্প

Cleaned: ফ্লোরিডায় হামলাকারী মানসিকভাবে অসুস্থ ট্রাম্প

Original: সেরাটা দিতে পারলে সিরিজ জিতবে বাংলাদেশ: মাশরাফি

Cleaned: সেরাটা দিতে পারলে সিরিজ জিতবে বাংলাদেশ মাশরাফি

After cleaning the data, we may choose the appropriate headline length to utilize in order to make each headline the same length. The greatest, lowest, and average lengths of headlines are shown in Figure 4. In addition, each category has a large number of terms. From each category, we choose words that are both distinct and related. This is known as data statistics, as seen in Figure 5:

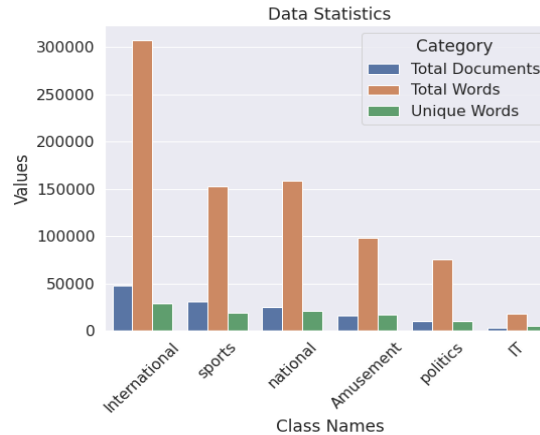


Fig-5: Dataset statistics

### 3.7 Stop word remove

A stop word is a widely used phrase, such as ".", ",", ":", "|", and so on, that a web index has already been designed to disregard while sorting portions for viewing and retrieving those as the result of a pursuit query. I don't need these keywords to eat up valuable database space or time to process. I may efficiently evacuate individuals for this purpose by keeping a list of keywords that we consider are stop words. Python's NLTK (Nltk Toolkit) has a collection of stop words from 16 distinct dialects. They may be found in the nltk information search.

### 3.8 Tokenization

Tokenization is the process of isolating the arguments from the phrase and referring to these single arguments as tokens. In such approaches, tokenization is necessary for learning the algorithm's input. The data was divided into two groups for labeling: positive and negative.

### 3.9 Implementation Requirements

We preferred Python as a programming language to use the machine learning technique. The Panda library is used to load the data, while a NLTK package is used for preprocessing. The complete implementation is written in Python in a jupyter environment.

## CHAPTER 4

### Experimental Results and Discussion

#### 4.1 Introduction

In the instance of our identification of negative and positive outcomes, the algorithms were able to achieve a very volatile and appreciating level of result. The methods were chosen and placed on our dataset frequency because, in Natural Language Processing, the dataset is the most important aspect of the entire operation. The algorithms we used were Linear Regression, Decision Tree, Random Forest, Multinomial Naive Bayes, GRU, Linear SVM, and RBF SVM. These are all advanced machine learning approaches that produced the desired outcome. And then for comparison purpose we have experimented with the deep learning approaches which were LSTM, Bi-LSTM, GRU methods. Then we have demonstrated a comparison perspective.

#### 4.2 Model Performance

In the instance of our identification of negative and positive outcomes, the algorithms were able to achieve a very volatile and appreciating level of result. The methods were chosen and placed on our dataset frequency because, in Natural Language Processing, the dataset is the most important aspect of the entire operation. The algorithms we used were Linear Regression, Decision Tree, Random Forest, Multinomial Naive Bayes, Linear SVM, and RBF SVM. These are all advanced machine learning approaches that produced the desired outcome. And then for comparison purposes we have experimented with the deep learning approach. Here, we have demonstrated our result diagrams, in which we implemented our dataset in order to achieve the highest accuracy. The dataset was contemplated into training for 80/20 split. The test case contains in total 20% of the total dataset. With the dataset preference we figured some of the most re-categorized curved which would represent the performance evaluation.

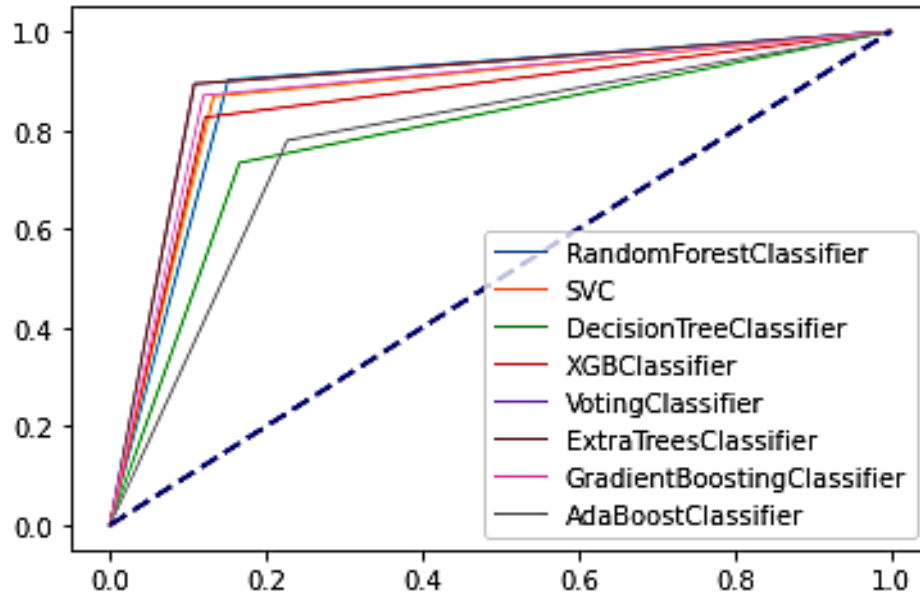


Fig 6. ROC Curve model performance

As different ML algorithms were used for the approach, we were able to acquire a quite fascinating result outcome. Each of the algorithms were working towards separating the real news from the fake news perspective.

RF	SVM	DT	XGB	VC	XTC	GB	ABC
0.87	0.86	0.77	0.85	0.88	0.89	0.89	0.77

Table 1.0 Highest accuracy evaluation

The accuracy evaluation is measured by evaluating a certain mathematical convolution measuring between the true negatives, true positive and the false negative, false positive. The term is given below,

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

The model names were then divided while the precision, recall, F1-score and Accuracy wise, the division was then recalled onto the table given below.

Model Name	Precision	Recall	F1-score	Accuracy
RF	0.88	0.87	0.88	0.87
SVM	0.87	0.87	0.87	0.86
DT	0.78	0.78	0.77	0.77
XGB	0.88	0.82	0.85	0.85
VC	0.89	0.90	0.89	0.88
XTC	0.89	0.89	0.89	0.89
GB	0.88	0.88	0.88	0.89
ABC	0.78	0.78	0.78	0.77

Table 1.1 The comparison between classifiers

Model	Real				Fake			
	Acc	P	R	f1	Acc	P	R	f1
RNN	0.94	0.96	0.96	0.96	0.94	0.86	0.85	0.85
LSTM	0.93	0.96	0.94	0.95	0.93	0.80	0.87	0.84
Bi-LSTM	0.94	0.96	0.95	0.96	0.94	0.83	0.87	0.85
GRU	0.92	0.96	0.94	0.95	0.92	0.78	0.85	0.81
Bert	0.95	0.96	0.98	0.97	0.95	0.85	0.72	0.78

Table 1.2 The comparison between Deep learning model

As from the table 1.1 we can count that the GB and XTC were able to acquire the highest of the accuracy of 89%. As for the F1 score the XTC altogether with VC with the 0.89. In case of recall the VC were capable enough to acquire the 90. In the concern of precision VC and XTC both

scored 89%. In overall performance VC and XTC were able to perform best in all average. In case of acquiring such result the reason behind it was in the dataset customization as the VC Dimension is a characterization algorithm, which can easily distinguish between such data. The client-server paradigm is used by XTC. Servers are either used as data collection interfaces for monitoring by routinely asking lists of signals or as process control, i.e., establishing specific process signals. Any client connecting over TCP can access XTC daemons, and messages are sent in ASCII format. Therefore, XTC must be viewed as highly vulnerable from a security perspective. The client-server paradigm is used by XTC. Servers are used either as data collection interfaces for monitoring by routinely asking lists of signals or as process control, i.e., establishing specific process signals. Any client connecting over TCP can access XTC daemons, and messages are sent in ASCII format. Therefore, XTC must be viewed as highly vulnerable from a security perspective.

The models' different confusion matrices were then assembled for a better understanding of the algorithm's performances.

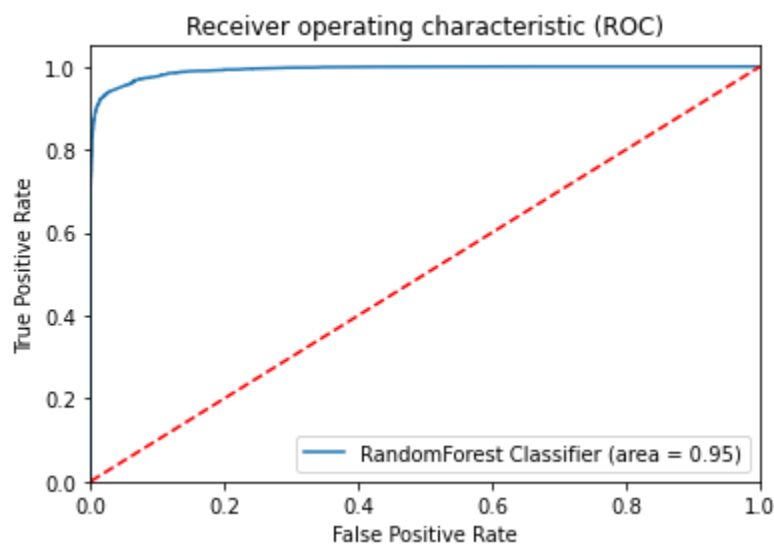


Fig 7. Random forest ROC curve

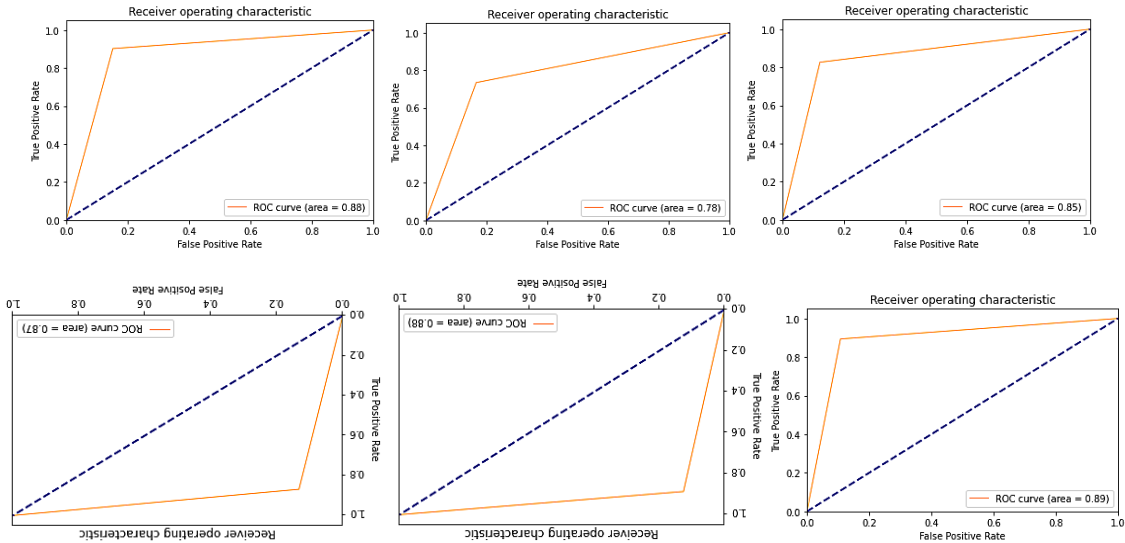


Fig 8. ROC curves of ML Classifiers

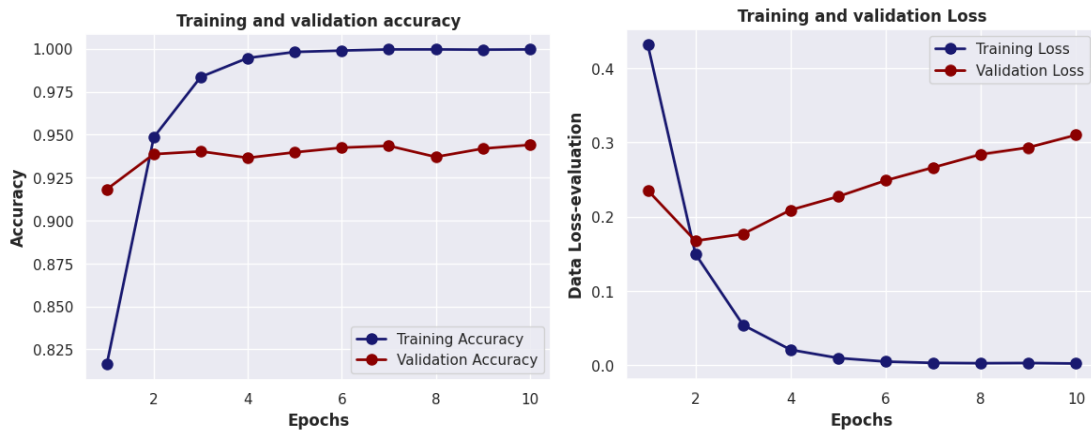


Fig 9. Training and Validation Accuracy & Loss Of RNN

In some cases, the validation accuracy drops below 85%, which is constant of the datasets multiclass maneuver. But as in our case we were working on binary classification. Which and why we got quite the acceptable training and validation accuracy,



Paper	Algorithm	Accuracy
[22] Hossain et al. worked a elaborate work on 50k dataset to detect fake news.	SVM	91%
[27] Tanvirul et al. worked to detect spam from Bengali text using.	MNB	82%
[28] In this research work Dense Neural Network was used to take out fake news.	DNN	85%
[30] In this paper they used a fifty thousand of data to analyze with a CNN-LSTM model.	CNN-LSTM	75%
[31] Shafayat et al. used a GNB method the evaluate the fake news detection.	GNB	87%
Our Proposed Model	RNN	94%

Table 1.3 Comparative Analysis (CA)

The RNN was able to get 94% of accuracy which is far more than the genuine machine learning algorithms. The main reason for such occurrences can be considered because of the dataset size. The neural network works as a net which filters the target class via layers. If the data size is greater then it can acquire more consistency. Which lacks more frequently the machine learning algorithms. These methods, however, have little chance of being widely used because they can't be adapted to other connections without a major reconfigure and a lot of labor. Instead, the primary driving force is to minimize such effort.

to employ learning-based techniques. a method that keeps track of protocol-specific sequences of thoroughly examined protocol communications employs separate Markov chains, much as we do,

to identify the same kinds of attacks which we are interested in. This finite-state machine-like model needs to be further tweaked using process- or p2p information, such the so-called "worth of relevance," to achieve an acceptable false-positive rate.

### **4.3 Summary**

Many words were rendered incomprehensible once special characters were removed, as these were also responsible for assessment in the event of sentiment expression. In some circumstances, I needed to take a different approach by re-processing them into a more positive perspective. Such circumstances compels that the model we are proposing the RNN to be our proposed model because of its high quantified accuracy achievement. The analysis report shows that with 94% accuracy the RNN model is standing atop the other analysis methods.

## **CHAPTER 5**

### **Impact on Society, Environment and Sustainability**

#### **5.1 Impact on Society**

Each individual feeling today can be linked to the words we view on a daily basis on various digital platforms. In this instance, these platforms must have a framework in place to differentiate among real feelings and manufactured antagonism. As a result, we've chosen to focus our efforts on one of the most important genres of all time: news. We intend to do this by bringing in a more distinct and diverse digital future. By which we can initiate the different concerns and rumors that impacts highly on negative basis on our contempered society.

#### **5.2 Impact on Environment**

The harm is caused by the massive distribution of bogus stories on social media. You should be more cautious if you are a person or company that delivers a lot of information, potentially with the help of life online management. It takes only seconds to increase your social media feeds. Take a look at the beginning of the story. And if you've never heard of it, search it up to see whether it's credible. If you don't have time, disregard it, especially if it appears to be satire, misleading information, or deliberate public relations. By not spreading false information, you may assist to minimize the spread of deception and news.

One of the key advantages of web-based living, as previously said, is the ability to offer information to a large number of people in a short period of time. Although this appears to be a great benefit in a crisis, perhaps it is a substantial burden because incorrect material may be disseminated in a split second. This can lead to major deception and terror. When news of the Queen's death spread, it was because queen having skipped a Yuletide banquet due to a normal sickness. This, along with other hoaxers spreading phony news, led many individuals to realize the Queen had dead.

### **5.3 Ethical Aspects**

Web-based living is one of the best ways to meet and interact with new people who share your interests since it allows you to search for groups that are interested in your perks and pastimes. This is excellent for meeting new people, but it's also wonderful for love interests and web dating, which has grown more popular than traditional face-to-face encounters owing to online life and Tinder-like characteristics.

The internet is a fantastic tool for quickly sharing news throughout the world, with "breaking news" tweets getting thousands of retweets in minutes. This may be extremely useful for updating people about necessary news such as weather reports and missing peoples.

As previously mentioned, internet connection has had a good impact on society in a number of ways, all at little added charge because all critical web-based life phases are free. Evaluate another thing or institution that has ever revolutionized your life as a result of the internet, but then consider their cost.

### **5.4 Sustainability**

- There are around 2.4 billion total dynamic web-based daily customers.
- 90% of major companies have at least dual web-based life cycles.
- In case they are unable to went into the online life catagories, 65 percent of individuals feel uneasy and uncomfortable.

## CHAPTER 6

### Conclusion and Future Work

#### 6.1 Conclusion

A machine learning-based model for news headlines was developed in this study. Bengali newspaper categorization. The majority of research in the literature takes another linguistic publication into account. For this classification technique, GRU,RNN & LSTM are the most powerful algorithms for finding a decent model. The results of the categorizations are mainly in line with previous research. Because we employed two methods for this categorization, the results might vary from one model to the next. For news categorization, we've chosen eight categories. The outcomes are independent of the categories. This approach yields a more accurate answer when there is more data, including balanced and dissimilar data. Various items of information Companies seek to classify news depending on what's been published in the newspaper. As a result, they may obtain the outcomes that they desire. Overall conclusions: further research is needed. This is a tiny dataset. As a consequence, we will be able to provide superior results if we use more than one dataset. Changing the model's characteristics also yielded various outcomes. The outcome should be altered while changing epochs. In addition, in the models, not employing the activation function causes an impact. There are several machine learning models available. Various models provide different outcomes.

#### 6.2 Recommendations

Experimental analysis is the most latest phenomenon in understanding public needs; it's a cheaper and more smart tool for studying how people feel about a particular problem and the brand influence of smaller blogging. In this example, we examined people's attitudes toward the aviation industry, as well as Airways' continuous troubles and how the public perceived them. The research confirmed our assumptions about how effective a Twitter assumption research approach is. The computation's Logistic Regression and Random Forest classifiers, together with two coding for better results, convincingly establish the mass group premise and, as a result, The airliner may easily evaluate the data to profit from this by striving to improve characteristics that are uncomfortable or disliked by the lead audience. This assignment has several choices, including:

### **6.3 Implication for Further Study**

As our works demonstrated above, we can say that our work has been a collection of various contrasts. The models we applied and the methods we followed can be the infrastructure of vary foundation of our future interest in natural language processing, machine learning and deep learning modules and experiments. The work on Bengali linguistic based datasets has been one of the toughest challenges in this field to overcome and we hope to give our full-fledged support to help obtaining the research height that would give us a new upbringing of social and counterproductive value among the progress of computer science.

## REFERENCES

- [1] Meparlad, Understanding Text Classification in NLP with Movie Review Example, AnalyticsVidhya, (2020).
- [2] Shahin, M. M. H., Ahmmed, T., Piyal, S. H., & Shopon, M. (2020, June). Classification of bangla news articles using bidirectional long short term memory. In *2020 IEEE Region 10 Symposium (TENSYP)* (pp. 1547-1551). IEEE.
- [3] Yang, Y., & Joachims, T. (2008). Text categorization. *Scholarpedia*, 3(5), 4242.
- [4] Hu, Y., Li, Y., Yang, T., & Pan, Q. (2018, November). Short text classification with a convolutional neural networks based method. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)* (pp. 1432-1435). IEEE.
- [5] Stein, R. A., Jaques, P. A., & Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. *Information Sciences*, 471, 216-232.
- [6] Omidvar, A., Jiang, H., & An, A. (2018, September). Using neural network for identifying clickbaits in online news media. In *Annual International Symposium on Information Management and Big Data* (pp. 220-232). Springer, Cham.
- [7] Cai, J., Li, J., Li, W., & Wang, J. (2018, December). Deep learning model used in text classification. In *2018 15th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)* (pp. 123-126). IEEE.
- [8] Shahi, T. B., & Pant, A. K. (2018, February). Nepali news classification using naïve bayes, support vector machines and neural networks. In *2018 International Conference on Communication Information and Computing Technology (ICCICT)* (pp. 1-5). IEEE.
- [9] Dhar, P., & Abedin, M. (2021). Bengali News Headline Categorization Using Optimized Machine Learning Pipeline. *International Journal of Information Engineering & Electronic Business*, 13(1).
- [10] Khushbu, S. A., Masum, A. K. M., Abujar, S., & Hossain, S. A. (2020, July). Neural network based bengali news headline multi classification system: Selection of features describes comparative performance. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.
- [11] Al-Tahrawi, M. M. (2015). Arabic text categorization using logistic regression. *International Journal of Intelligent Systems and Applications*, 7(6), 71.
- [12] Zia, T., Abbas, Q., & Akhtar, M. P. (2015). Evaluation of Feature Selection Approaches for Urdu Text Categorization. *International Journal of Intelligent Systems & Applications*, 7(6).
- [13] Gambäck, B., & Sikdar, U. K. (2017, August). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online* (pp. 85-90).

- [14] Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4580-4584). IEEE.
- [15] Kostadinov, S. (2017). Understanding GRU networks. *Towards Data Science. Towards Data Science, Towards Data Science, 16*.
- [16] Bangladesh protidin, <https://www.bd-protidin.com> (2021).
- [17] Doinik Jugantor, <https://www.jugantor.com> (2021).
- [18] Daily Inqilab, <https://www.dailyinqilab.com> (2021).
- [19] Elgabry, O. (2019). The ultimate guide to data cleaning. *Towards to data science*.
- [20] Vala Ali Rohani, Shahid Shayaa, and Ghazaleh Babanejaddehaki. 2016. Topic modeling for social media content: A practical approach. In *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*. IEEE, 397–402.
- [21] Alfred Hermida. 2016. Social media and the news. *The SAGE handbook of digital journalism* (2016), 81–94.
- [22] Rasmus Kleis Nielsen and Kim Christian Schrøder. 2014. The relative importance of social media for accessing, finding, and engaging with news: An eightcountry cross-media comparison. *Digital journalism* 2, 4 (2014), 472–489.
- [23] Elisa Shearer and Amy Mitchell. 2021. News use across social media platforms in 2020. (2021).
- [24] Margaret Van Heekeren. 2020. The curative effect of social media on fake news: A historical re-evaluation. *Journalism Studies* 21, 3 (2020), 306–318.
- [25] Yoel Roth and Del Harvey. 2018. How Twitter is fighting spam and malicious automation. *Twitter [blog]*, June 26 (2018).
- [26] Agata Giełczyk, Rafał Wawrzyniak, and Michał Choraś. 2019. Evaluation of the existing tools for fake news detection. In *IFIP International Conference on Computer Information Systems and Industrial Management*. Springer, 144–151.
- [27] Kai Shu, Deepak Mahudeswaran, and Huan Liu. 2019. FakeNewsTracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory* 25, 1 (2019), 60–71.
- [28] H. Ahmed, I. Traore, and S. Saad, “Detection of online fake news using n-gram analysis and machine learning techniques,” in *Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pp. 127–138, Springer, Vancouver, Canada, 2017.
- [29] Iskandar, B. Terrorism detection based on sentiment analysis using machine learning. *J. Eng. Appl. Sci.* 2017, 12, 691–698
- [30] Sarker, I.H. A machine learning based robust prediction model for real-life mobile phone data. *Internet Things* 2019, 5, 180–193. [CrossRef]
- [31] Johnston, A.H.; Weiss, G.M. Identifying Sunni extremist propaganda with deep learning. In *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, Honolulu, HI, USA, 27 November–1 December 2017.
- [32] Hossain, M.Z.; Rahman, M.A.; Islam, M.S.; Kar, S. BanFakeNews: A Dataset for Detecting Fake News in Bangla. *arXiv* 2020, arXiv:2004.08789.



[33] Chakraborty, P.; Seddiqui, M.H. Threat and Abusive Language Detection on Social Media in Bengali Language. In Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, 3–5 May 2019.

[34] Dinakar, K.; Reichart, R.; Lieberman, H. Modeling the detection of textual cyberbullying. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, 17–21 July 2011

[35] Aulia, N.; Budi, I. Hate Speech Detection on Indonesian Long Text Documents Using Machine Learning Approach. In Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence, Bali, Indonesia, 19–22 April 2019.

[36] Hammer, H.L. Detecting threats of violence in online discussions using bigrams of important words. In Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference, The Hague, The Netherlands, 24–26 September 2014.

## BENGALI NEWS HEADLINE CATEGORIZATION USING ML APPROACH

### ORIGINALITY REPORT

<b>25%</b> SIMILARITY INDEX	<b>19%</b> INTERNET SOURCES	<b>13%</b> PUBLICATIONS	<b>15%</b> STUDENT PAPERS
--------------------------------	--------------------------------	----------------------------	------------------------------

### PRIMARY SOURCES

<b>1</b>	<b>dspace.daffodilvarsity.edu.bd:8080</b> Internet Source	<b>4%</b>
<b>2</b>	<b>www.mecs-press.org</b> Internet Source	<b>4%</b>
<b>3</b>	<b>Submitted to Daffodil International University</b> Student Paper	<b>1%</b>
<b>4</b>	<b>Submitted to University of Witwatersrand</b> Student Paper	<b>1%</b>
<b>5</b>	<b>Submitted to Letterkenny Institute of Technology</b> Student Paper	<b>1%</b>
<b>6</b>	<b>dergipark.org.tr</b> Internet Source	<b>1%</b>
<b>7</b>	<b>wpsites.ucalgary.ca</b> Internet Source	<b>1%</b>
<b>8</b>	<b>Submitted to University of Hertfordshire</b> Student Paper	<b>1%</b>

**polynoe.lib.uniwa.gr**