Bengali News Classification Using Different Machine Learning and Deep Learning Algorithm

Submitted by
Md Mahamodul Islam
172-35-2152
Department of Software Engineering
Daffodil International University

Supervised by
Md Sanzidul Islam
Lecturer
Department of Software Engineering
Daffodil International University

This Project report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering.

APRIL 2021

APPROVAL

This Research titled **“Bengali News Classification Using Different Machine Learning and Deep Learning Algorithm”**, submitted by **Md Mahamodul Islam** to the Department of Software Engineering, Faculty of Science and Information Technology, **Daffodil International University**, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

Chairman

Md Anwar Hossen
Assistant Professor
Department of Software Engineering
Daffodil International University

Internal Examiner 1

Asif Khan Shakir
Senior Lecturer
Department of Software Engineering
Daffodil International University

Internal Examiner 2

Professor Dr M Shamim Kaiser
Institute of Information Technology
Jahangirnagar University

External Examiner

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Md Sanzidul Islam, Lecturer, Department of Software Engineering**, Faculty of Science and Information Technology, Daffodil International University. I also declare that neither this research nor any part of this research has been submitted elsewhere for the award of any degree.

Supervised by:

Md Sanzidul Islam
Lecturer
Department of Software Engineering
Daffodil International University

Submitted by:

Mahamod

Md Mahamodul Islam
ID: 172-35-2152
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

ACKNOWLEDGEMENT

First and foremost, I would like to express my heartfelt gratitude to almighty God for his wonderful grace, which has enabled me to effectively complete the final year thesis.

I am extremely thankful and desire to express my deep gratitude to **Md Sanzidul Islam**, Lecturer, Department of Software Engineering, Faculty of Science and Information Technology, Daffodil International University, Dhaka. His in-depth knowledge and strong enthusiasm, combined with encouraging guidance, aided me in my machine learning and deep learning-based research, finally I have successfully completed this work on "Bengali News Classification Using Different Machine Learning and Deep Learning Algorithm". His never-ending patience, intellectual direction, continuous guidance, consistent and enthusiastic supervision, constructive criticism, valuable advice, and reviewing numerous poor manuscripts and improving them throughout each level paved the way to finish this job.

I would like to thank you from the bottom of my heart to **Dr. Imran Mahmud**, Professor and Head, Department of Software Engineering, Faculty of Science and Information Technology, DIU, for his invaluable assistance and advise in completing my project, as well as my heartfelt gratitude to other faculty member and the staff of Department of Software Engineering, Daffodil International University.

I would want to express my gratitude to all of my well-wishers, friends, family, and seniors for their support and inspiration. This research is the result of a lot of hard effort as well as a lot of inspiration and help.

Finally, I must express my gratitude for my parents' unwavering support and patience.

ABSTRACT

In this modern era, Artificial Intelligence has emerged as the next data science powerhouse. The use of Machine Learning, Deep Learning, and Computer Vision algorithms in data analytics has become a popular trend since their introduction. However, applying Support Vector Machine, Naïve Bayes, Convolution Neural Network, Long Short-Term memory in different Bengali text classification tasks and study the performance of these models is yet to be explored. Hence, in this paper, we have proposed different machine learning ad deep learning based model building in order to classify 7 types of news category of Bengali newspaper data. I have compared their all-accuracy level between all of the building model SVM, Naïve Bayes, CNN, LSTM, CNN-LSTM and between all of them LSTM and CNN have achieved fairly high accuracy with the containing of a 0.5(504266) million datasets.

TABLE OF CONTENTS

CONTENTS **PAGE**

Board of Examiners…………………………………………………………....... i

Declaration……………………………………………………………………… ii

Acknowledgment………..……………………………………………………... iii

Abstract……………………………………………………………………………. iv

CHAPTER **PAGE**

CHAPTER 1: INTRODUCTION 1-7

1.1 Introduction…………………………………………………………………… 1-2

1.2 Objective ………………………………………………………….…………... 2-3

1.3 Motivation…………………………….…………………………………...... 4

1.4 Rational of the Study………………………….……………………………... 5

1.5 Research Questions………………………………………………………….... 5

1.6 Expected Outcome…..……………………………………………………… 6

1.7 Layout of the Report………………………………………………………. 7

CHAPTER 2: BACKGROUND 8-10

2.1 Introduction…………………………………………................................ 8

2.2 Related Works ……………………………….………………………… 8-10

2.3 Challenges…………………………………………………………………. 10

CHAPTER 3: RESEARCH METHODOLOGY 11-25

3.1 Introduction…………………………………………………………… 11

3.2 Research Subject and Instrumentation……………………………… 11

3.3 Workflow ……………………………………………………………... 11-13

3.4 Data Collection Procedure…………………………………………… 14

3.5 Data Insight Details…………………………………………………… 15

3.5.1 Data Pre-Processing…………………………………………………. 16

3.5.2 Data Preparation………………………………………………….…. 16

3.6 Understanding of the Algorithm……………………………………… 16

3.6.1 Naïve Bayes Classifier……………………………………………… 16-17

3.6.2 Support Vector Machine……………………………………………. 17-18

3.6.3 Neural Networks……………………………………………………. 18-19

3.6.4 Convolution Neural Networks……………………………………… 19-20

3.6.5 Recurrent Neural Networks…………………………………………. 20-21

3.6.6 Long Short-Term Memory…………………………………………... 21-22

3.7 Training the Model …………………………………………………… 23-24

3.8 Implementation Requirements………………………………………… 25

CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION 26-35

4.1 Introduction …………………………………………………………. 26

4.2 Performance Evaluation ……………………………………………... 26-29

4.3 Result Discussion ……………………………………………………… 30-35

CHAPTER 5: SUMMARY AND CONCLUSION 36

5.1 Introduction …………………………………………………………. 36

5.2 Conclusion...…………………………………………………………… 36

5.3 Future Works…………………………………………………………… 36

REFERENCES 37

APPENDIX: A 38

APPENDIX: B 38

ACCOUNTS CLEARANCE 39

LIST OF FIGURES

FIGURES **PAGE**

Fig 3.3: Workflow of my approach. 13

Fig 3.4: A primary sight of dataset 14

Fig 3.5: Total percentage data as category wise 15

Fig 3.6.2: Support vector Machine 17

Fig 3.6.3: Neural Networks 18

Fig 3.6.4: Structure of CNN 19

Fig 3.6.5 Structure of RNN 20

Fig 3.6.6 Structure of LSTM 21

Fig 4.2.1: Training &Testing accuracy of LSTM 26

Fig 4.2.2: Training &Testing accuracy of CNN 27

Fig 4.2.3: Training &Testing accuracy of LSTM-CNN 27

Fig 4.2.4: Training &Testing loss of LSTM 28

Fig 4.2.5: Training &Testing loss of CNN 28

Fig 4.2.6: Training &Testing loss of LSTM-CNN 29

Fig 4.3.1: Normalized confusion matrix of Naïve bayes 30

Fig 4.3.2: Without Normalized confusion matrix of Naïve bayes 31

Fig 4.3.3: Normalized Confusion Matrix of SVM 31

Fig 4.3.4: Without Normalized confusion matrix of SVM 32

Fig 4.3.5: Normalized Confusion Matrix of LSTM 32

Fig 4.3.6: Without Normalized Confusion Matrix of LSTM 33

Fig 4.3.7: Normalized Confusion Matrix of CNN 33

Fig 4.3.8: Without Normalized Confusion Matrix of CNN 34

Fig 4.4.9: Normalized Confusion Matrix of LSTM-CNN 34

Fig 4.4.10: Without Normalized Confusion Matrix of LSTM-CNN 35

LIST OF TABLES

TABLES	PAGE
Table 3.5: Bengali News Dataset	15
Table 4.3.1: Confusion Matrix	30
Table 4.3.2: Accuracy of the Model	35

CHAPTER 1
INTRODUCTION

1.1 Introduction

Bengali nation is the only nation in the world who sacrificed their life just to speak in their mother tongue. Many language martyrs died in 1952, because of their mother language. My nationality is Bengali by birth. Bengali is my mother language as I am a Bangladeshi resident. Bengali is the official language in Bangladesh. In some of Indian regions this language is well known. Bengali is the official language of West Bengal and Tripura and its position is second from the 22cadastre languages in Indian regions. In this day and age, 228million nearby speakers everyday communicate in the Bengali language. Additionally, 37million unfamiliar people talk every day in this language. Bengali is the 5th most ordinarily communicated language on the planet from the statistics report of speakers. It has the 7th position when we talk about the most used language. News is the substance of current conditions and activities in every part of the nation or the world, published in newspapers and distributed in the electronic media or in any other media. In the present computerized world, the number of people who read newspapers is less than internet readers. E-news is the electronic version of news which is published on the internet and people can get the news by internet browsing. Because of a good developing network structure and huge users of the internet, the numbers of E-news readers are rapidly increasing. As E-news readers are increasing all over the world there is a similar scenario in Bangladesh. E-news readers are loved to browse those sites which give them the most recent updates with regular news. From the Google trend data, it is clear that most of the online readers are leaving the physical paper of the dailies like Kaler Kantha, Naya Diganta, Manab Zamin, Janakantha, Prothom Alo and so on moving forward to their online news portals. Google data also indicate that most web news readers are shifted to those online news portals where they find the daily news, breaking news and time to time updated news like bdnews24.com, banglanews24.com [1]. The abundance of unformatted data in textual documents which entices the people of natural language processing (NLP) to bring out some meaningful information from that data and analyze them find something useful for improving the online unstructured textual data in the future. The most and improving the unstructured content quality text categorize is the important where there are several applications available like searching, browsing, and organizing textual records are mostly depend on the text classification. In last few years. Analyzing and extricate significant attribute for precisely categorize the textual documents some machine learning and statistical methods have been applied. For example, TF-IDF [15], Bag of words, Word embedding [16] using this feature have been extracted from the text documents for the purpose of train some supervised learning algorithm by applying this text classify into different category, for instance, Naïve bayes, Logistic Regression, KNN and nowadays some researchers are using deep learning algorithm. Deep learning algorithm has been developed foundation of machine learning algorithm which can utilize various layers to steadily extract more significant level of feature from the intense input [14].

In other language text classification, there are available a large dataset for applying model and getting better output whereas In Bangla, most of the researchers working on a small amount of dataset which cannot always extract the all-valuable information and would not able to be use in large portal. In this paper, we analyze the all-previous problems for classifying textual data and developed a dataset which consist of 0.5 million data with seven different categories and comparing between some models like Naïve bayes, SVM, CNN (Convolutions Neural Networks), and LSTM (Long Short-Term Memory).

1.2 Objective

We are in the midst of a technological revolution right now. We use technology everywhere in today's digital world. In the field of Natural Language Processing (NLP), we need to move forward our Bengali Language as long as other languages who are doing a great job in NLP field. This NLP is needed every day for making easier in our life as like news online paper reader should always find easily as their preferable news according to news content category. My main objective is to classify Bengali news text into different type of categories. There are seven different types of news available in dataset all-banglaesh, politics, national, international, sports, entertainment, economics-business.

I want to classify the Bengali news by using machine learning and deep neural network. So, we can describe these goals in a list like this:

- My goal is to analyze how to classify Bengali news into different category and find out best feature from the data along with their behavior.
- To develop a system that will be able to categorize the Bengali news as their desirable category.
- To visualize some analytical analysis of news textual data for better understanding as for finding the better outcome from the data.
- As I have collected containing with different category Bengali news almost 1 million data which will be available as public dataset for the purpose of future research and betterment of Bengali News.

1.3 Motivation

From the beginning of my university life, I was so keen on the different applications which were developed by Artificial Intelligence till now I am curious about this area. Earlier I decided to work on different recommender system but this sector is well developed so I realized doing something about my mother language Bengali. I read some of research articles related with different language news classification then decided I can do some research on Bengali news classification which are very informative quantitative for our research community for develop our Bengali language for betterment user experience and recognize worldwide. I named it with the title of "**Bengali News Classification Using Different Machine Learning and Deep Learning Algorithm**" Besides this, observed that today's world is so much focusing on developing new technologies and making user experience better day by day. People anticipate that the system will propose the good alternatives for them. To design a system capable of making news recommendations as user interested wise, it must be capable of making decisions by itself. These piqued our curiosity in doing research-based work. Our work is entirely focused on deep learning and machine learning approach.

1.4 Rational of the study

There is no denying that hundreds of papers have been published in the field of Natural Language Processing and Text Classification. However, just a few studies have been conducted on Bengali news categorization. As a result, our work is a novel technique that incorporates a variety of machine learning and deep learning methods as well as data pre-processing. To construct a more efficient classifier application in the field of natural language processing, I am putting forth my best effort to create my unique models for this research utilizing various algorithms.

Natural language processing is a cutting-edge method it can analyze textual data or speech data find out the actual meaning of the context. The domain of natural language processing (NLP) entails programming computers to accomplish cognitive control using the natural languages that people use. Speech and textual content could be used as outputs and inputs in an NLP system. Text Classification refers a large sequence of information context and find out the keyword from whole context and help to identify the other sequence of information. Sometimes it can extract some misleading context which can be withdrawal by repeating the large training data.

1.5 Research Questions

Completing this assignment was quite difficult for me. The researchers would like to suggest the following queries to describe these sentiments and consequences in order to have a practical, effective, and precise response to the situation.

- Can I collect the huge amount of dataset of Bengali textual data?
- Is it possible to pre-process data after collecting making ready for models?
- Is it possible to improve the text classification system throughout this work?
- How online news portal and research community will be benefited from this work and dataset?

1.6 Expected Outcome

Several points are mentioned in this section, and they represent our minimal intended output. The goal of this research project is to develop an algorithm or a complete efficient methodology for categorizing Bengali news utilizing the training dataset's constructed model.

- Bengali news can be classified into different categories
- An online news portal can be benefited by using this framework.
- Online news reader also benefitted by using this framework if it is adopted by news portal system.
- From publishable public dataset research community can do more creative works with the Bengali textual data.

1.7 Layout of the Report

Chapter one has provided a background of the project with purpose, motivation, research questions, and projected conclusion. This part outlines the entire arrangement of this report.

The chapter two presented whatever has been accomplished in this arena. The second chapter's last portion demonstrates the depth that has resulted from their field's limitations. Finally, the research's major difficulties or hurdles are discussed.

The third chapter discusses the theoretical aspects of this study project. This chapter explains on the quantitative methods used in this study in order to address the theoretical portion of the research. In addition, this chapter demonstrates the Deep Learning and Machine Learning-based model construction classifier's procedural methodologies. Confusion matrix analysis is provided in the end portion of this chapter to validate the model as well as to demonstrate the accuracy label of the classifier.

The experimental findings, performance assessment, including consequence analysis are presented in Chapter 4. This chapter contains several experimentation graphical confusion matrixes to aid in the completion of project.

The fifth chapter presented the study's summarization, future development, and conclude. This section is useful for enhancing the entire project report in accordance with the recommendations. The chapter concludes by demonstrating the limitations of our work, which may be of interest to all those who wish to work in this topic in the future.

CHAPTER 2
Background Study

2.1 Introduction

In this part, I will explain about similar works, an overview of the research, and some of the research's obstacles. I will cover other study papers and their works, as well as their methodology and correctness, under the linked works area. I will offer an overview of our connected efforts in the research overview section. I will go into how we improved the accuracy level in the difficulties section.

2.2 Related Works

From the flourished of Deep Learning, individuals everywhere on the world performing different applications by utilizing Deep Learning algorithm put together with respect to their application and research papers. Numerous researchers from Bangladesh and everywhere on the world dealt with different issues identified with Bengali text classification. Before my examination works numerous researchers experimented on different algorithm dependent on Deep Learning and Machine Learning.

In 2014 Chy and et. all used RSS crawler for crawling news sites then process few techniques for cleaning text in particular, punctuation removal, digit removal, tokenization, and use Bengali dictionary for steaming. Finally, each document converted into vector and applied supervised learning algorithm Support Vector Machine for categorize the documents and achieved prosperous results, showed their results in the precision-recall graph [3]. Researchers' main focusing point was doc2vec, 'it is a neural network-based method for encoding a document's depiction in a low-dimensional vector' performs better, contextually correct than LSA and LDA modeling technique where doc2vec has an accuracy of 91.0%, which is higher than other approaches. The precision is 85% and 84% of LDA and LSA, accordingly [4].

Researchers utilized semantics procedures and Information Extraction (IE) in their examination model for Bangla textual data source [5] in another exploration paper from

2015. Such researchers focused on few Bangla textual data elements, for example, individuals and areas utilizing Traditional Natural Language Processing (NLP) procedures with semantics [5].

In 2018 a paper utilized Bi-LSTM-CNN model which exploit the circle structure for acquiring the context data also formulate the left and right settings of each word through the Convolutional Neural Network (CNN) to detail the printed articulation of the word, which is even more absolutely conveyed the semantics of the substance and they obtained an extraordinary score in Bi-LSTM-CNN contrasted with others deep learning strategies [6]. Deep learning achieved astounding achievements in field of Image Processing and Speech Recognition [7], however in the field of natural language processing it is continuously creating an impact. Deep learning technology step by step supplanted traditional machine learning techniques and fulfilled a standard innovation in textual data grouping or classifications [8]. The semantic representation of texts can be effectively composed using a neural network-based approach. When compared to conventional approaches, neural networks can collect more qualitative knowledge about features and may be less influenced by data sparsity and they achieve 96.49% accuracy in RCNN model compared to others model [9]. In 2010, Ranjan and et. all implement a system using Neural network and LSTM (Long Short-Term Memory) proceed toward and they extracting feature vector for every class using feature weighting and feature selection algorithm preprocessing a public dataset with 20 newsgroups finally they apply their LSTM model. They achieved loss occurred was 0.200 in the 25 epochs [10]. However, they could use some other algorithms for better performance as it is public dataset. Three different categories of news available for their designed model where text preprocessing is used for organize the data, HMM (hidden Markov Model) for feature extraction, SVM for text classification and acquired 92.66% accuracy using HMM-SVM model although dataset was not good enough [11]. In 2020 a group of researchers have collected a dataset consist of 300 data with different categories the applied data preprocessing like English word, Bangla digit remove, and applied some machine learning and deep learning algorithm where CNN score the best compared with Naïve Bayes, Random Forest, Logistic Regression, Linear SVM, Bi-LSTM although they did not remove stop words from the text [12]. In 2018, authors compared performance of four machine learning algorithm SVM, Stochastic gradient descent, Logistic Regression, Multinomial Naive Bayes with four different datasets of 10027, 42370,60000, 84906 size.

They applied four algorithms on their all dataset, observing that accuracy of prediction model was decrementing with the increment of training data amount and owing to the huge number of features in large datasets where classifiers model overfitting with the training data [13].

2.3 Challenges

The most difficult part of this project was gathering as well as shaping the massive dataset into a clean and well-formatted dataset. To cleaning the dataset had to follow several steps like stop words removal, English text remove and punctuation remove. Finally for machine learning algorithm had to make the dataset into tokenization as well as for deep learning had to make into vectorization. After training with machine learning and deep learning models it took a lot of time for training all of the models. So, for achieving the final output of all models had to wait a long period of time as the dataset was much larger. Although there were some available datasets which were not much larger for that reason getting a good accuracy level for Bengali text classification had to take the challenges collecting this huge dataset. For working this work, I had to start from the very scratch like data collection, data cleaning, data pre-processing, building all of the five models and had to start from my own motivation.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

In this section I will go over the steps of our unique methodology of categorizing Bengali news. There are some key points like data collection, data cleaning, data pre- processing along with related equation, diagram, figure, table, and explanation are also included in the selected framework. Self-developed five models as Naïve bayes, Support Vector Machine, LSTM, CNN, LSTM-CNN based model building used and collected of my dataset is been used in this work. The chapter concludes with an explanation of this work statistical hypotheses, as well as a solid fundamental understanding of all of the algorithms employed in this study.

3.2 Research Subject and Instrumentation

A research subject is a study field that has been evaluated and explored in order to clarify concepts. Not just for building, but also for model development, data collection, data cleaning, data implementation or processing, and model training and testing. Instrumentation, on the other hand, refers to the technology and approach we employed. We utilized the Windows platform and the Python programming language, which included several libraries including as NumPy, pandas, skit-learn, matplotlib, TensorFlow, and Keras. Regarding data science and machine learning tasks, Anaconda is a genuinely free version of the Python and R programming languages., was utilized for all of the training and testing.

3.3 Workflow

The procedure for this study includes data collecting, data cleaning, and data pre-processing, data tokenization and vectorization, model building.

Step 1 - Data Collection: Data was collected from different Bengali famous newspapers and make a large dataset by processing some processing gathering all newspapers together then labeling them with their category along with data collection I also collected the data source, text title. Collecting data from various source was so tough as the I had collected a large number of datasets.

Step 2 - Data Processing: After collecting data from numerous sources, all data was analyzed category by category. There are several data sets with distortion, flaws, and irrelevant data. We manually evaluate those data first, then go on to the next phase with the selected dataset.

Step 3 - Data Cleaning and pre-processing: After completing processing phase here according to class wise data has been shaped in a format. For training purpose, there were some English words, punctuation, special characters, white space, and digit number all had be removed from the dataset for making the dataset ready for training and testing period. This pre-processing phase was a lengthy process as this is large dataset for running all of the programming it took a lot of time.

Step 4 - Model Building: For model building there were two part one was for machine learning based model where data needed to converted into vectorization and for deep learning data converted into tokenization and finally for training and testing this dataset for acquiring a good output and showing the differentiate behavior from different , used a total five models by using machine learning and deep learning algorithm besides will show the comparison between them who are doing better in training and testing phase.

Step 5: Performance Evaluation: All of the graphical and theoretical result have been graphed and explained in this section. Following training and testing, these processes produced a number of accuracy graphs along with testing loss and accuracy, generated the confusion matrix for all models also displayed their accuracy and performance.

Step 6: Conclusion and Future Work: A concluding and a development plan for the long term will be included in this section.

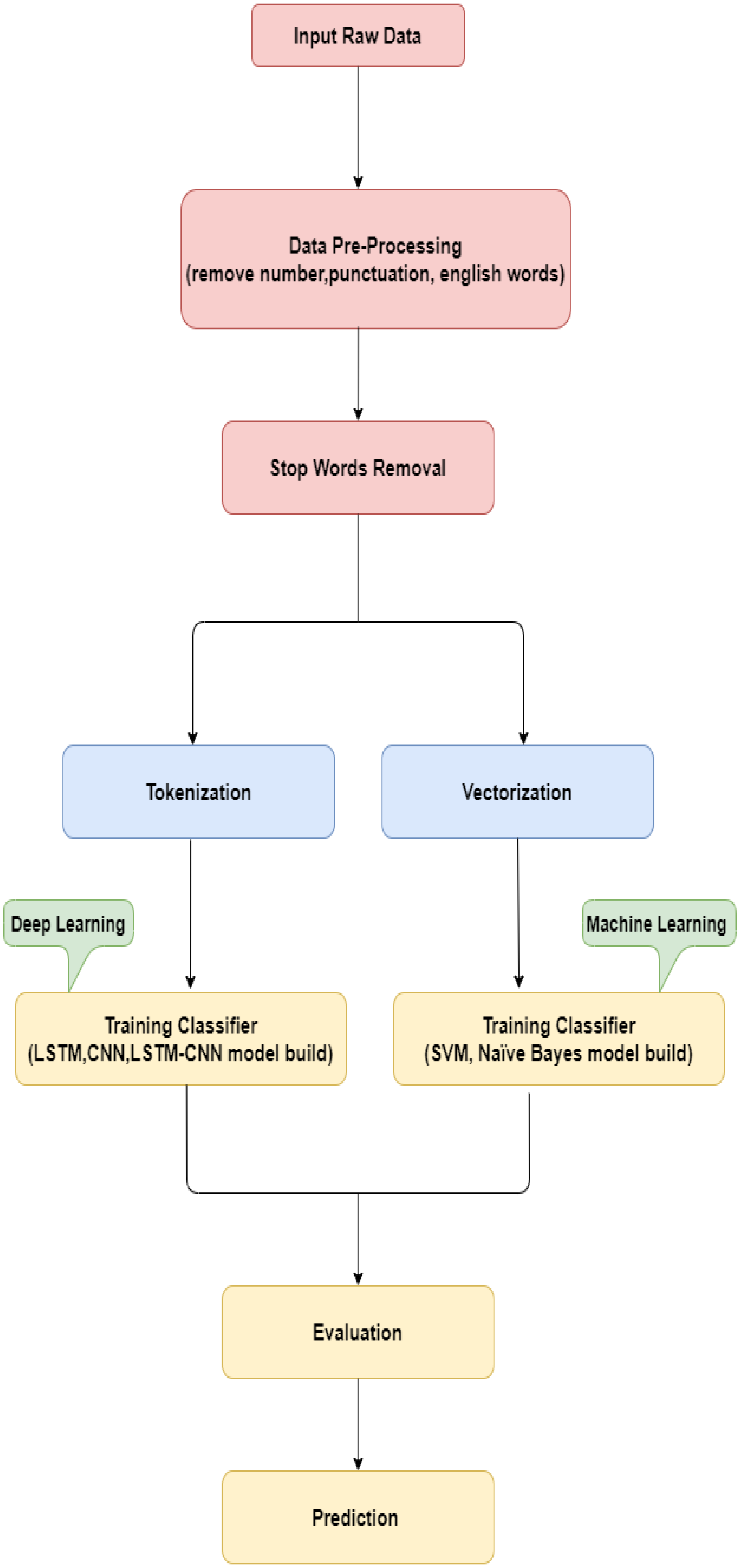
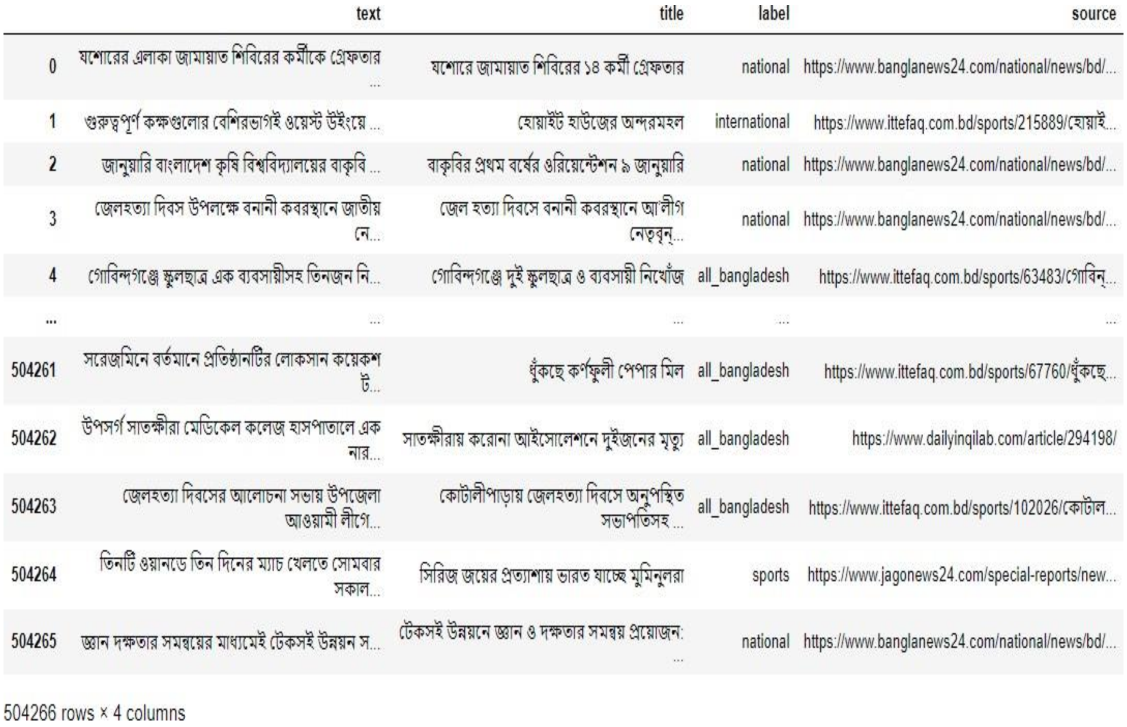


Fig 3.3: Workflow of my approach.

3.4 Data Collection Procedure

Most important part is data for any machine learning problem, there are least amount of dataset of Bengali text available which are not much large so we have collected a dataset consist of 1 million and named this dataset as Potrika. It is largest dataset available till now with Bangla text. For collecting this data scrapy framework is been used. Potrika dataset is isolated news articles well-structured and organized data was collected from different sources of famous Bangla newspapers.



	text	title	label	source
0	যশোরের এলাকা জামায়াত শিবিরের কর্মীকে গ্রেফতার ...	যশোরে জামায়াত শিবিরের ১৪ কর্মী গ্রেফতার	national	https://www.banglanews24.com/national/news/bd/...
1	গুরুত্বপূর্ণ কক্ষগুলোর বেশিরভাগই ওয়েস্ট উইংয়ে ...	হোয়াইট হাউজের অন্দরমহল	international	https://www.ittefaq.com.bd/sports/215889/হোয়াই...
2	জানুয়ারি বাংলাদেশ কৃষি বিশ্ববিদ্যালয়ের বাকৃবি ...	বাকৃবির প্রথম বর্ষের ওরিয়েন্টেশন ৯ জানুয়ারি	national	https://www.banglanews24.com/national/news/bd/...
3	জেলহত্যা দিবস উপলক্ষে বনানী কবরস্থানে জাতীয় নে...	জেল হত্যা দিবসে বনানী কবরস্থানে আ'লীগ নেতৃবৃন্...	national	https://www.banglanews24.com/national/news/bd/...
4	গোবিন্দগঞ্জে স্কুলছাত্র এক ব্যবসায়ীসহ তিনজন নি...	গোবিন্দগঞ্জে দুই স্কুলছাত্র ও ব্যবসায়ী নিখোঁজ	all_bangladesh	https://www.ittefaq.com.bd/sports/63483/গোবিন্...
...	...	...	...	...
504261	সরেজমিনে বর্তমানে প্রতিষ্ঠানটির লোকসান কয়েকশ ট...	ধুঁকছে কর্ণফুলী পেপার মিল	all_bangladesh	https://www.ittefaq.com.bd/sports/67760/ধুঁকছে...
504262	উপসর্গ সাতক্ষীরা মেডিকেল কলেজ হাসপাতালে এক নার...	সাতক্ষীরায় করোনা আইসোলেশনে দুইজনের মৃত্যু	all_bangladesh	https://www.dailyinqilab.com/article/294198/
504263	জেলহত্যা দিবসের আলোচনা সভায় উপজেলা আওয়ামী লীগে...	কোটালীপাড়ায় জেলহত্যা দিবসে অনুপস্থিত সভাপতিসহ ...	all_bangladesh	https://www.ittefaq.com.bd/sports/102026/কোটাল...
504264	তিনটি ওয়ানডে তিন দিনের ম্যাচ খেলতে সোমবার সকাল...	সিরিজ জয়ের প্রত্যাশায় ভারত যাচ্ছে মুমিনুলরা	sports	https://www.jagonews24.com/special-reports/new...
504265	জ্ঞান দক্ষতার সমন্বয়ের মাধ্যমেই টেকসই উন্নয়ন স...	টেকসই উন্নয়নে জ্ঞান ও দক্ষতার সমন্বয় প্রয়োজন: ...	national	https://www.banglanews24.com/national/news/bd/...

504266 rows × 4 columns

Fig 3.4: A primary sight of dataset

3.5 Data Insight Details

This dataset has total 4 attributes title, body, label, source and for some data variation and training purpose I have reduced the dataset and final version of Potrika dataset consist of 0.5 (504266) million data with 7 different categories sports, national, international, all-bangladesh, politics, entertainment and economic-business will be used in our experiment with different algorithms.

Category	Total Data
Sports	86749
International	84742
National	84431
All Bangladesh	81176
Politics	66142
Entertainment	53975
Economic business	47051
Total	**504266**

Table 3.5: Bengali news Dataset

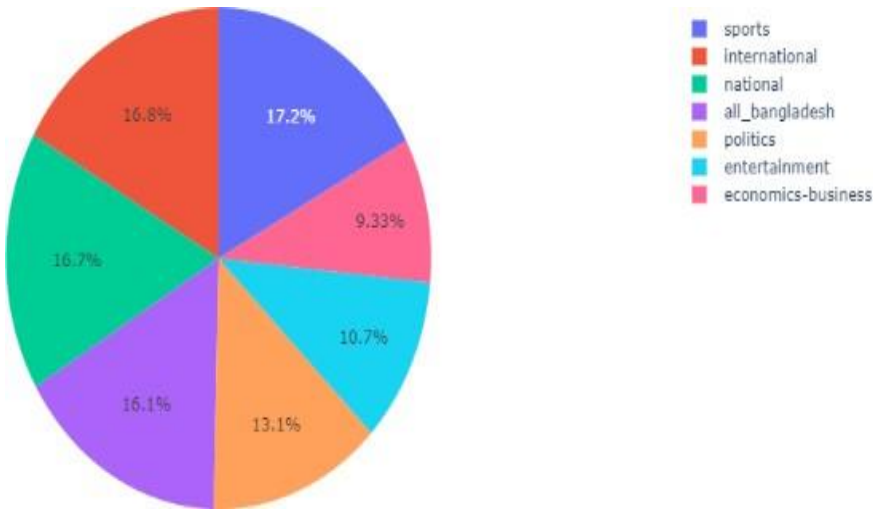


Fig-3.5: Total percentage data as category wise

3.5.1 Data Pre-Processing

For training and testing our data we need preprocessing the data first of all I remove the all-English word from the text and remove all digits English and Bangla also any kind of symbols also has been removed. We remove the all punctuations as it does not make any sense for our perdition. Secondly, we dispel the all stopwords from the textual document which will make more accurate in our training and testing process as stop words are important whose always available in every category text and it making the confuse for the prediction output.

3.5.2 Data Preparation

As till now, this dataset is ready for training into model but before training and testing need not convert the data into numerical value as computer only can identified the digit. So, here for machine learning model building converted the whole dataset into vectorization and for deep learning model transformed the dataset into tokenization.

3.6 Understanding of the Algorithm

Hence, there are total five model building so I use different five algorithm SVM, Naïve Bayes, SLTM, CNN, LSTM-CNN. Here I will explain the all of algorithm working method and related some other method will also be described.

3.6.1 Naïve Bayes Classifier

Naïve Bayes is a method for classification problems dependent on Bayes hypothesis with a presumption of freedom among predictors. A Naive Bayes classifier, in essential words, suggests that the presence of one element in a class is unimportant to the presence of some other component. This classifier used for large volume of dataset and work better. The Bayes theorem is a probability hypothesis that applies to contingent probabilities. The likelihood that something will occur if something different has effectively happened is known as restrictive likelihood. Utilizing earlier data, restrictive probability will compute the probability of an event.

This is contingent probability equation is: $P(X|Y) = P(Y|X). P(X) / P(Y)$

Here, P(X): The likelihood of theory H being valid. This is known as the earlier probability. P(Y): The likelihood of the proof. P(X|Y): The likelihood of the proof given that theory is valid. P(Y|X): The likelihood of the speculation given that the proof is valid.

3.6.2 Support Vector Machine

A Support Vector Machine (SVM) is a mechanism for classifying objects based that uses a confined hyperplane to represent it. All things considered; computation produces an ideal hyperplane that organizes innovative models supplied named getting ready data (supervised learning). The hyperplane is an important line that dividing the plane into two different halves along with every class with their own side that all is happen in two-dimensional space as like Fig-1.

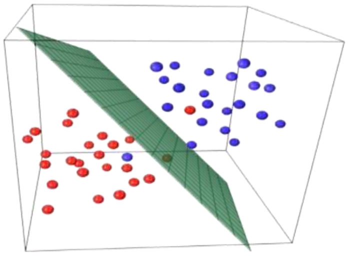


Fig 3.6.2: Support vector Machine

In the SVM algorithm, we are hoping to augment the edge between the information focuses and the hyperplane. The loss function that expands the edge is pivot misfortune.

$$C (X, Y, f(X)) = \{0 \text{ if } Y * f(X) >= 1, \text{ else } 1-Y* f(X)$$

The expense is 0 if the anticipated worth and the real worth are of a similar sign. On the off chance that they are not actually, it finds out the misfortunate value and additionally adding a regularization boundary into expense work. The goal of this regularization boundary is to adjust edge augmentation as well as misfortune. In wake of adding regularization boundary, the expense capacities look as beneath

$$\boldsymbol{m}\mathbb{i}n_{w}\lambda\|w\|^2 + \sum_{i=i}^{n}(1 - y_i\langle x_i, \omega\rangle)_+$$

Loss function for SVM

Taking fractional subordinates regarding the weights to discover the angles which updating weights. When there is a misrepresent incorporate the misfortune close by the regularization limit to carry out gradient bring up to date.

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

Gradient’s Update

3.6.3 Neural Networks

Neural networks are a combination of algorithm which specify to extract the relationship between a set of data by processing neurons which actually mimic the operation and that is actually do of our human mind. It identifies the numerical pattern which adopted in vectors where this numerical pattern can able to identify or extract the main feature information from textual data or image, also sequential data, and sound wave. All of this feature info accommodated in vectors which process as like our human brain neuron process information finally produce an output.

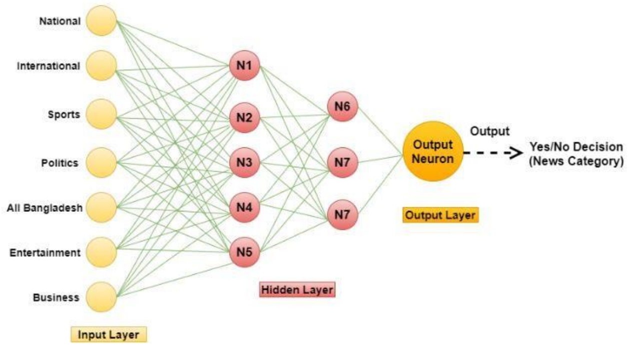


Fig 3.6.3: Neural Networks

Artificial Neural Network normally process a large amount of data simultaneously by following process of neurons. In this part this organize all of this neuron in different layers. A node layer comprises the input layer where accommodate vector data is fed into input layer from all of this node in input layer are connected to hidden layer along with containing some weight value with all nodes and threshold link with every node. In artificial neural network there can be a lot of hidden layers where for activating the node the weight value of node be higher than threshold and forwarding data into the network's upcoming layer. Output layer provide a result from the data of input layer.

3.6.4 Convolution Neural Networks

A Convolutional Neural Network (CNN) is a type of artificial neural architecture that analyzes data using perceptron in machine learning algorithms for supervised learning applications. CNNs use 3-dimensional layers where just a portion of the neurons are associated with the past layer. CNNs are shaped by stacking various layers that change the information; kernel (convolutional layer), pooling layer, rectified linear unit (ReLU) layer, and fully connected layer. Finally, from the fully connected layer it is apply in the neural network neurons. Local connections layering and spatial invariance from where the architecture of deep convolutional neural networks was inspired. In CNN architecture, a pooling layer is regularly embedded between progressive convolution layers(kernels).

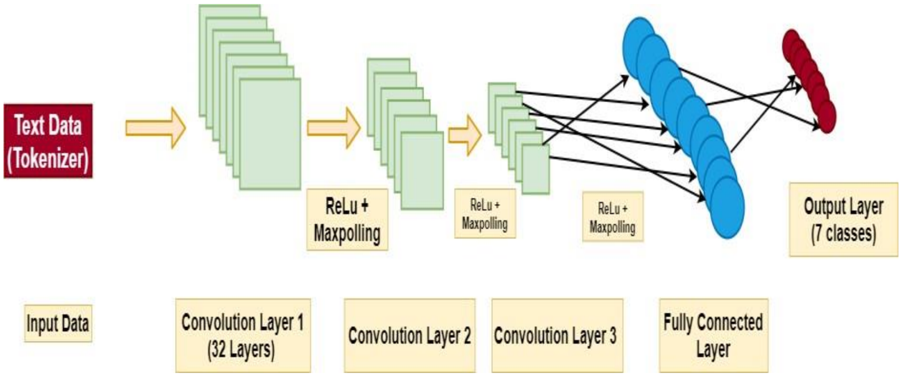


Fig 3.6.4: Structure of CNN

while controlling overfitting the pooling or subsampling layer lessens the quantity of boundaries and computational requirements. CNNs are extracted the data from the input and reshaping the input data prepare the data for applying any method like Artificial Neural Network.

3.6.5 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a type of neural network that can handle both sequential and time series data it contains loop's structure. It can be called collection of networks that are linked with each other and their frequently feature a chain like architecture makes all of the network for co-operative for getting a meaningful result. All of sequential data and time series data converted into vector sequences in input layer. The output from a specific layer always relies on the previous input. It utilizes the previous input data in the current layer for providing meaningful output for the next layer. It has internal memory which help them process the sequence of information into every phase also memorize the previous input data for providing better out in the next input data. All of inputs of RNN is dependent each other for stepping into the next movement for final output.

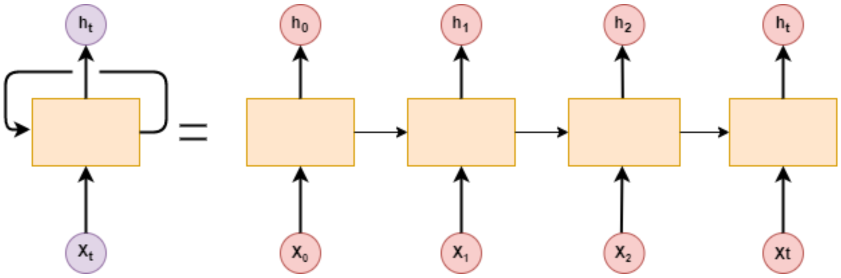


Fig 3.6.5 Structure of RNN

Firstly, it proceeds with X (0) input where it providing an outputs h (0). Next loop structure it follows the from previous output h (0) and x (1) are input data. It continuing the process till the h(n-1) and X(n) phase while processing continuing loop structure in training it keeps the record in the internal memory

The is the initial state of equation:

$$h_t = f(h_{t-1},x_t)$$

It generally operates tanh as use of activation function:

$$h_t = \tanh (W_{hh}h_{t-1}+ W_{xh}x_t)$$

Here, weight is denoted by W, where W_{hh} represents the weight of previous hidden input, W_{xh} – weight of the present input phase, and h which is represents the single hidden vector. For specifying the final output state denoted by y_t.

$$y_t = W_{hy}h_t$$

3.6.6 Long Short-Term Memory

LSTM is a specific variant of RNN that was developed to ensure the constraints of RNN that’s solve the RNN problems it is most expertise for vanishing the gradient problem. As the mistake propagates through the network, it must pass into unraveling temporal loop - the concealed states are connected in time by weights consumption. whereas this weight is put multiple times on top of itself, the gradient soon decreases. Finally, the weights of every states nodes on the outlying left are renovating gradually than those on the outlying right

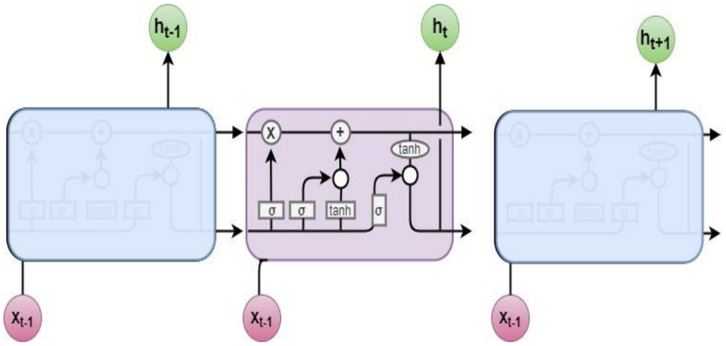


Fig 3.6.6 Structure of LSTM

Because the weights of the outlying left layers specify the inputs states to the outlying right layers, this has an effect on the different outcomes of relating with this issue. Furthermore, the entire network's training experience hardship, which is known as the vanishing gradient problem.

LSTM consist of three gates which are input gate, output gate, forget gate. It processing the information in the all states using back-propagation. LSTM solves the sequential data problem by providing this gate.

Input gate

For updating cell state input gate are available where previous and present state input go through sigmoid function and analyze the important value for adopted into the input state memory and analyzing decide values between 0,1 where tanh function augment the weight value 1 to -1 from this tanh weight scaling value decide to keep the output.

$$i_t = \sigma (W_i * [h_{t-1}, x_t] + b_i)$$

$$C_t = \tanh (W_c * [h_{t-1}, xt] + b_c)$$

Forget gate

This gate decide which information should keep or throw out and sigmoid function makes the all decision for passing all information from previous (ht-1) and current (xt) state. Output value values comes out between 0 and 1 for every cell in state (Ct-1) and value which is near to 0 make the forget and near to 1 make the keep information.

$$f_t = \sigma (W_f * [h_{t-1}, x_t] + b_f)$$

Output gate

The output gate which takes the decision in upcoming hidden state what would text sequence. This hidden state always apprehends sequence of information from the previous state and it also makes the predations. Here novel updated cell into tanh function and make the multiply between tanh output and sigmoid function output, finally the decide what sequence information should hidden state contain.

$$o_t = \sigma (W_o * [h_{t-1},x_t] + b_o)$$

$$h_t = o_t * \tanh(C_t)$$

3.7 Training Model

In this section, I demonstrate the Naïve Bayes classifier, SVM (Support Vector Machine), Recurrent Neural Network (RNN), LSTM (Long Short-Term Memory), including Convolutional Neural Networks (CNN), algorithm-based model accomplishment, the process we followed for trained all of this model.

I have proposed those models and completed the all tasks as required and I use the same dataset, same category for individual models for omit the bias output. Dataset has been shuffled randomly for all of the model. Form the table:1 category wise total data was the same for every models.

Naïve Bayes Classifier Model

For training and testing purpose splitting the dataset 10% data used for testing from total 0.5 (504266) million data. Creating a Count Vectorizer to gather the unique elements in sequence from the dataset. Dataset has been shuffled randomly for better understanding in the training period.

Support Vector Machine Model

For training and testing purpose splitting the dataset 0.20% data used for testing from total 0.5 (504266) million data. Creating a TfidfVectorizer to gather the unique elements in sequence from the dataset which are more useful and enable encrypt new documents by learning the vocabulary and inverse document frequency weightings. Dataset has been shuffled randomly for better understanding in the training period. We use linear kernel for data separation by using this, data is separated by single line. SVM linear kernel faster than any other kernel in this algorithm.

LSTM Model

LSTM is Deep learning-based algorithm so in this model we use 0.10% data for testing from the dataset. Our LSTM model has 4 layers, firstly tokenized the all data then in the embedding layer takes the all-tokenized word as input. Secondly spatial dropout layer where has been applied 20% dropout, it reduces the overfitting in the model.

Thirdly we make used of total 100 hidden size and in dropout layer applied 20% recurrent dropout also applied 20%. Finally, Dense where use category number total 7 and to extract the output result of likelihood we utilized softmax activation function in model implementation phase. We run the model using 128 batch size and 10 epochs for training.

CNN Model

For this Convolution neural network (CNN) model 0.10% used for the testing from dataset. Embedding layer takes the all-tokenized data as input conv1d layer applied for shorter the vector size. maxpoolind1d applied for taking the maximum values from tensor over the window where pool sized is being defined and strides change the position of the frame. Finally, without flatten the all output for classifying in the dense where softmax has been used for getting the output of probability. For running the model used 10 epochs and 128 batch size.

LSTM-CNN Model

we use 0.10% data for testing from the dataset. This LSTM-CNN model has five layers, firstly tokenized the all data then in the embedding layer takes the all-tokenized word as input. Secondly conv1d layer applied for shorter the vector size. Thirdly maxpooling1d applied for taking the maximum values from tensor over the window where pool sized is being defined and strides change the position of the frame. Fourthly LSTM has been used where Relu was the activation function if the output is negativities, it flattens that outputs to Zero. Finally flatten the all output for classify the category where class no total 7 as well as to extracting output of probability here we utilized softmax activation function into the dense of this model. We run the model using 128 batch size and 10 epochs.

3.8 Implementation Requirements

Following a thorough examination of all relevant statistical or theoretical ideas and methodologies, a list of prerequisites for such a text classification project was compiled. The following items are likely to be required:

Hardware/Software Requirements

- ✓ Operating System (Windows 7 or above)
- ✓ Hard Disk (minimum 500 GB)
- ✓ Ram (Minimum 12 GB)

Developing Tools

- ✓ Python Environment
- ✓ Spyder (Anaconda3)
- ✓ Jupyter Notebook
- ✓ Google Colab

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In this section, I described the output and behavior process of Bengali news classification of all models. The overall process of all model divided into few steps like all model's confusion matrix, training, testing accuracy, training and testing loss. Finally, their output result and their final accuracy in which portion or which categorical data are working better of all model.

4.2 Performance Evaluation

When model is experimented by using the training dataset also assessing using training data at this point it is entitled with training accuracy. Similarly, When the train segment of the modal is constructed, the testing phase begins. For determining model accuracy or assessing performance, the model evaluated a few previously unknown data from testing dataset, and the effectiveness of these previously unknown data is ascribed with test accuracy. For this model, we created a graph that demonstrates the relation between train and test accuracy

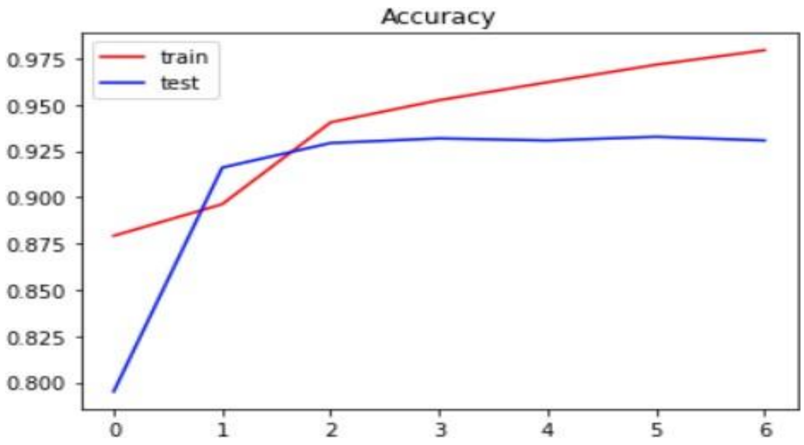


Fig 4.2.1: Training &Testing accuracy of LSTM

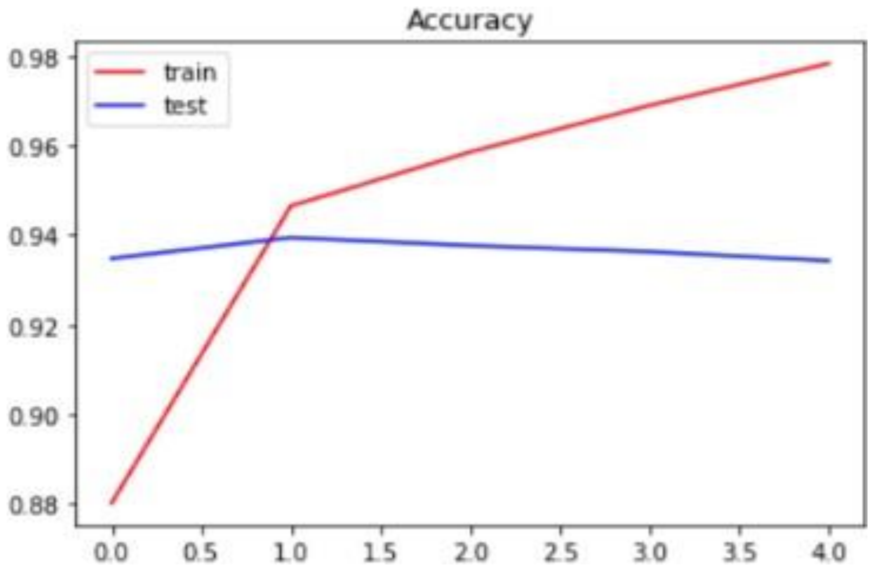


Fig 4.2.2: Training &Testing accuracy of CNN

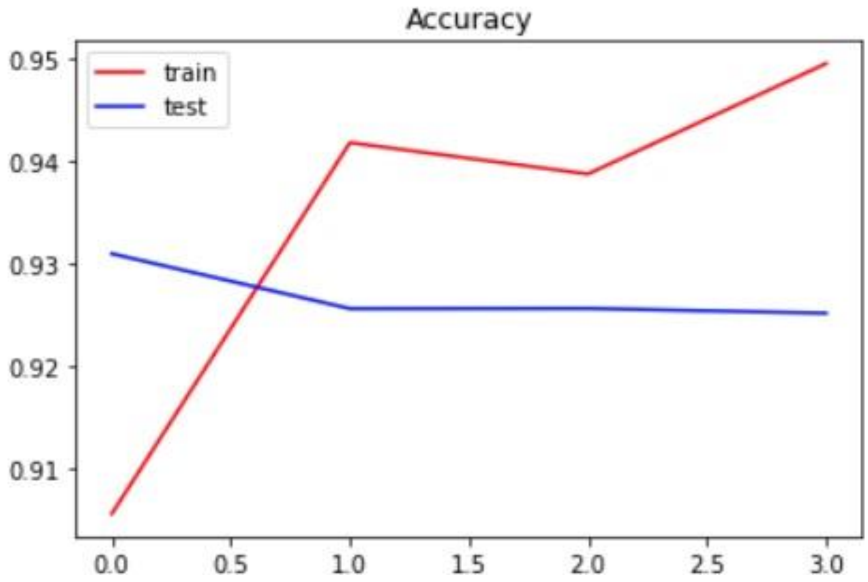


Fig 4.2.3: Training &Testing accuracy of LSTM-CNN

The blunder while anticipating on preparing dataset is called preparing misfortune. The error occurs during the testing phase, when the model is evaluated with unknown data; this error is known as test loss. The plot of train loss vs test loss for this model

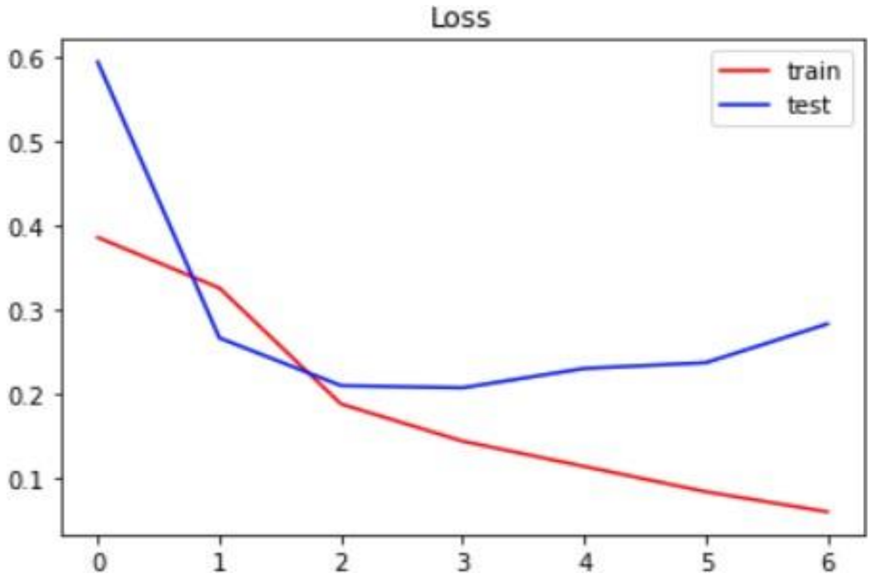


Fig 4.2.4: Training &Testing loss of LSTM

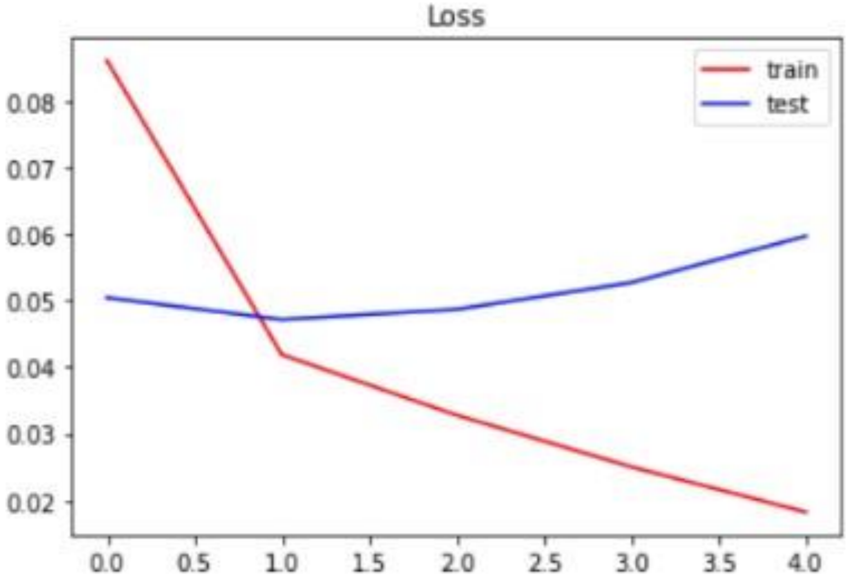


Fig 4.2.5: Training &Testing loss of CNN

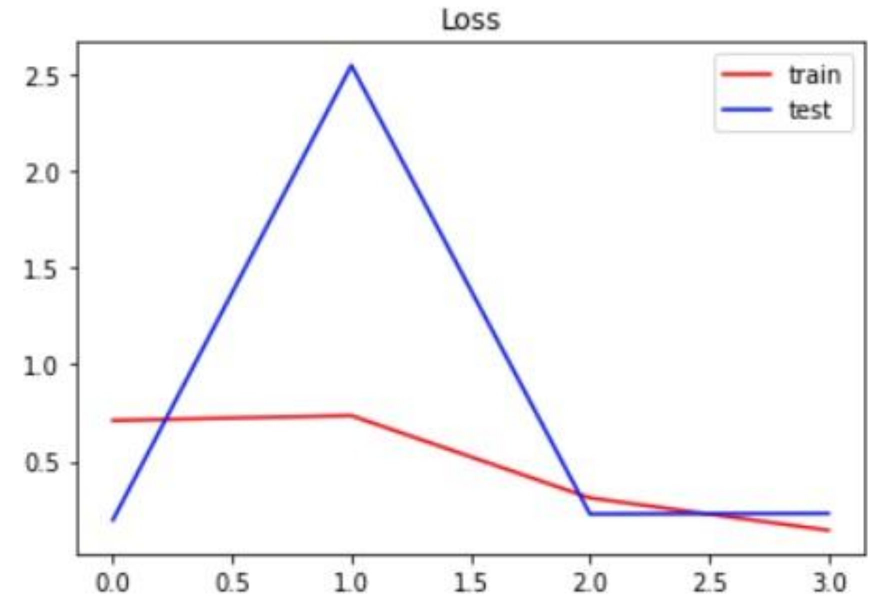


Fig 4.2.6: Training &Testing loss of LSTM-CNN

4.3 Result Discussion

We calculated confusion matrix and normalized confusion matrix for the all of our models. For the classification test case it is important how is model behaving and their predicted accuracy score.

The confusion matrix is a method for evaluating a classification algorithm's effectiveness. Whether you have an unbalanced number of samples from each category or if the dataset has more than 2 groups, classification accuracy alone would be deceptive. As a consequence, the aforementioned method was employed to present the result for better comprehension. The number of positive and negative cases measured and predicted is shown in a confusion matrix. This also allows to see how well the model is behaving.

True Positive	False Positive
False Negative	True Negative

Table 4.3.1: Confusion Matrix

Here sports (1), international (2), national (0), all_bangladesh (3), politics (4), entertainment (6), economics-business (5) for Normalized confusion matrix.

Naïve bayes Model

After completing our experiment, we achieved 0.845% accuracy in this model which is quite good for this model. From fig: 4.31 can see the normalized confusion matrix.

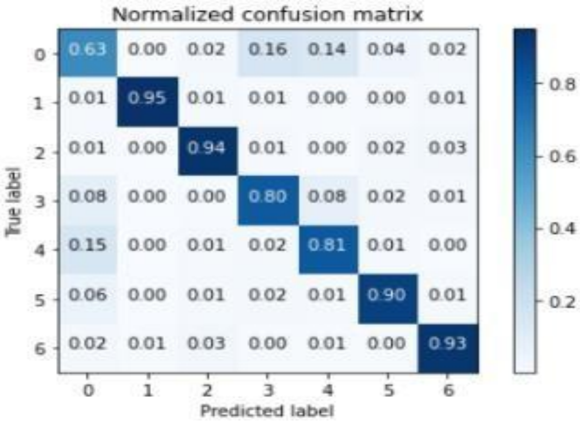


Fig 4.3.1: Normalized confusion matrix of Naïve bayes

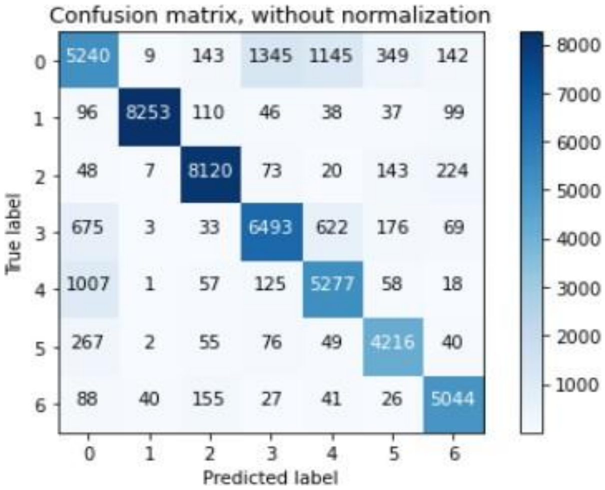


Fig 4.3.2: Without Normalized confusion matrix of Naïve bayes

From fig:4.3.1 we can see that best accuracy we are getting from sports (0.95), international (0.94) accordingly whereas national category data performed so poor (0.63).

SVM

Here we achieved 90.9% accuracy for this algorithm which is pretty good for this algorithm. From Fig: 4.3.2 normalized confusion matrix. All of the individual category performed very well.

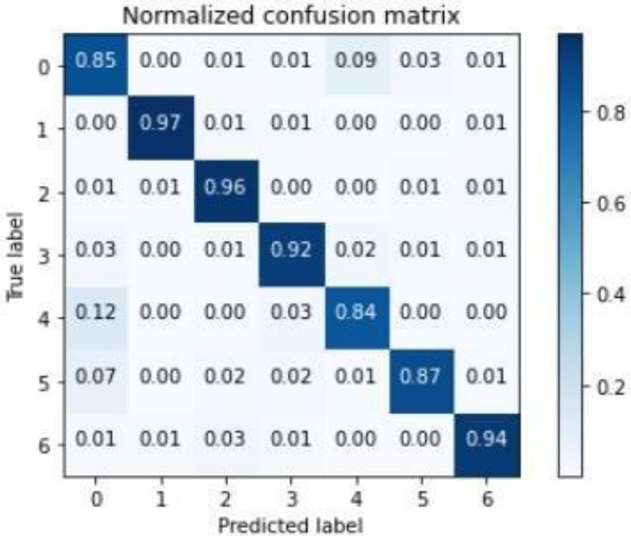


Fig 4.3.3: Normalized Confusion Matrix of SVM

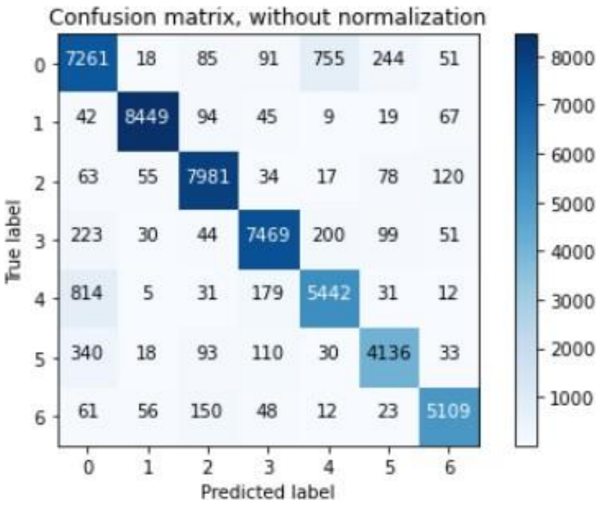


Fig 4.3.4: Without Normalized Confusion Matrix of SVM

LSTM

We run the model using 128 batch size and 10 epochs but in the 7th epochs model achieved 0.980% accuracy and stopped. Final overall accuracy of this model is 0.933% and loss is 0.280%.

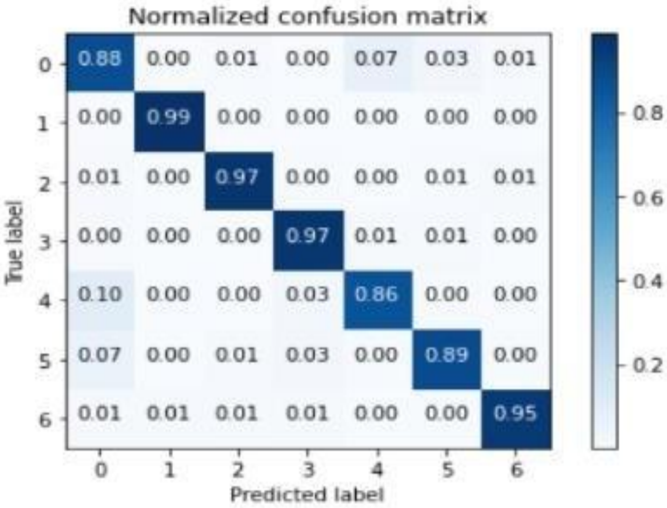
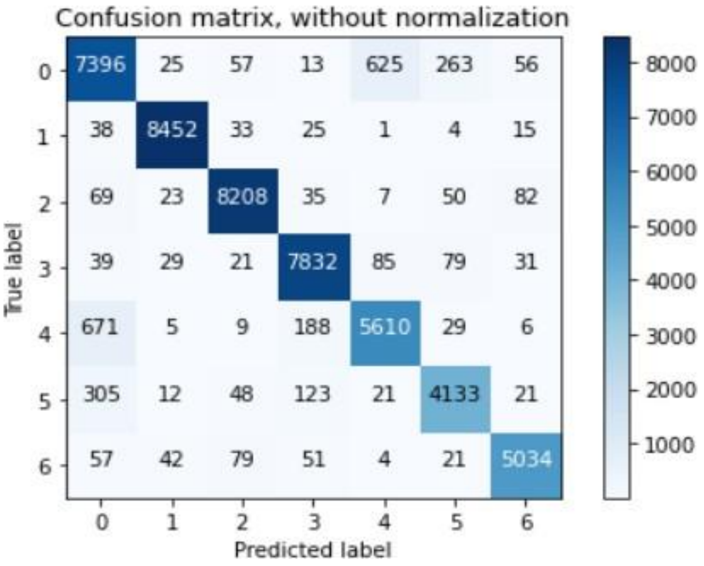


Fig 4.3.5: Normalized Confusion Matrix of LSTM

Here for this model category wise performance is pretty good where sports (1) got achieved the 0.99% accuracy whereas lowest here politics (4) which is 0.86% from fig:12.

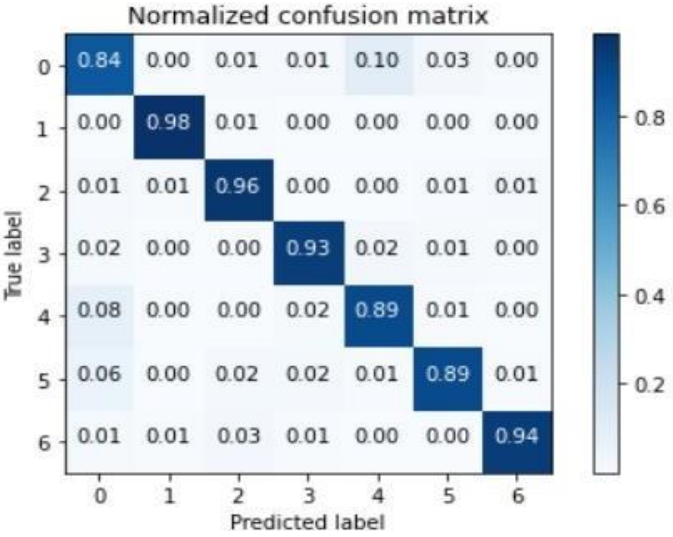


	0	1	2	3	4	5	6
0	7396	25	57	13	625	263	56
1	38	8452	33	25	1	4	15
2	69	23	8208	35	7	50	82
3	39	29	21	7832	85	79	31
4	671	5	9	188	5610	29	6
5	305	12	48	123	21	4133	21
6	57	42	79	51	4	21	5034

Fig 4.3.6: Without Normalized Confusion Matrix of LSTM

CNN

For running the model used 10 epochs and 128 batch size unfortunately 5th epochs model achieved 0.9803% and stopped. Final accuracy for this model is 0.933% and loss is 0.061%.



	0	1	2	3	4	5	6
0	0.84	0.00	0.01	0.01	0.10	0.03	0.00
1	0.00	0.98	0.01	0.00	0.00	0.00	0.00
2	0.01	0.01	0.96	0.00	0.00	0.01	0.01
3	0.02	0.00	0.00	0.93	0.02	0.01	0.00
4	0.08	0.00	0.00	0.02	0.89	0.01	0.00
5	0.06	0.00	0.02	0.02	0.01	0.89	0.01
6	0.01	0.01	0.03	0.01	0.00	0.00	0.94

Fig 4.3.7: Normalized Confusion Matrix of CNN

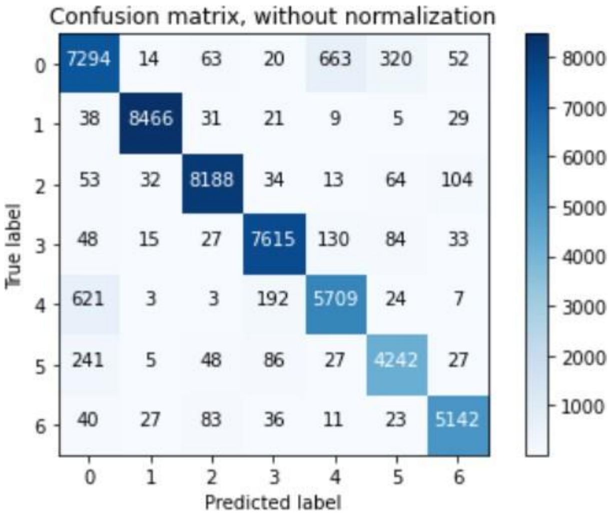


Fig 4.3.8: Without Normalized Confusion Matrix of CNN

From the experiment of CNN model here fig:4.3.4 shows the category wise accuracy where sports (1) achieved 0.98% was the highest on other hand national (0) achieved lowest accuracy 0.84%.

LSTM-CNN

For running the model used 10 epochs and 128 batch size unfortunately 5th epochs model achieved 0.9803% and stopped. Final accuracy for this model is 0.933% and loss is 0.061%.

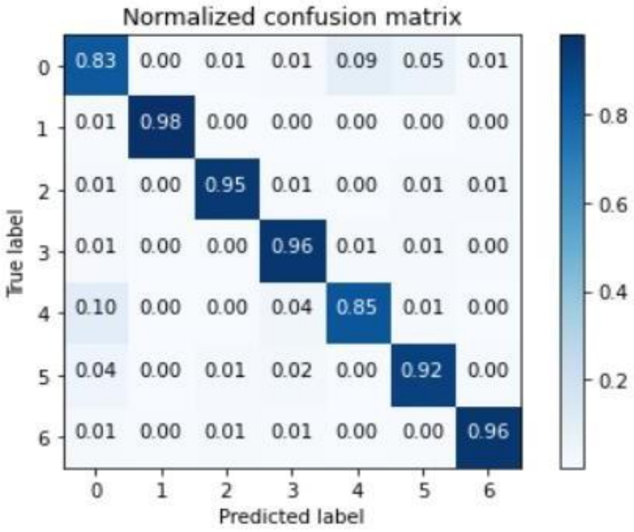


Fig 4.4.9: Normalized Confusion Matrix of LSTM-CNN

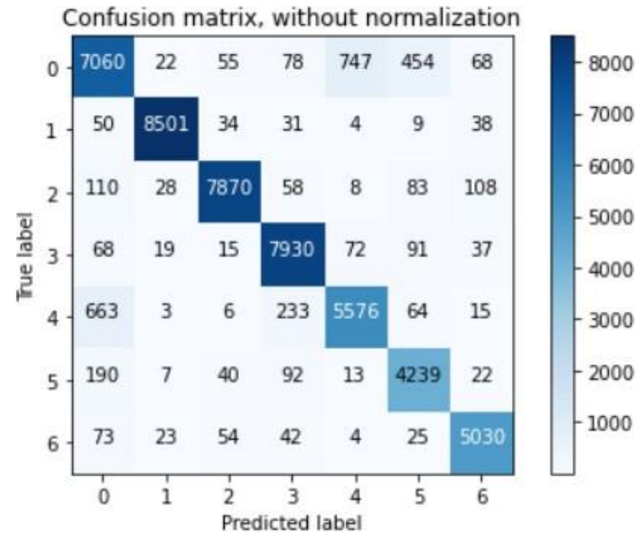


Fig 4.4.10: Without Normalized Confusion Matrix of LSTM-CNN

From fig:4.45 shows the final output of the normalized confusion matrix where sports (1) achieved the highest accuracy (0.98%).

Comparison of the Model

From table:2 shows the accuracy level between these all models where from machine learning two models Naïve Bayes and SVM where SVM score the most 90.9% whereas based on deep learning algorithm LSTM and CNN scores the most 93.3% and LSTM-CNN scores bit less from them.

Model	**Accuracy**
Naive bayes	84.5%
SVM	90.9%
LSTM	93.3%
LSTM-CNN	92.4%
CNN	93.3%

Table 4.3.2: Accuracy of the model

For this experiment where dataset and category data were same for all of the models. So finally, from all of this model working experience and validating their training, testing, loss accuracy as well as analyzing all of the model’s confusion matrix and their perdition output here individual LSTM and CNN accuracy is most which is 93.3%.

CHAPTER 5

CONCLUSION, RECOMMENDATION AND FUTURE WORKS

5.1 Introduction

There is no question that several researches worked on Bengali news categorization, which is one of the most significant sectors of natural language processing in Bengali. It has become significant in a range of applications. Many types of technologies are now available that can employ text categorization to provide the user with a positive user experience., so this approach will discover a novel enrichment in Bengali community and other stakeholders what is my main goal is enhancement the Bengali language to worldwide and using this text classification into different news portal and enriching user experience.

5.2 Conclusion

In In this paper, I used machine learning-based Nave Bayes and support vector machines where the vectorization, discreate feature and linear kernel were used as well as for deep learning-based LSTM, CNN, and LSTM-CNN, where the Embedding layer, Spatial Dropout layer, LSTM layer, and dense layer, maxpooling layer, and conv1d layer were used to build the classification model. CNN and LSTM had the highest accuracy of 93.3 percent among all of the models. All of this model can categorize the seven types of news. We built an extensive system that can categorize Bengali textual input data and provide prediction performance.

5.3 Future Works

Seven classes news was applied for all of the models Our future objective is to construct a superior neural organize and advance the informational index with more class of information. Enhancing more classification for building the model which will assist with making a framework.

REFERENCES

[1] Takie Mohammad Jubayer, “DIGITAL AGE NEWSPAPER MODEL FOR BANGLADESH”, Academia.

[2] Nayan Banik, Md. Hasan Hafizur Rahman, “GRU based Named Entity Recognition Sys-tem for Bangla Online Newspapers”, International Conference on Innovation in Engineer-ing and Technology (ICIET) 27-29 December, 2018

[3] Abu Nowshed Chy, Md. Hanif Seddiqui, Sowmitra Das, “Bangla News Classification us-ing Naive Bayes classifier”, 16th Int'l Conf. Computer and Information Technology, 8-10 March 2014, Khulna, Bangladesh.

[4] Rabindra Nath Nandi, M.M. Arefin Zaman, Tareq Al Muntasir, Sakhawat Hosain Sumit, Tanvir Sourov and Md. Jamil-Ur Rahman, “Bangla news recommendation using doc2vec”, International Conference on Bangla Speech and Language Processing (ICBSLP), 21-22 September, 2018.

[5] Md. Hanif Seddiqui, Md. Nesarul Hoque, Md. Hasan Hafizur Rahman, “Semantic Annota-tion of Bangla News Stream to Record History”, 18th International Conference on Com-puter and Information Technology(ICCIT), 21-23 December,2015.1987 [Digests 9th An-nual Conf. Magnetics Japan, p. 301, 1982].

[6] Chenbin Li, Guohua Zhan, Zhihua Li, “News Text Classification Based on Improved Bi-LSTM-CNN” , 2018- 9th International Conference on Information Technology in Medicine and Education.

[7] Xuefeng Xi and Guodong Zhou, “A Survey on Deep Learning for Natural Language Processing,” ACTA AUTOMATICA SINICA, vol.42(10), 2016, pp.1445-1465

[8] Tang D, Qin B and Liu T, “Deep learning for sentiment analysis: successful approaches and future challenges,” John Wiley & Sons Inc,2015.

[9] Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao, “Recurrent Convolutional Neural Networks for Text Classification” Vol. 29 No. 1 (2015): Twenty-Ninth AAAI Conference on Artificial Intelligence.

[10] Prof. Mr. Nihar M. Ranjan, Yash R .Ghorpade, Gauri R. Kanthale, Adishree R. Ghorpade, Abhishek S. Dubey, "Document Classification using LSTM Neural Network", Computer Engineering, Sinhgad Institute of Technology and Science, India,Journal of Data Mining and Management Volume 2 Issue 2.

[11] Krishnalal G, S Babu Rengarajan, K G Srinivasagan, "A New Text Mining Approach Based on HMM-SVM for Web News Classification", 2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 19.

[12] Mohammad Rabib Hossain, Soikot Sarkar, Moqsadur Rahman, "Different Machine Learning based Approaches of Baseline and Deep Learning Models for Bengali News Categorization", International Journal of Computer Applications (0975 – 8887) Volume 176 –No. 18, April 2020.

[13] Ronald Tudu,Shaibal Saha,Prasun Nandy Pritam,Rajesh Palit,"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).

[14] Li Deng, Dong Yu, “Deep Learning: Methods and Application”, published by Microsoft, 2014.

[15] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, Aug. 1988.

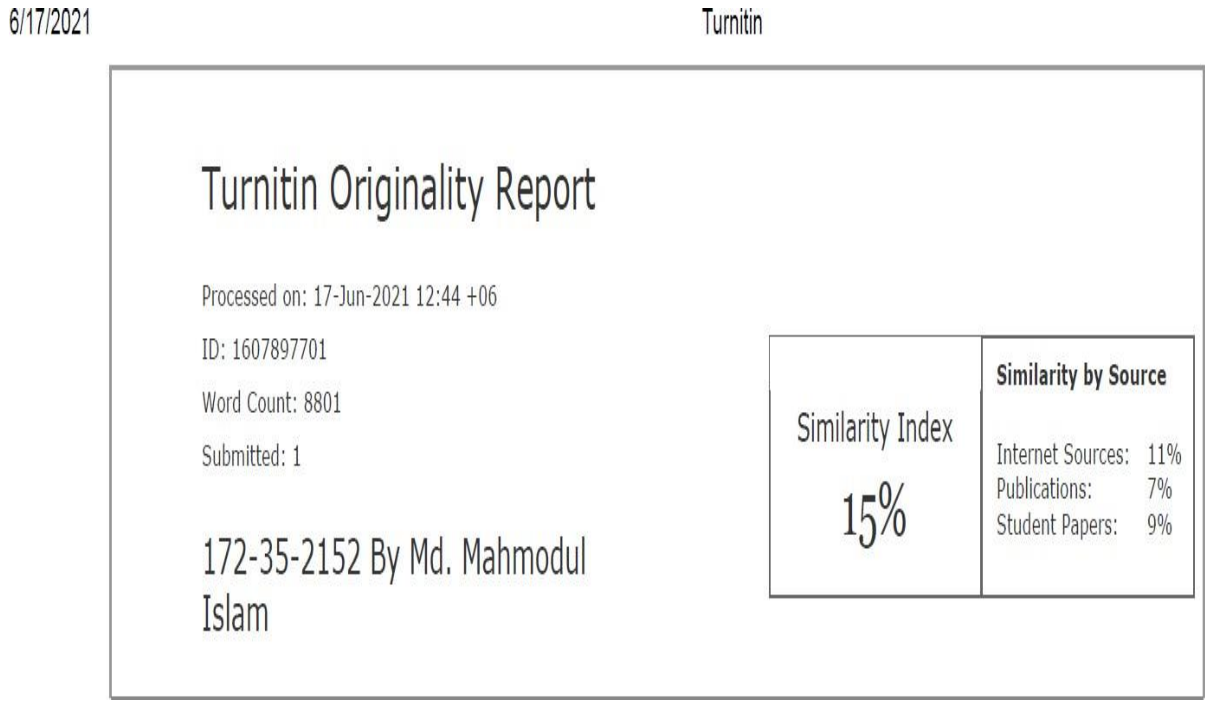
[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119.

APPENDIX: A

To finish the project, I had to deal with a slew of issues, the first of which was determining our project's research strategy. It was not a conventional research; it was a research-based project, and there had been previous work in this field that was primarily focused on machine learning algorithms and a tiny number of datasets. Another issue was that data collecting was a significant barrier for me. There was a dataset available that was not much larger, therefore I collected a large dataset over a long period of time and constructed a best fit model. It's fascinating to work with such a vast dataset.

APPENDIX: B

Plagiarism Report



6/17/2021 Turnitin

Turnitin Originality Report

Processed on: 17-Jun-2021 12:44 +06

ID: 1607897701

Word Count: 8801

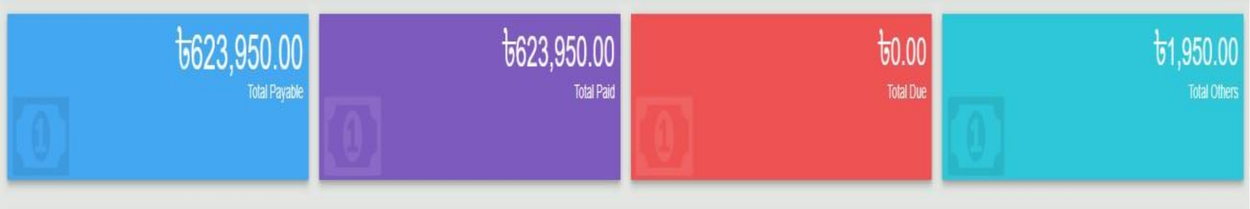
Submitted: 1

172-35-2152 By Md. Mahmodul Islam

Similarity Index	Similarity by Source
15%	Internet Sources: 11% Publications: 7% Student Papers: 9%

Plagiarism Report Drive Link:
https://drive.google.com/file/d/1OZiIW8jko6lE7HJjWZAPd5kY3G3GyeXh/view

Accounts Clearance:



৳623,950.00
Total Payable

৳623,950.00
Total Paid

৳0.00
Total Due

৳1,950.00
Total Others