

BENGALI DOCUMENT CATEGORIZATION USING DEEP LEARNING APPROACH

BY

**Sanjida Akter Akhi
ID: 183-15-11863**

AND

**Mus.Fatima Tuzzahura Talukder Sathi
ID:183-15-11891**

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

Supervised By

Mr. Saiful Islam
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Md. Jueal Mia
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

AUGUST 2022

APPROVAL

This Project titled “**Bengali Document categorization using Deep learning approach**”, submitted by **Sanjida Akter Akhi** , Id- 183-15-11863 and **Mus.Fatima Tuzzahura Talukder Sathi**, Id- 183-15-11891 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on August 2022.



Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



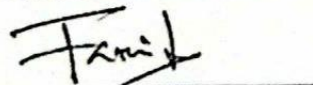
Nazmun Nessa Moon (NNM)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Mr. Faisal Imran (FI)
Assistant pprofessor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner




Dr. Dewan Md Farid
Professor
Department of Computer Science and Engineering
United International University

External Examiner

DECLARATION

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


Supervised by:



Mr. Saiful Islam
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:


Md. Jewel Mia
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:


Sanjida Akter Akhi
ID: 183-15-11863
Department of CSE
Daffodil International University


Mus. Fatima Tuzzahura Talukder Sathi
ID: 183-15-11891
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We are really grateful and wish our profound indebtedness to **Mr. Saiful Islam, Sr. Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Deep Learning*” to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr.Touhid Bhuiyan, Professor and Head**, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

The rapid growth of social media and microblogging sites not only gives places for increasing free expression and individual voices, but also allows people to engage in anti-social conduct such as online harassment, cyberbullying, and hate speech. Several initiatives, mainly for highly resourced languages like English, have been proposed to leverage this data for social and antisocial behavior analysis, document categorization, and sentiment analysis by predicting scenarios. But when it comes to sub-sided languages such as the Bengali, Hindi, Urdu and many others, the researchers in the outgrowing field of Natural Language Processing suffers from a great amount of deal because of the lack of basic components and materials. In the case of our experiments, we have used a dataset of news data consisting of a total of 19137. The CNN-BiLSTM deep learning approach was used in the case of categorizing different classes. The main purpose of this work was to determine between the classes which could be helped in order to help the user's concussion to help individuals to identify in which categories the data resembles.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval Page	I
Declaration	Ii
Acknowledgements	Iii
Abstract	iv
List of Figures	ix
List of Tables	x
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 Introduction	1
1.2 Motivation	2
1.3 Rational of Study	2
1.4 Research Question	2
1.5 Expected Outcome	3
1.6 Report Layout	3
CHAPTER 2: BACKGROUND STUDY	5-9
2.1 Introduction	5
2.2 Related Works	6
2.3 Comparative Analysis and Summary	7
2.4 Scope of the Problem	8
2.5 Challenges	9

CHAPTER 3: RESEARCH METHODOLOGY	10-23
3.1 Research Subject and Instrumentation	10
3.4 Data Collection Procedure	12
3.6 Static Analysis	19
3.7 Proposed Methodology	20
3.8 Implementation Requirements	23
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	24-27
4.1 Experimental Setup	24
4.2 Experimental Results and Analysis	25
4.3 Result Discussion	26
CHAPTER 5: IMPACT on SOCIETY, ENVIRONMENT AND SUSTAINABILITY	28-29
5.1 Impact on Society	28
5.2 Impact on Environment	28
5.3 Ethical Aspects	28
5.4 Sustainability Plan	29
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	30-31
6.1 Summary of the Study	30

6.2 Conclusion	30
6.3 Implicatin for Further Study	31
APPENDIX	32
REFERENCES	33-34

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1.1: LSTM Cell Base Structure	10
Figure 3.3.1: Flow Chart of Bangla Documentation Categorization	12
Figure 3.4.1: The Starting Part of Our Data	13
Figure 3.5.1: Collected Newspaper Data Static Diagram	13
Figure 3.6.1: Pie Chart of Dataset Distribution Split	14
Figure 3.7.1: Dataset Length Frequency Distribution	15
Figure 3.9.3.1: Normal News and After Cleaning News	17
Figure 3.9.4.1: Tokenizer Info	18
Figure 3.10.1: Static Analysis	19
Figure 3.11.1.1: Diagram of LSTM	20
Figure 3.11.2.1: Diagram of CNN	21
Figure 3.12.1: Pad Sequencing	22
Figure 3.12.2: Model Structure of CNN-BiLSTM	22
Figure 4.2.1: Training and Validation Accuracy	24
Figure 4.3.1.1: Precision, Recall & F1-score on deep learning model	26
Figure 4.3.2.1: Confusion matrix of CNN-BiLSTM model	27

LIST OF TABLES

TABLES	PAGE NO
Table 3.8.1: Our Collected Data from Online News	16

CHAPTER 1

INTRODUCTION

1.1 Introduction

Microblogging systems and social networking sites have risen tremendously in recent years, allowing its users to express themselves[1] Thoughts Simultaneously, they have permitted antisocial conduct [2], digital harassing, stalking, false political and religious rumors, and racial hatred activities [3,4]. The anonymity and mobility provided by such media have facilitated the cultivation and spread of hate speech [5], eventually leading to hate crime in a variety of areas such as religious, political, geopolitical, personal, and gender abuse. This is true for all people, regardless of language, geography, or race. While the Internet originated as mostly an English phenomenon, it today contains material in numerous languages. Languages are becoming extinct at an alarming rate in the real world, and the globe may lose more than half of its linguistic variety by the year 2100. Languages are dynamic, spoken by groups whose lives are molded by an ever-changing world. The globe has around 7,099 live languages, which are continually changing. A third of such languages are currently endangered, with fewer than 1,000 speakers left in many cases. Meanwhile, only 23 languages account for more than half the world's population.

In our work, we utilize definitive Bangla news texts from many online newspapers portals to comprehend the feelings of these articles. Where we will arrange the texts into 12 gatherings. These 12 kinds of information will be perceived while our model can be effectively educated. We're achieved in this work through and through Bangla text. We have fostered a framework that can separate effectively between the information given in Bangla.

Bengali, a rich and diverse languages, is spoken in Bangladesh, the second most prevalent language in India, and the seventh most common language in the world, with about 230 million users (200 million native speakers) [6]. Sentiment analysis takes advantage of the polarity represented in texts, and lexical resources are frequently employed to seek up certain terms whose existence might be predictive. Linguistic

characteristics, on the other hand, use syntactic information as features such as Part of Speech (PoS) and certain dependence connections. The practice of organizing documents into multiple classes or categories, known as content-based

categorization, is evolving as a result of the continual expansion of digital data. Aside from this, sentiment analysis in Bengali is increasingly being regarded as a difficult job, with earlier techniques attempting to discern the general polarity of Bengali texts. Which and why we needed the outcome to give machines the capability to understand the different variations in linguistic data.

1.2 Motivation

Whenever it comes to NLP research, scientists have long been dominated by the ability to develop and research on systems in order to understand and predict human behavior. This requirement can only be understood through language. Readers are simply drawn to the news articles of their advantage. The reader has to explore all the news to find their desired news. For this reason they have to pass lots of functions. So it will be great if they get all their news at the same place. That is why we wanted to focus on this study genre for future advancement. In terms of document categorizing, it is frequently mistaken with the phrase Contemplating. So, in order to make a clear distinction between our work, we became compelled to concentrate on this particular issue. By this means while concentrating on giving our dataset a head up progression towards categorizing, we specifically chose the deep learning approach to put a contrast to the base of our work.

1.3 Rational Study:

With the help of deep learning there is a lot of work that has been done in the last few years. But in Bangla language it has limitations. We can see that most of the research was conducted with different techniques and algorithms. So, with the help of our proposed method will execute the document categorization with less cost. On the other hand, we'll get good accuracy.

1.4 Research Question

- How have we collected data?
- How hard was it to preprocess data?
- Did it give better accuracy?
- How will it give more perfect performance than other works?
- Which algorithm we've used?
- How will it impact our real life?
- Should we use more algorithms?

1.5 Expected Outcome

- People can easily find the news by their category.
- It will help people to identify the news by document.
- It'll be used to find Bangla news correctly with a deep learning approach.
- With the help of this model we can develop the existing model.
- The researcher will get help from it for better improvement.

1.6 Report Layout

This report varied in a total of six different chapters. Which are capable of extending the understanding of “Document Categorization” more briefly.

In the first chapter, we'll mention introduction, motivation, rational study, research questions and the last one is the expected outcome.

In the second chapter, we'll brief about some related works, which types of challenges that we had faced and about the research summary.

In the third chapter, we'll talk about our research subject and instrumentation, workflow of the model, how we've collect our data, collected data insight details, data pre-processing techniques, how we build and train our model, LSTM,, CNN-LSTM and what requirements that we need to implement our research.

In the fourth chapter, we'll talk about the accuracy that we got , the evaluation of our model and the comparison with other models.

In the fifth chapter , We'll describe its impact on our society,impact on our environment and sustainability.

In the sixth chapter, which is our last chapter, we'll mention the conclusion and our future works.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

Although other major languages have had extensive and well-studied linguistic analyses, only a few approaches have been explored for Bengali [6], owing to a lack of a systematic method of text selection, annotated corpora, name lexicon, morphological analyzers, and an overall research outlook. Existing work on Bengali NLP focuses primarily on document categorization [7] using classifier or using the N-gram technique to categorize newspaper cnns [8], analysis of the data preparation [9] for aspect-based sentiment [10], and word vectors for document classification [11].

For one-to-four-character n grams, an LR-based technique is provided for identifying tweets classified as racist, sexist, or neither [12] and for separating hateful and unpleasant but not hateful retweets utilizing L2 regularization [13]. The authors used word-level n-grams as well as numerous PoS, emotion, and tweet-level metadata elements for the latter use case. However, the precision of these solutions is inadequate [14]. Davidson et al. [15] employed DNN techniques using two binary classifiers: one to predict the existence of abusive communication in general and another to distinguish the kind of abusive speech or developing a classifier made of two distinct networks for hate speech detection [3].

2.2 Related Work

The majority of existing classifiers are supervised [16]: LR is the learning rate.while other algorithms, such as SVM, NB, and RF are also available.used. Despite the introduction of a variety of features, nothing is known about them.about the multimodality of many types of variables in a single classifier by concatenating all of them, most approaches just 'use them all.'all feature types into sparse, high-dimensional feature vectors are prone to overfitting, particularly when dealing with brief messages like tweets. SomeTo cut costs, we used an automated statistical feature selection

technique. Others accomplished this by hand and optimized the feature space, while others did it automatically. Numerous research has already utilized neural network approaches to study hate speech; Badjatiya et al. performed extensive trials with various deep learning architecture to acquire meaningful phrase segmentation to manage poisonous remark recognition [19].

Though each type of connection has demonstrated efficacy for general-purpose text categorization, few studies have looked into merging them. Both architectures are combined into a single network [17], with the exception that when combined, and especially when transferred learning is used, both neural networks in classification accuracy a classifier by itself [18]. Our strategy, on the other hand, is based on a single word. Embedded similar to fastText, but with a significant focus on low-resource settings by minimizing the amount of needed parameters the training time required, while also improving performance and robustness across similar classification tasks. Text processing technologies must handle a lot of document operations quickly and correctly in order to handle this volume of data. Human language processing's core issue of text classification is crucial to numerous applications, including topic categorization, text analytics, filtration, and online search [20]. Online consumer content that is harmful, harmful, or insulting, such as hate speech, profanity, harassment, etc., is referred to as toxic comments. [21,22]. In a similar spirit, the terms ``Abusive `` and ``Hateful `` are changed to "Toxic," while the term "Normal" is changed to "non-Toxic." We eliminate the cases of "Spam" since, in accordance with the suggested category of "hate-based rhetoric" in [23], they do not constitute poisonous remarks. Following comparison, Twitter 18k has 18,625 tweets overall, 5,814 of which are toxic and 12,811 of which are not. Twitter 42k has 36,609 regular tweets and 5,705 harmful tweets. The fourth dataset, referred to as Wiki [24], was compiled from the Wikipedia Discussion page and annotated using a multi-label classification technique. Muhammad Ashfaq Khan, Rezaul Karim and Yangwoo Kim propose an effective examination system[1], which is in fact a dynamic AI method converged with Spark-based straight models, Multilayer Perceptron (MLP) and LSTM, utilizing a two-stage overflow structure to improve the prescient precision. Our proposed design empowers us to coordinate huge information examinations in a versatile and proficient manner. Another outstanding work that was for direct hate speech,

informal or angrier. Mai ElSherief, Elizabeth Belding, Dana Nguyen, Vivek Kulkarni and William Yang Wang work for hate speech[2]. Also extend how they might interpret online disdain discourse by zeroing in on a to a great extent ignored yet vital part of disdain discourse - - its objective: either coordinated towards a particular individual or substance, or summed up towards a gathering sharing a typical safeguarded trademark. Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgilio A. F. Almeida, and Wagner Meira Jr. 2018. Characterizing and Detecting Hateful Users on Twitter[4]. They create and utilize a vigorous procedure to gather and clarify scornful clients which doesn't rely straightforwardly upon vocabulary and where the clients are explained given their whole profile. This is an example of Twitter's retweet diagram containing 100, 386 clients, out of which 4, 972 were clarified. We additionally gather the clients who were restricted in the three months that followed the information assortment. We show that disdainful clients contrast from typical ones as far as their action designs, word use and as well as organization structure. They acquire comparative outcomes looking at the neighbors of scornful versus neighbors of typical clients and furthermore suspended clients versus dynamic clients, expanding the vigor of our examination. They see that derisive clients are thickly associated, and in this manner plan the disdain discourse identification issue as an undertaking of semi-managed learning over a chart, taking advantage of the organization of associations on Twitter. Ziqi Zhang, David Robinson, and Jonathan Tepper(2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Neural Network[5]. They present another technique based on a profound brain network joining convolutional and gated recurrent networks. We lead a broad assessment of the technique against several baselines and cutting edge on the biggest assortment of publicly available Twitter datasets to date, and show that contrasted with previously revealed results on these datasets.

2.3 Research Summary

We studied 24 research papers which are using Machine Learning & Deep learning approaches. For our research we took help from the online news portal. We collected data from different online newspapers. Those are: jugantor, prothom alo, news24, Jagonews24.com, Bangla tribune, bdnews24.com, BBC Bangla, Dhaka post,

bd24live.com. In our research, we used CNN- BiLSTM as a deep learning algorithm in our research. We tried to find out the accuracy. Out of these two algorithms, we got good results. From other research paper we got to know that most of the worked for abuse word,bad word.

2.4 Scope of the Problem:

This study discusses and gives evaluations in presumption analysis of news data from several primary genres. The purpose of this test is to see if data can be ordered to display the classes Art, Science-Tech, Economics, Environment, International, Accidents, Opinions, Crime, Entertainment, Sports, Education, Politics, or Accidents, Opinions, Crime, Entertainment, Sports, Education, and Politics. In order to classify, a total of 12 classes were taken. Clients can converse and exchange their facts, opinions, and suppositions through informal communication destinations in the online world. The analysts also determined the value of a documentary text if it is certain subjective investigations, such as conclusion inquiry, to more easily appreciate the trial's outcome. It would put individuals ahead of the curve in this field of observation in terms of dynamicity and automation.

2.5 Challenges:

The most difficult challenge for us is to collect data. We've no idea how it'll happen. After that choose some online news portal which was in bangla language. Then started collecting data. 19 thousand data collected in a different category was not easy work. On the other hand, we didn't know how to do pre-processed data, how to tokenize, how to remove other words & punctuation. Moreover we had no knowledge about the CNN-BiLSTM process. We had a little knowledge about Python but that was not enough. We practiced more and more on python , CNN and LSTM algorithms. As it was totally new and unknown so it became a big challenge for us.

We have considered the slant analysis based on voyager inputs in regards to carrier organizations in this study. Our suggested method revealed that both element determination and over-inspecting methods are equally important in improving our

results. Using highlight choosing algorithms, we were able to recover the best selection of highlights while also reducing the number of calculations required to create our classifiers. It has, however, reduced the skewed appropriation of classes observed in several of our smaller datasets without creating overfitting. Our findings show that the suggested model has a high level of grouping precision when it comes to predicting how the 12 classes would be structured. Managing Bengali text and processing it for model training was also a difficult challenge. As can be observed, several of the applied classifiers have outperformed the others.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this part, I will quickly describe the steps I took to accomplish our study project. To hide the complex Deep learning algorithms, we donned several Python programming languages. The algorithms were chosen and deployed on our dataset on a regular basis since, in Natural Language Processing, the dataset is the most important component of the entire process. CNN-BiLSTM was the algorithm we used. Our objective in this study was to find an uncommon rule-based method. Figure 3.1.1 depicts the method through which we consider the operation while seeing a flow chart.

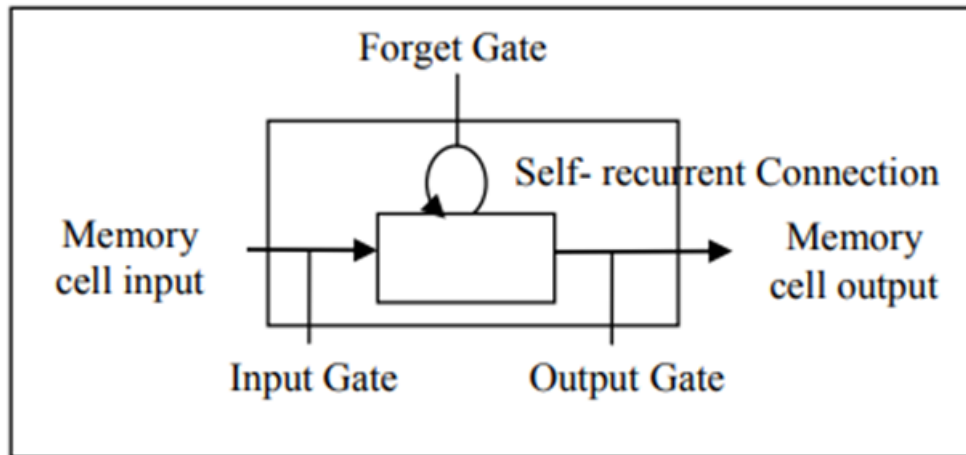


Figure 3.1.1 LSTM cell base structure[27]

The word - based output is provided to the model. The incremental ML model utilized in just this app has a first tier that integrates variables for the verbal ability, number of features of the sentences. As we only want one output in the end, the LSTM, which has 100 neurons per layer, is the next phase. This is preceded by a deep network with sigmoid. We have used the Adam method for adaptable estimation, binary crossed variance for loss calculation, and finally a fall out layer in between to avoid overfitting. The model was subsequently evaluated and trained.

3.2 Research Subject and Instrumentation

In terms of the name, we went with Document Categorization to symbolize the whole picture. Because the area of DL and NLP is expanding as a result of technical advancements, the space that this research will represent is relatively considerable when compared to previous efforts. The computer that was required to finish my project model was constructed to a high specification in order to manage such a long period of intensive study. To develop the system the programming language we used was named python. We import various kinds of libraries like Skit-learn, Keras, TensorFlow, matplotlib, Pandas, and NumPy

3.3 Workflow

To complete our research ,we've followed some steps. These steps are very important and one step comes from another one.

- **Data Collection:** We've collected our data from different online news portals. Such as: jugantor, prothom alo, news24, Jagonews24.com, Bangla tribune, bdnews24.com,BBC Bangla,Dhaka post, bd24live.com.Our collected amount of a certain dataset contains 19137 data's which build up sentences. The sentences are profoundly collected from numerous sources such as news portal, media sector etc. The dataset was divided into 12 different classes.
- **Data Distribution:** After all of the scrubbing and removing of all the untamed data, the dataset was finally ready to be divided into classes for training and testing.
- **Data Preprocessing:** After collecting data now we'll proceed to our next step. As we've collected our data from online so there are lots of punctuations, unique words, special characters . With the help of python we'll handle these words.
- **Model Building:** After preprocessing steps we'll complete our model building part. For achieving accuracy now we'll train & test data.
- **Evaluation:** After all steps now we'll evaluate the results.

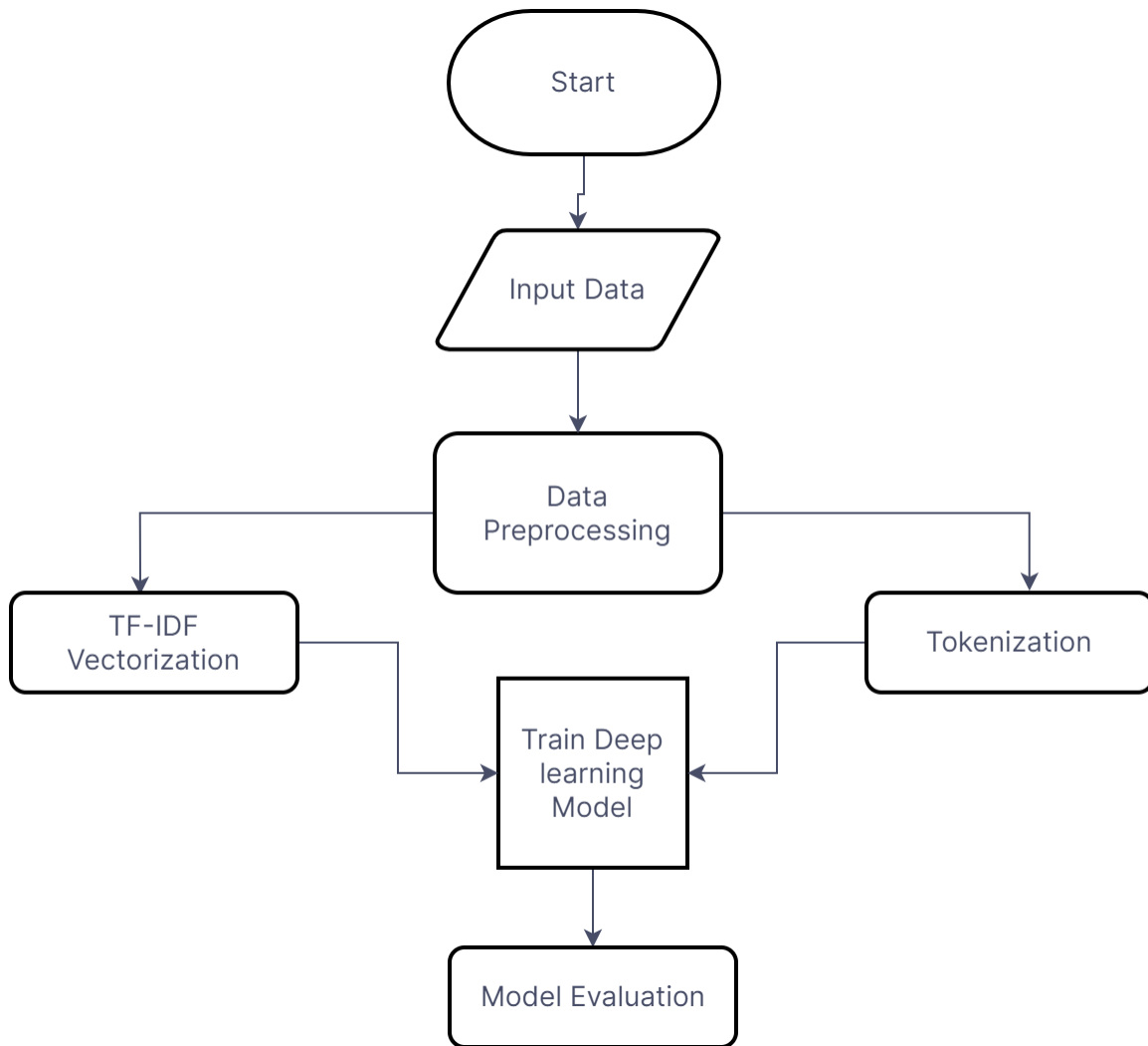


Figure 3.3.1: Flow Chart of Bangla Documentation Categorization using Deep Learning

3.4 Dataset Collection Procedure

Data is the most important part for research. Because without the data we couldn't train and test data for the project. We collected our data from the online news portal. Such as : BBC Bangla,Jugantor,Prothom alo,Bdnews24.com. We've added 3 columns in our excel sheet .News,news type and label.

	label	text	is_valid
0	opinion	বাংলাদেশের রাজনীতিবিদদের সহিংস, লোভী, জঘন্য, ন...	True
1	opinion	এই প্রতিবেদনে ভবন ধসের কারণ হিসাবে নয়টি বিষয়কে...	True
2	opinion	গতকাল রোববার নৌ নিরাপত্তা সপ্তাহের শুরুর দিনেই...	True
3	opinion	আমাদের মহান ভাষা আন্দোলন থেকে শুরু করে ১৯৬২ স...	True
4	opinion	চট্টগ্রাম, মার্চ ১৫ (বিডিনিউজ টোয়েন্টিফোর ডটকম...	True

Figure 3.4.1: The starting part of our Data

3.5 Data set Distribution:

We've collected 19000 data for our research. Divided into 12 classes. Those are: Art, Science_Tech, Economic, Environment, International, Accident, Opinion, Crime, Entertainment, Sports, Education, Politics. The chart of our data distribution is given below:

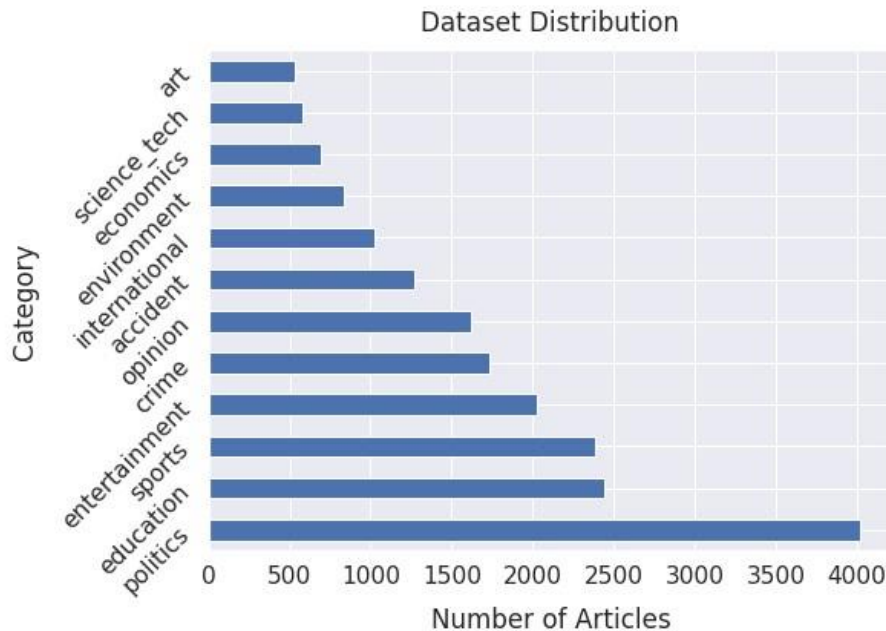


Figure 3.5.1: Collected Newspaper Data Statistic Diagram

3.6 Dataset Distribution Split

After all of the scrubbing and removing of all the untamed data, the dataset was finally ready to be divided into classes for training and testing. The whole quantity of data was divided in half, 90/10. The train data size was 13679 bytes, whereas the test data size was 1900 bytes. A pie chart is given below

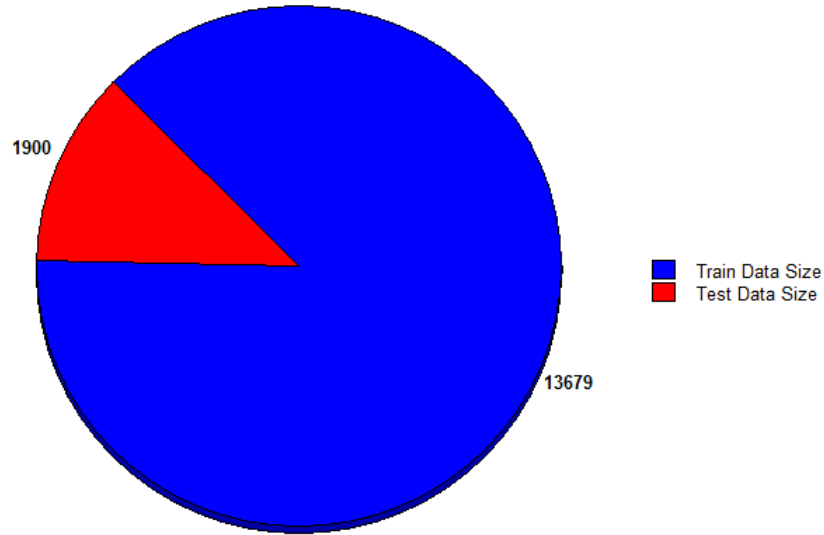


Figure 3.6.1: Pie Chart Dataset Distribution Split .

3.7 Length Frequency Distribution

The impact of report length on the classification performance is likewise concentrated as a piece of this work. It can be asserted that messages containing long or potentially many sentences may have more data which can guide arrangement calculations to work better. In that case, classifying longer reports ought to get superior performance than of more limited archives.

Here, the length of the reports are estimated in words.

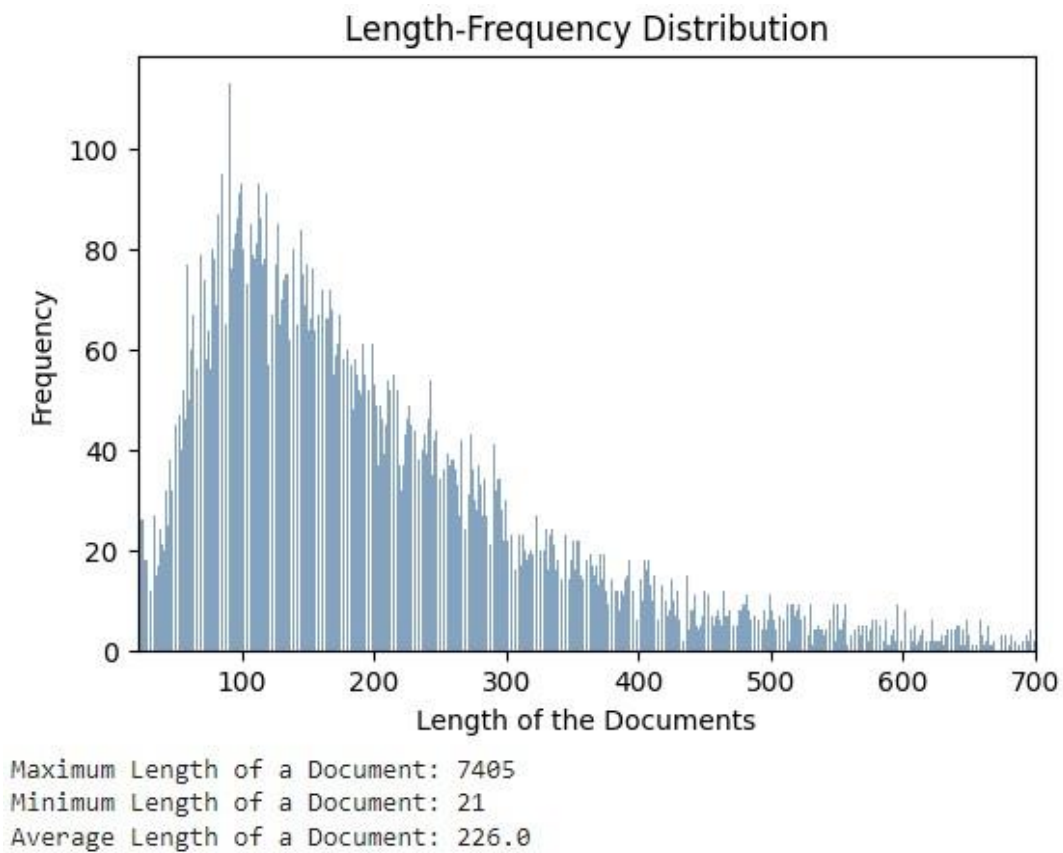


Figure 3.7.1: Dataset length frequency distribution

3.8 Data Details

Our data set contains 3 columns, News types, News and Label. We've collected data into 12 classes. As this was based on the Bangla language so it was a little bit difficult to collect. We've collected 19 thousand pieces of news . The news collecting information is given below.

Table 3.8.1: Our Collected Data Set From Online news

News Type	Data
Politics	4012
Education	2368
Sports	2360
Entertainment	2012
Crime	1718
Opinion	1619
Accident	1015
Environment	830
Economics	688
International	1264
Art	532
Science_Tech	581

3.9 Data Pre-Processing

To prepare the dataset for the train case of the algorithms, we had to delete the junk character from the entire dataset. Which are special characters ("!", "@", "#", "\$", "percentage ", "*") numerical characters (1, 2, 3, 4, 5, 6, 7, 8, 9, 0), white space, and duplicate characters? The duplicate character might possibly be kept because it was reconciled many times when accepting the data set. It is critical that the dataset be raw while training the machine so that it can recognize the differences between the classes.

3.9.1 Stop Word Remove

A stop word is a commonly used word, such as ".", ",", ":", "||," and so on, for which a web index was employed in the case of alteration to make the model miss these little usurp symbols. So that the model may perform to its greatest potential in order to get a higher turnout result.

3.9.2 Whitespace Remove

There is a lot of white space in the Bangla Language. So for removing those we've to use python regular expressions.

3.9.3 Special Character & Punctuation

For different types of punctuation it became challenging to work on the Bangla language. Because this language cleaning process is more difficult and time-consuming.

	text	label	cleaned
0	বাংলাদেশের রাজনীতিবিদদের সহিংস, লোভী, জঘন্য, ন...	opinion	বাংলাদেশের রাজনীতিবিদদের সহিংস লোভী জঘন্য নোংর...
1	এই প্রতিবেদনে ভবন ধসের কারণ হিসাবে নয়টি বিষয়কে...	opinion	প্রতিবেদনে ভবন ধসের নয়টি বিষয়কে চিহ্নিত হয়েছে ...
2	গতকাল রোববার নৌ নিরাপত্তা সপ্তাহের শুরুর দিনেই...	opinion	গতকাল রোববার নৌ নিরাপত্তা সপ্তাহের শুরুর দিনেই...
3	আমাদের মহান ভাষা আন্দোলন থেকে শুরু করে ১৯৬২ স...	opinion	মহান ভাষা আন্দোলন ১৯৬২ সালের শিক্ষা আন্দোলন ...
4	চট্টগ্রাম, মার্চ ১৫ (বিডিনিউজ টোয়েন্টিফোর ডটকম...	opinion	চট্টগ্রাম মার্চ ১৫ বিডিনিউজ টোয়েন্টিফোর ডটকম দ...
...
19132	শুক্রবার রাতে-- উৎসবের ফেইসবুক পেইজে এ ঘোষণা দ...	entertainment	শুক্রবার রাতে উৎসবের ফেইসবুক পেইজে ঘোষণা ইভেন্ট...
19133	বলাকা সিনেওয়ার্ল্ড (৩, মিরপুর রোড, ঢাকা) বলাকা...	entertainment	বলাকা সিনেওয়ার্ল্ড ৩ মিরপুর রোড ঢাকা বলাকা ১ প...
19134	এ যেন একেবারে হাতে লাল টিসি ধরিয়ে দিয়ে বলা নি...	entertainment	একেবারে হাতে লাল টিসি ধরিয়ে দিয়ে নিজস্ব হ...
19135	গ্লিটজকে শাবনুর বললেন, "নিশ্চয় নানা কথা রটি...	entertainment	গ্লিটজকে শাবনুর নিশ্চয় নানা কথা রটিয়েছে কানে এসে...
19136	ঢাকা, অক্টোবর ১৮ (বিডিনিউজ টোয়েন্টিফোর ডটকম)- ...	entertainment	ঢাকা অক্টোবর ১৮ বিডিনিউজ টোয়েন্টিফোর ডটকম দেশে...

Figure 3.9.3.1: Normal News and After cleaning News

3.9.4 Tokenization

Tokenization is the process of breaking down a sentence's argument into its constituent parts, which are referred to as tokens. Tokenization is critical in such procedures because it allows the algorithm to train on the data.

```
===== Tokenizer Info =====  
Words --> Counts:  
হয়েছে      20402  
এক         18196  
বিডিনিউজ      16267  
টোয়েন্টিফোর 15116  
জানান      13797  
কথা        13436  
ডটকম      12605  
ঢাকা       11994  
oo         11130  
সংবাদ     10051  
  
Words --> Documents:  
এক         9970  
হয়েছে     8455  
ঢাকা       7359  
বিডিনিউজ      7253  
জানান      7180  
কথা        7000  
টোয়েন্টিফোর 6766  
ডটকম      6463  
গত         6091  
সময়       5486
```

Figure 3.9.4.1: Tokenizer info

3.10 Statistical Analysis

1. In the dataset total 18999 data is presented.
2. The dataset is divided into 12 classes.
3. 13679 data is used for the train.
4. 1900 data is used for the test.

Highest accuracy achieved 95%.

There is a pie chart of our static analysis is given below:

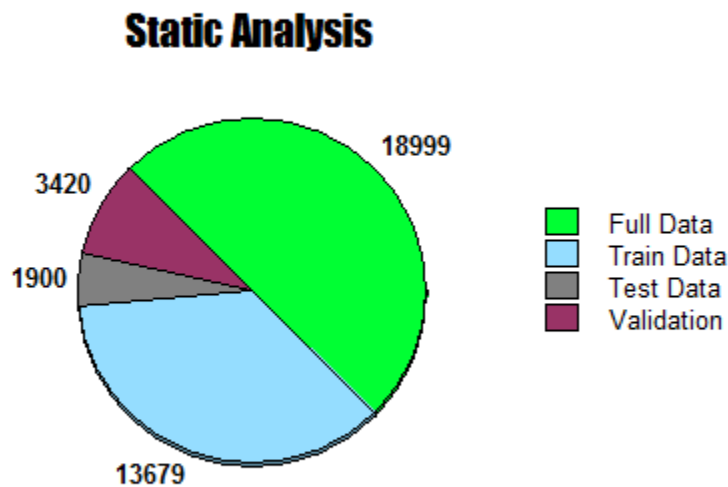


Figure 3.10.1: Static Analysis

3.11 Proposed Methodology

For our research we've used CNN-BiLSTM algorithms. But first of all we have to know about two algorithms. Such as: CNN & LSTM. There are many algorithm but we've used CNN-BiLSTM algorithm.

3.11.1 LSTM

Long Short-Term Memory (LSTM) networks are a kind of repetitive brain network equipped for learning and requesting reliance on grouping expectation issues (Jason, 2021). This is a conduct expected in complex issue spaces like machine interpretation, discourse acknowledgment, and the sky's the limit from there. LSTMs are a complicated area of profound learning. It may very well be difficult to get your hands around what LSTMs are, and the way in which terms like bidirectional and grouping-to-succession connect with the field. The word - based output is provided to the model. The

incremental ML model utilized in just this app has a first tier that integrates variables for the verbal ability, number of features of the sentences. As we only want one output in the end, the LSTM, which has (100 neurons per layer, is the next phase. This is preceded by a deep network with sigmoid. We have used the Adam method for adaptable estimation, binary cross variance for loss calculation, and finally a fall out layer in between to avoid overfitting. The model was subsequently evaluated and trained.

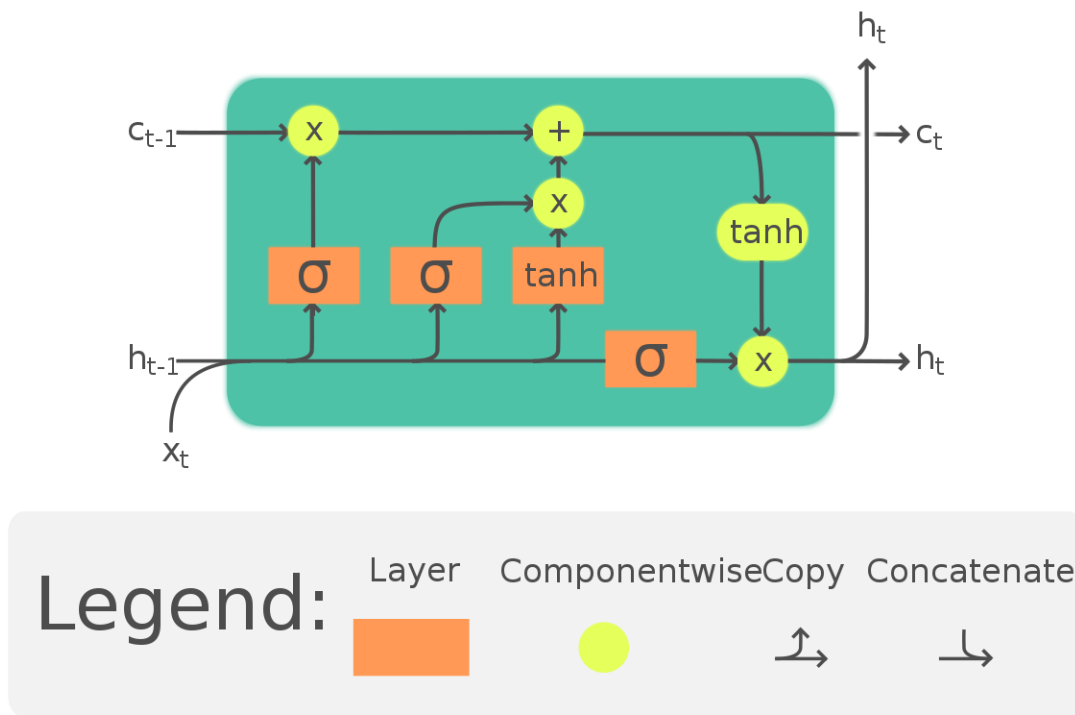


Figure 3.11.1.1 Diagram of LSTM[25]

3.11.2 CNN

Convolutional brain organization (CNN) is a piece of profound brain network that has become controlling in a few PC vision errands for instance breaking down visual pictures [21]. CNN naturally distinguishes the significant elements with no human communication [21]. To this end CNN would be an optimal answer for PC vision and picture order issues. CNN is likewise computationally effective. CNN uses 3-layered layers where the past layer is associated with a modest quantity of the neurons [21]. It

applies extraordinary convolution and pooling activities and acts boundary sharing [21]. This licenses CNN models to run on any gadget, building them more captivating [21]. Lately CNN is utilized for text classification and numerous scientists involved CNN for text order. Thus, we picked the CNN network for our Bangla text arrangement based Bangla brief tale classification issue.

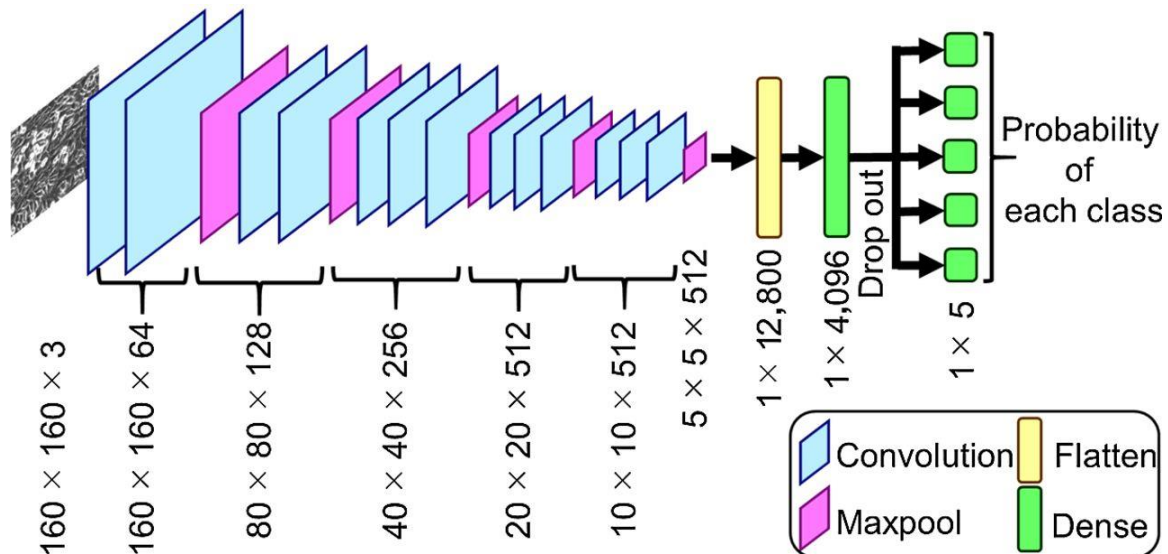


Figure 3.11.2.1 Diagram of CNN[26]

3.12 Model Development

In our research we used deep learning algorithms called CNN-BiLSTM. We can characterize a CNN LSTM model in Keras by first characterizing the CNN layer or layers, enveloping them by a Time distributed layer and afterward characterizing the LSTM and yield layers. We have two methods for characterizing the model that are the same and just vary as an issue of taste. We can characterize the CNN model first, then add it to the LSTM model by enveloping the whole succession of CNN layers by a Time distributed layer, as follows: We know that Tokenization is the process of breaking down a sentence's argument into its constituent parts, which are referred to as tokens. Tokenization is critical in such procedures because it allows the algorithm to train on the data. After tokenization we found 188977 unique tokens. After tokenization we do pad sequencing to get our desire length.[21]

```
#===== Pad Sequences =====
corpus = keras.preprocessing.sequence.pad_sequences(sequences, value=0.0,
                                                    padding='post', maxlen= padding_length)
print("\n\t\t\t\t==== Paded Sequences =====\n",dataset.cleaned[10],"\n",corpus[10])
```

Figure 3.12.1 Pad Sequencing

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, 300, 128)	640000
conv1d (Conv1D)	(None, 296, 128)	82048
max_pooling1d (MaxPooling1D)	(None, 59, 128)	0
bidirectional (Bidirectional)	(None, 59, 128)	98816
bidirectional_1 (Bidirectional)	(None, 59, 128)	98816
dense (Dense)	(None, 59, 28)	3612
dense_1 (Dense)	(None, 59, 14)	406
flatten (Flatten)	(None, 826)	0
dense_2 (Dense)	(None, 12)	9924
Total params: 933,622		
Trainable params: 933,622		
Non-trainable params: 0		

Figure 3.12.2: Model Structure of CNN-BiLSTM

We used 9 layers in our CNN-BiLSTM model to train our data

1. Embedding Layer: This is the input layer of our model. We were using unique tokenized words, Embedding dimension = 128, and input length = 300.
2. Conv1D Layer: In this layer we were using filters = 128, kernel Size=64.
3. Maxpooling1D: After the 1D convolution layer we used Maxpooling1D for dimensionality reduction. Here we were using pool Size=2.
4. Two BiLSTM Layers = 64.

5. Dropout Layer: Dropout is a regularization method. We were using this layer for reducing overfitting and improving model performance. In our model we were using dropout=0.7.
6. Two dense Layers = 28,14
7. Dense Layer: This is the output layer of our model where we used class no = 12 and activation function = 'softmax'.

3.13 Implementation Requirements

We've used Jupyter Notebook for our data train testing steps. Simple to change over: Jupyter Notebook permits clients to change over the note pads into different arrangements like HTML and PDF. It likewise utilizes online devices and nbviewer which permits you to straightforwardly deliver a freely accessible scratch pad in the program. It is very easy to use and also we can use it like Google Colab. Based on our project we need the below mentioned requirements:

Hardware and Software Requirements:

- Internet Browser
- Operating System
- 16GB RAM and Intel core i78th generation.
- 2 TB Hard Disk.
- Windows 10

Development Requirements:

- Python 3.8
- Numpy and Pandas
- NLTK
- Jupyter Notebook

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In the instance of our identification of multiple classes, the algorithms are able to get a very volatile and acceptable level of result. The diagram of classification by our chosen algorithm model is shown in the graph below. Because the dataset is the most important aspect of the Natural Language Processing technique, the model was chosen and deployed on our dataset on a regular basis. When it comes to working with language datasets, the LSTM is more than enough. These are all advanced deep learning techniques that produced a better-than-expected outcome.

4.2 Experimental Setup

The dataset was quite efficiently and deliberately transformed in the scope of the deep learning technique to work with quite sufficient elements. The limitations of the LSTM were removed during the dataset preprocessing phase. The model worked just accurately while categorizing the data classes. The differentiation of classes were enough to help in case of the models perspective. The accuracy diagram is given below in terms of the showcase. The accuracy of 95% was achieved via our work contrast.

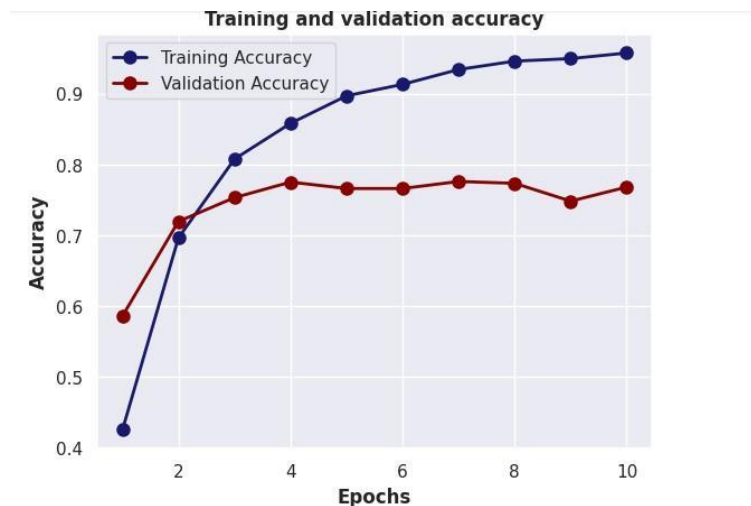


Figure 4.2.1: Training and Validation Accuracy

From the accuracy plot it is observed that the validation accuracy has not improved more than 85%, it is due to multiclass imbalanced classification problem. Moreover by proper tuning the vocabulary size the model performance can be improved.

4.3 Result Discussion

Now we'll give our explanation of how we've got our results on the deep learning approach. Our desired outcomes detailed in this piece depend on "F1 - score", which can be estimated by the consonant mean of accuracy worth and review esteem[23]

$$\textit{Precision} = TP/(TP+FP)$$

$$\textit{Recall} = TP/(TP+FN)$$

**TP= True Positive, FP= False Positive, FN= False Negative*

If we assume a class 'a' then,

Tp= Tp means 'Predicted Positive, Actual Positive'. That means It is correctly categorize to class a.[23]

FP= FP means 'Predicted Positive, Actual Negative'. It correctly categorize to class a but do not belongs to Class a.[23]

FN= FN means 'Predicted Negative, Actual Positive'. It is not correctly categorize to class a but it is belongs to class a.[23]

We can also measure the F1-score by using the below given equation:

$$F1=2*(\textit{precision}*\textit{recall})/(\textit{precision}+\textit{recall})$$

4.3.1 Result Discussion on Deep Learning Model

With the help of a bar chart ,we're going to show three different categories on CNN-LSTM deep learning model. We'll show Precision, F1-score and Recall, which we get from our project. As we've 12 classes then we'll mention our result in 12 classes sequentially.

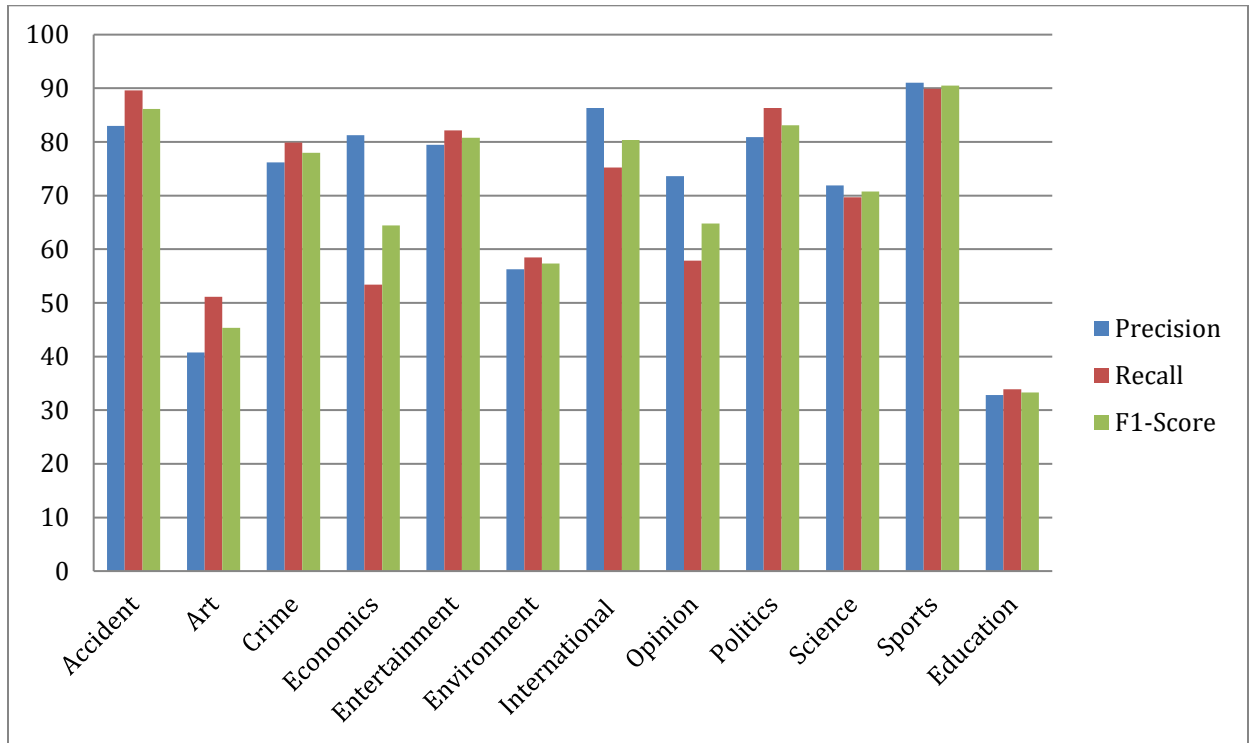


Figure: 4.3.1.1: Precision, Recall & F1-score on deep learning model

From the given pie chart of our 12 classes. We got to know the highest value and lowest value from our CNN-LSTM model . We have accomplished a decent exactness of 95% on this basic crossover brain network for Bengali record order task. This accuracy can be additionally improved by doing hyperparameter tuning and by utilizing more sophisticated network engineering with a huge dataset. By observing precision,recall and f1-score we can see that all the classes are classified reasonably well except **Art and Environment**.

4.3.2 Confusion Matrix

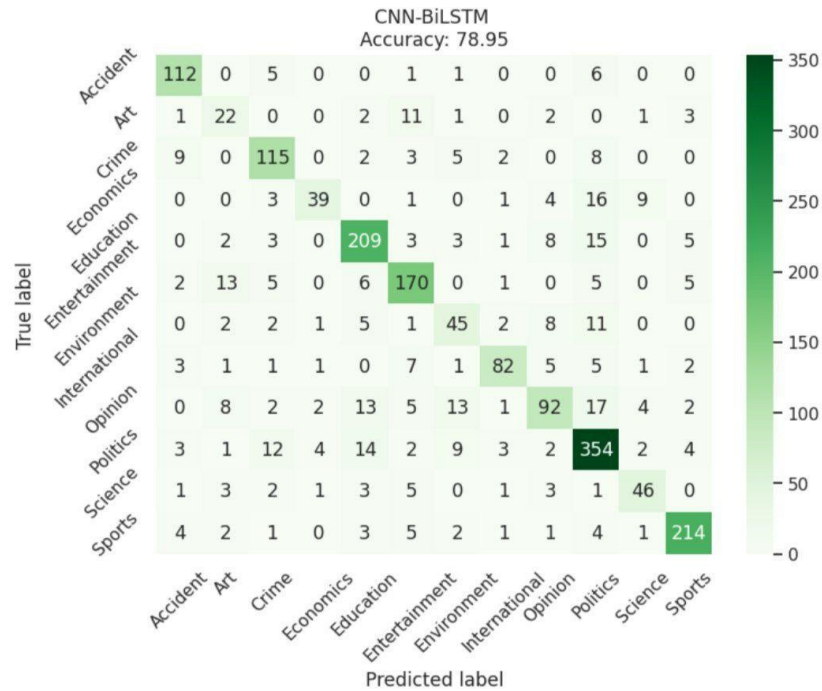


Figure 4.3.2.1 Confusion Matrix

The confusion matrix provides a good understanding about how many documents are correctly classified in each class and which classes get confused during classification. Here, we can see that the Art, Entertainment, Politics category gives a larger number of false classified result. When special characters were removed from texts, many of them became unintelligible, as they were also responsible for evaluation in the case of expression classification.

CHAPTER 5

IMPACT ON SOCIETY,ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Every human feeling may be linked to the words we view on a daily basis on various online platforms in the digital world. In this case, it is critical for these platforms to have a mechanism in place to discern which are genuine emotions and which are pre-programmed aggressiveness. This is why we've decided to focus our efforts on detecting one of the most fascinating genres of all time, such as Art, Science-Tech, Economics, Environment, International, Accidents, Opinions, Crime, Entertainment, Sports, Education, Politics, or Accidents, Opinions, Crime, Entertainment, Sports, Education, and Politics, or Accidents, Opinions, Crime, Entertainment, Sports, Education, and Politics. By doing so, we can expect to create a more definitive and diverse digital era.

5.2 5.2 Impact on Environment

It is fairly usual these days on various internet sites to share a social or political issue using a highly definitive word of mixed ridicule. However, categorization allows for the representation of a societal perspective on terminology in a variety of disciplines. Which is why we need be more cautious about technical conclusiveness, so that frameworks may completely comprehend which is which. Because there have been countless reports of individuals being able to comprehend various environmental flaws by employing comedy as the primary method. As a result, categorization has been shown to be both advantageous and detrimental to people's perceptions of the environment.

5.3 Ethical Aspects

The internet's media outlets have now become accessible to people of all ages. As a result, the conditions of the user limitations are no longer valid. Because there are insufficient security measures to distinguish between moral and social perspectives. One must be able to comprehend the overall context of a notion conveyed through platforms. In many circumstances, this has been shown to be harmful to people's moral ideals. The

algorithm might block legitimate communication while permitting a harmful one if it doesn't grasp the correct categories. Document Categorizations may be entertainment toward a sentimental experience, but too much of it might drive a user to withdraw from social platforms due to the overwhelming rules and constraints imposed on this procedure.

5.4 Sustainability Plan

- There are over 2.3 billion active internet-based life clients worldwide.
- At least two internet-based life cycles are present in 91 percent of large business brands.
- When they can't access their online life profiles, 65 percent of individuals feel uneasy and uncomfortable.
- As we were working with the Bangla language .Then it'll be very beneficial to the user.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

In the instance of our identification of multiple classes, the algorithms are able to get a very volatile and acceptable level of result. The diagram of classification by our chosen algorithm model is shown in the graph below. Because the dataset is the most important aspect of the Natural Language Processing technique, the model was chosen and deployed on our dataset on a regular basis. When it comes to working with language datasets, the LSTM is more than enough. These are all advanced deep learning techniques that produced a better-than-expected outcome.

6.2 Conclusions

We believe that this work is capable of providing a new addition to the work's notion of the ongoing developmental age of BNLNLP in the wide globe where emotional intelligence is becoming more aligned with all the turbulence between our daily life structure and habits. Categorization has long been a hotly debated issue among scholars in this field, and we hope that our contribution will encourage others. So that we might create a world in which technology aids one's mind in experiencing disaster. The model was able to achieve outstanding performance in orders accustomed. The achieved accuracy was 95% in total. The dataset was well processed in order to achieve this height of result. We would also recommend the deep learning approach towards such integrated research problems; machine learning might be preferable towards it if the dataset isn't well accommodated.

6.3 Recommendations

- Remove any bias value from the dataset.
- The percentage of data categories must be divided evenly.
- To get a better outcome, create a neural network.

- For categorization, parameter adjustment is required.

6.4 Implication for Further Study

Because of the rapid growth of information available on the internet and in online social media, businesses may now use conclusion analysis to gain insight into their consumers' feelings about their products or services. In current writing, document classification is typically based on little social media data, with only a few days' worth of data. Unless social media material is routinely retrieved, this obstacle prevents the acquisition of factually significant and meaningful .

APPENDIX

Abbreviations:

CNN= Convolutional Neural Networks.

LSTM= Long Short Term Memory

BiLSTM= BiDirectional Long Short Term Memory

BNLP = Bangla Neural Language Processing

REFERENCES

1. Muhammad Khan, Md Karim, and Yangwoo Kim. (2018). A Two-Stage Big Data Analytics Framework with Real World Applications Using Spark Machine Learning and Long Short-Term Memory Network.
2. Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth M. Belding. (2018). Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. From, <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17910>
3. Rohan Kshirsagar, Tyus Cukuvac, Kathleen McKeown, and Susan McGregor (2018). Predictive Embeddings for Hate Speech Detection on Twitter. arXiv preprint arXiv:1809.10644.
4. Manoel Horta Ribeiro, Pedro H. Calais, Yuri A. Santos, Virgilio A. F. Almeida, and Wagner Meira Jr. 2018. Characterizing and Detecting Hateful Users on Twitter. From <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17837>
5. Ziqi Zhang, David Robinson, and Jonathan Tepper (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Neural Network. In ESWC. Springer, 745–760.
6. MS Islam. (2009). Research on Bangla language processing in Bangladesh: progress & challenges. In 8th Intl. Language & Development Conf. 23–25
7. Md Islam, Fazla Elahi Md Jubayer, Syed Ikhtiar Ahmed, et al. (2017). A comparative study on different types of approaches to Bengali document categorization. arXiv preprint arXiv:1701.0869.
8. Munirul Mansur. 2006. Analysis of n-gram based text categorization for Bangla in a newspaper corpus. Ph.D. Dissertation. BRAC University
9. Md Atikur Rahman and Emon Kumar Dey. 2018. Datasets for Aspect-Based Sentiment Analysis in Bangla and Its Baseline Evaluation. Data 3, 2 (2018), 15.
10. SakhawatHosainSumit, Md Zakir Hossan, Tareq Al Muntasir, and Tanvir Sourov. 2018. Exploring Word Embedding For Bangla Sentiment Analysis. In Intl. Conf. on Bangla Speech and Language Processing (IMSLP), Vol. 21. 22
11. Adnan Ahmad and Mohammad Ruhul Amin. (2016). Bengali word embeddings for solving document classification problems. In 19th IEEE Intl. Conf. on ICCIT. 425–430.
12. eerak Waseem and Dirk Hovy. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proc. of the NAACL student research workshop. 88–93
13. WafaAlorainy, Pete Burnap, Han Liu, and Matthew William (2018). Cyber Hate Classification: Othering Language And Paragraph Embedding. arXiv preprint arXiv:1801.07495
14. Md Islam, Fazla Elahi Md Jubayer, Syed Ikhtiar Ahmed, et al. (2017). A comparative study on different types of approaches to Bengali document categorization. arXiv preprint arXiv:1701.08694.
15. Thomas Davidson, Dana Warmesley, Michael Macy, and Ingmar Weber (2017). Automated hate speech detection and the problem of offensive language. arXiv preprint arXiv:1703.04009 .
16. Xiaodong Wei, Hongfei Lin, Liang Yang, and Yuhai Yu (2017), A convolutional-lstm based deep neural network for cross-domain MOOC forum post classification. Information 8, 3 ,92.

17. Joni Salminen, Hind Almerekhi, and Milenkovic (2018) Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media.. In ICWSM. 330–339
18. R. IzsÁt'ak. (2015).Hate speech and incitement to hatred against minorities in the media. UN Humans Rights Council, A/HRC/28/64 .
19. Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, pages 759–760,
20. Charu C Aggarwal and ChengXiangZhai. (2012). A survey of text classification algorithms. In Mining text data. Springer, 163–222.
21. Learn About CNN-LSTM,available at<<<https://machinelearningmastery.com/cnn-long-short-term-memory-networks/>>>last last accessed on 15-08-2022 at 02:30pm.
22. Irene Kwok and Yuzhou Wang (2013).. Locate the hate: Detecting tweets against blacks. In Twenty-seventh AAAI conference on artificial intelligence.
23. Learn about Precision,recall and f1-score,available at<<blog.nillsf.com/index.php/2020/05/23/confusion-matrix-accuracy-recall-precision-false-positive-rate-and-f-scores-explained/>> last accessed on 15-08-2022 at 02:30pm
24. Vikas S Chavan and SS Shylaja (2015). Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pages 2354–2358. IEEE.
25. Brendan Kennedy, Drew Kogon, Kris Coombs, Joe Hoover, Christina Park, Gwenyth Portillo-Wightman, Aida Mostafazadeh, Mohammad Atari, and MortezaDehghani (2018).. A typology and coding manual for the study of hate-based rhetoric.
26. ConversationAI (2017). Toxic comment classification challenge: Identify and classify toxic online comments.
27. Learn about LSTM, available at<<https://en.wikipedia.org/wiki/Long_short-term_memory>> last accessed on 15-08-2022 at 02:30pm
28. Learn about CNN,available at<<<https://en.wikipedia.org/wiki/CNN>>> last accessed on 15-08-2022 at 02:30pm.

E. J. [Signature]
12.09.22

our report

ORIGINALITY REPORT

25%
SIMILARITY INDEX

19%
INTERNET SOURCES

10%
PUBLICATIONS

14%
STUDENT PAPERS