

Long Short-Term Memory Deliberation for Bengali Document Classification

BY

Rafid Imtiaz Chowdhury
ID:183-15-11936

Md. Jannatul Naeem
ID:183-15-11927

Provaker Sarker Anik
ID:183-15-12035

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

Supervised By

Abdus Sattar
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University\



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

23 January 2023

APPROVAL

This Project titled "Long Short-Term Memory deliberation for Bengali Document Classification", submitted by Md. Rafid Imtiaz, Md. Jannatul Naeem & Provaker Sarker Anik to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on January 2023.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman




Dr. Md. Zahid Hasan

Associate Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Fahad Faisal

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Ahmed Wasif Reza

Associate Professor

Department of Computer Science and Engineering
East West University

External Examiner

DECLARATION

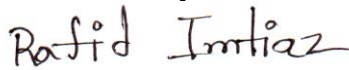
We hereby declare that, this project has been done by us under the supervision of **MR Abdus Sattar, Assistant Professor**, Dept. of CSE, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree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

Supervised by:



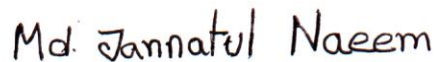
Abdus Sattar
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Submitted by:



Rafid Imtiaz Chowdhury

ID:183-15-11936



Md. Jannatul Naeem

ID:183-15-11927



Provaker Sarker Anik

ID:183-15-12035

Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to Abdus Sattar, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Long Short-Term Memory Deliberation for Bengali Document Classification” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive, criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

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

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

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

ABSTRACT

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

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

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgments	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 Motivation	2
1.3 Research questions	2
1.4 Expected output	2-3
1.5 Report layout	3
CHAPTER 2: BACKGROUND STUDIES	4-6
2.1 Introduction	4
2.2 Related work	4-5
2.3 Research summary	5
2.4 Scope of the problem	6
2.5 Challenges	6
CHAPTER 3: RESEARCH METHODOLOGY	7-12
3.1 Introduction	7
3.2 Research subject and instrumentation	8
3.3 Data collection	8-10
3.4 Dataset Distribution	10
3.5 Preprocessing	10-11

3.6 Stop Word Remove	11
3.7 Tokenization	12
3.8 Statistical Analysis	12
3.9 Implementation requirements	12
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	13-15
4.1 Introduction	13
4.2 Model Performance	14-15
4.3 Summary	15
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY	16-17
5.1 Impact on Society	16
5.2 Impact on Environment	16
5.3 Ethical Aspects	17
5.4 Sustainability	17
CHAPTER 6: CONCLUSION AND FUTURE WORK	18-19
6.1 Conclusion	18
6.2 Recommendations	18-19
6.3 Implication for further study	19
REFERENCES	20-21

LIST OF FIGURES

FIGURES	PAGE
	NO
Figure 3.2: Data statistics Diagram	9
Figure 3.3: Length frequency distribution diagram	10
Figure 3.4: Dataset Distribution based on sentiment class	10
Figure 4.2.1: Confusion matrix	14
Figure 4.2.2: Training and Validation Accuracy and Loss	15

LIST OF ABBREVIATION

NLP Natural Language Processing

ML Machine Learning

LSTM Long Short-Term Memory

CNN Convolutional Neural Network

CHAPTER 1

Introduction

1.1 Introduction

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

1.2 Motivation

We can argue that estimation examination is not an NLP problem. It explains each aspect of NLP, such as coreference aims, invalidation handling, and word sense disambiguation, which adds to the difficulty because these are topics that aren't handled in NLP. It's also useful to understand that experiencing the examination is a very restricted NLP problem, because the framework doesn't need to fully comprehend the semantics of each phrase or archive, simply a few characters of positive or negative suppositions, and their major aspects or themes. In this aspect, opinion polling is a fantastic phase for NLP researchers to achieve significant headway on all fronts of NLP while also having the capacity to arrive at a massive sensible conclusion. In this paper, I examine the dataset utilized in this study, which was gathered from some well-known news portal and use a sequential model algorithm to make predictions. The dataset includes 136812 data, which were pre-processed before being used to train the model. This paper proposes a wide variety of models and strategies.

1.3 Research Questions

- Which algorithm is better for Document Classification?

1.4 Expected Output

In brand or celebrity administration, social media is a wonderful tool for getting customer feedback and exhibiting data, but there hasn't been an acceptable solution for consequently arranging the massive number of tweets, which allows room for study. This idea is based on comparing the presentation of distinct slant grouping approaches and developing a new conclusion

characterization method for dealing with the order of tweets concerning carrier administrations. Three approaches are introduced in this research, including an overall approach that includes CNN and LSTM. The collecting characterization strategy takes the five classifiers' arrangement results into account and uses the greater part vote procedure to determine the final evaluation expectation. This theory includes an assessment and investigation of several idea characterization methodologies. As the AI conclusion layout nears, the next step is model preparation, which comes after component selection. However, other traditional approaches, such as the Lexicon-based methodology, do not incorporate AI operations, therefore no prior preparation is necessary for these methodologies. In this section, I discuss the conclusion grouping tactics that I use in my work. To define the sentiment, ml classification techniques are utilized, with the algorithm predicting sentiments based on the review content. All of them are assigned a number that indicates the language that contains various document class names.

1.5 Report Layout

The report has six chapters. Every chapter describes the different aspects of the "**Bangla Document Classification**". Every chapter has different parts described in detail.

Chapter 1: Introduction

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

Chapter 2: Background Studies

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

Chapter 3: Research Methodology

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

Chapter 4: Design Specification

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

Chapter 5: Implementation

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

Chapter 6: Conclusion and Future Scope

The end is drawn and my commitments are portrayed.

CHAPTER 2

Background Studies

2.1 Introduction

The main objective of emotion research is to divide the task into positive or negative amplitude in order to separate parental attitudes or details. The purpose of this study is to improve customer penetration, revenue, and branding. Techniques, as well as various fields including finance, economy, and some spam detecting stock exchange, purchasing and selling goods, as well as several other businesses. Since they can react quickly and provide people the opportunity to profit from the required behavior or decision-making, effective intuition analyses might have a significant impact in many fields, including policy, governance or organization, campaigns, and enterprises. Cost-effectively may be acquired neural networks. many emails, comments, and assessments totaling thousands. Text classification techniques should be broadened to include all sizes of businesses. There are several critical situations that organizations must be aware of and act upon as promptly and effectively as feasible. To be able to swiftly identify crucial characteristics, computer information retrieval should often and in real-time mimic the designer labeling. The idea of text classification is not new in the field of natural language processing. The Bangla text is still being worked on, nevertheless. Online news is categorized in this industry in a wide variety of ways. In the era of online news sources, people depend on this issue. The suggested study, which is based on the Bangla language, has as its objective this categorization. Certain Bangla datasets make use of some of the study materials that are included in our literature survey section. Compared to other machine learning approaches, our hybrid modeling approach is more successful.

2.2 Related Work

To categorize brief material, Yang Li developed an SVM KNN technique [2]. They used CNN, SVM, NB, RNN, and LSTM machine learning classifiers. Finally, the SVM + CNN (SVMCNN) classifier produced superior results. Using the SVM KNN method, they were able to achieve findings that were around 90% accurate. Another relevant researcher, Tej Bahadur Shahi, made predictions for self-acting Nepalese article multi classification [4]. He also completed her research in order to select a classifier model and artificial neural. Multi-layer connectivity is utilized with

machine learning classifiers including such SVM and Naive Bayes. The neural network, on the other hand, is in a bit of an awkward predicament. Nepali news text categorization was 74.65% in favor of SVM, including RBF, during the process. However, having 73 percent efficiency, the neural net is ranked second in terms. The entire volume of Nepali news text categorization data is 4964, with 20 different sorts of news. All deep learning models, such as neural networks, are hungry for data with a high numerical value. To classify Bangla news headlines, Prasenjit Dhar and Md. Zainal Abedin applied the best machine learning concepts [6]. As machine learning classifiers, they employed SVM, Naive Bayes, and Adaboost. They were able to attain an accuracy rate of roughly 81%. Sheikh Abujar proposed a neural network-based Bengali news multi-classification system with comparable performance [7]. They prepare over 86 thousand news headlines. As machine learning techniques, they employed SVM, NB, Random Forest, Logistic Regression, and Neural Network. They were able to reach an accuracy of around 90% using Neural Network methods. Bjorn Gamback focused on text categorization for hate speech [8]. The convolutional neural network is something he wants to emulate. With the help of CNN, they were able to attain a score of 86.68 percent. They use a different method of word embedding that raises this number by 7.3 percent when using the SoftMax function and max pooling. Even so, the values are raised immediately. Word embedding is required to prepare the data for analysis. According to Roger Alan Stein, word embedding minimizes the system's worst performance [9]. Amin Omidyar employed clickbait web data from the media in their research [10], which they subsequently analyzed with such a machine learning classifier and a neural network.

2.3 Research Summary

A collection of data was gathered from a number of well-known social media networks. The dataset included a diverse group of remarks, which were subsequently analyzed. The dataset is then pre-processed in preparation for model adjustment. All extraneous punctuations were deleted from the dataset during pre-processing. My data is separated into five classes, however when I was labeling the dataset, I saw that the server document class was simply a sub-class of the hazardous class. It wasn't even close to being mentioned as a labeling problem in this situation. All extraneous punctuation marks were eliminated from the dataset after pre-processing in order for the model to better grasp the raw data producing more accurate recognition.

2.4 Scope of the problem

This study discusses and gives findings from a hypothetical assessment of Twitter posts linked to US airline companies. The goal of this investigation is to determine whether tweets may be classified as good, negative, or neutral. Clients can converse and exchange their facts, opinions, and suppositions through informal communication destinations in the online world. Carrier tweets are becoming increasingly popular and are being used as a dataset to assess client concerns. To construct a model, the researchers used classifiers KNN, Logistic Regression, and Random Forest to categorize extreme through explanation. The experts compiled tweets from adjacent planes about their experiences with the services provided by local airlines in the Philippines. The analysts also determined the value of a tweet based on whether it is positive, nonpartisan, or negative, and provided quantitative and subjective investigations, as well as conclusion examination, to better understand 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

We have considered the slant assessment 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 subset 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 events in the classes. Managing Bengali text and processing it for model training was also a difficult challenge. As can be other classes. When applied to all datasets, the neural network model, for example, has exhibited a high degree of anticipation and security. While LSTM has demonstrated a commendable level of execution across all assessment metrics.

CHAPTER 3

Research Methodology

3.1 Introduction

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

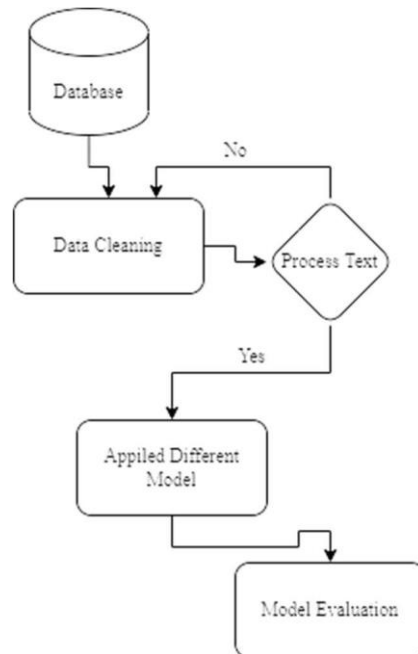


Figure 3.1 Model Architecture

3.2 Research Subject and Instrumentation

The name of the issue that I suggest is " Bengali Document Classification ". In the field of Natural Language Processing, this is a prominent study area. To this point, I have analyzed the way toward doing an estimation inquiry in Bangla using a specified and theoretical approach. A notable learning model necessitates a computer with a high structure and a variety of instruments. Under the fundamental instrument for this model, there is an example of a concept analyzer.

Hardware and Software:

- 8GB RAM and Intel core i5 5th generation.
- 2 TB Hard Disk.

Tools:

- Windows 10
- Python 3.7
- Sklearn
- NumPy
- Pandas
- Seabearn
- NLTK
- Jupiter Notebook

3.3 Data Collection

Scraping was used to acquire data from numerous Bangla newspapers. Our collection contains over a million records. We gathered information from publications such as Bangladesh pratidin [17], dainik jugantor [18], daily inqilab [19], kalerkantho, and others. These are Bangladesh's most widely read newspapers. We gathered data from these newspapers, which aids in determining which kinds of data are most frequently accessed by readers. For collecting data from websites,

we utilized Chrome Web Scrapper and Python tools. In our dataset, there are three columns. These are the headlines, the category, and the name of the newspaper. The data is open to the public. 1 The following graph depicts the headline dispersion of each category. This set of data is unbalanced.

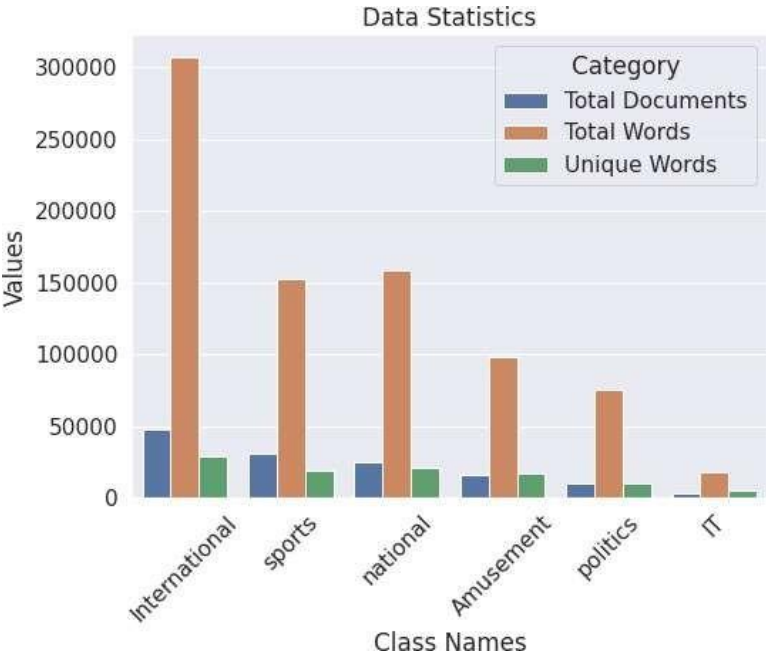


Figure 3.2 Data statistics Diagram

The headlines' maximum, minimum, and average lengths were then defined, and a data graphic was developed to represent those variables. The maximum length of a headline was 118 characters, the shortest length of a headline was 3 characters, and the average length of a headline was 21 characters. Following that, a context is provided for the headline length distribution diagram.

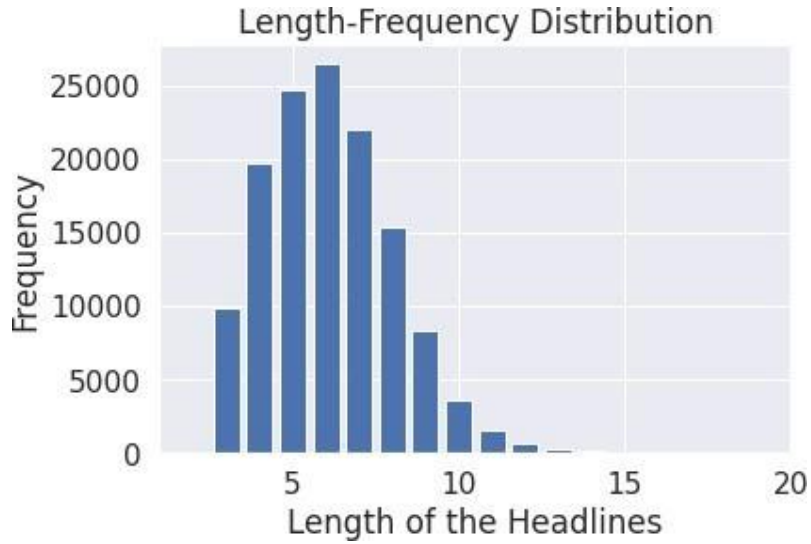


Figure 3.3 Length frequency distribution diagram

3.4 Dataset Distribution

After all of the cleaning and clearing 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, using an 80/20 split. The train data size was 11041 bytes, whereas the test data size was 276bytes.

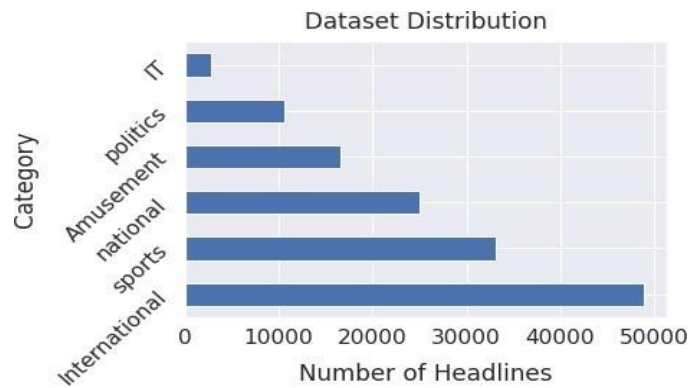


Figure 3.4 Dataset Distribution based on class

3.5 Preprocessing

Because the crude information contains clamor and copies, or the initial information is not appropriate for investigation procedures, information pre-handling is incredibly important in information examination. This is becoming increasingly important in content characterization

since content information is not the same as numeric information, implying that we need convert content information into a breakable configuration. If the characterization techniques are regulated learning arrangement approaches and model assessments are involved, this also requires information to be indicated. Aside from that, real content material frequently contains several grammatical errors, contractions, and pictures, resulting in incorrect arrangement outcomes. Individuals put "THX," "Much appreciated," or "Thanks" to show genuine thanks in informal community postings, where the first is a contraction and the last is a mistake. Even though it is obvious to most humans that these terms represent the same thing, AI calculations have significant difficulties in making sense of them.

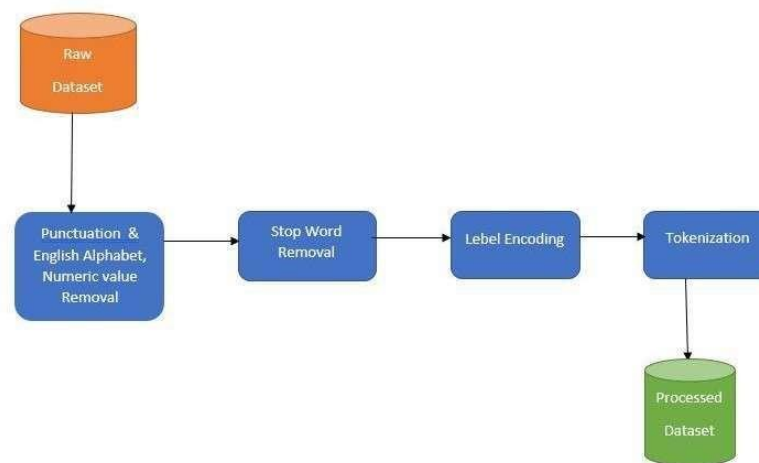


Figure 3.5 Data Preprocessing

3.6 Stop word remove

A stop word is a commonly used term, such as ".", ",", "'", and so on, that a web index has been configured to ignore, both while sorting parts for viewing and when recovering them as the result of a pursuit request. I wouldn't need these terms to take up space in our database or take up a lot of processing time. For this, I may effectively evacuate them by storing a list of terms that we regard to be stop words. In Python, the NLTK (Natural Language Toolkit) provides a list of stop words stored in 16 different dialects. They may be found in the NLTK data index.

3.7 Tokenization

While in the discipline of NLP, tokenization mostly refers to the act of dividing and maybe distinguishing a dataset for classification. The procedure for the train was completed in this study, and test cases for the model were developed.

3.8 Statistical Analysis

1. A total of 136812 data points is included in the dataset.
2. The data is separated into five categories.
3. The train is based on 95552 data.
4. The test is based on 13272 data points.
5. The accuracy rate was 83 percent.

3.9 Implementation Requirements

Python was utilized as a programming language to build the neural network model. The Panda library is used to load the dataset, while the NLTK library is used for preprocessing. For those neural network models, I also utilized the Kera's library. The entire system is developed on a Jupiter notebook environment.

CHAPTER 4

Experimental Results and Discussion

4.1 Introduction

Document Categorizations remarks are becoming a growing source of concern for many individuals throughout the world. I used a technique to remove poisonous remarks from a dataset gathered from different news portal. We applied LSTM deep learning model and gained good accuracy.

Based on the network's recurrent architecture, recurrent neural networks (RNNs) have been developed to handle time sequence data. Nevertheless, if the distance between both the unit containing the necessary data and the unit where RNNs struggle to learn how to link the information when the amount of information needed increases because of the gradient bursting or disappearing problem. So, by adding three gates (input gate, forget Long-Short Term Memory (input and output gates) (LSTM) Networks are set up to enhance the initial RNNs. RNN and LSTM have recently demonstrated their achievements in time sequence data processing fields like action, speech recognition, and language translation Image captioning and recognition.

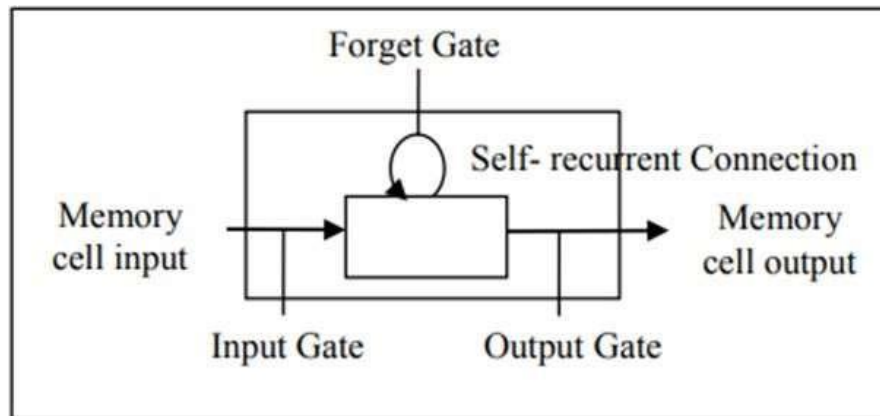


Figure 4.1 LSTM single cell structure

4.2 Model Performance

As we added the dataset to our convolutional neural network model, we can observe the training and validation accuracy, as well as the training and validation loss, on the provided accuracy assessment graph.

Rather than that, we were able to obtain an overall accuracy of 83 percent. In the case of the Bengali text for NLP assessment, we had to deal with a slew of problems due to the Bengali text's abundance of difficult terminology. In order to comprehend such material, which and why other classifiers were a complexity. The following is a diagram of the confusion matrix.

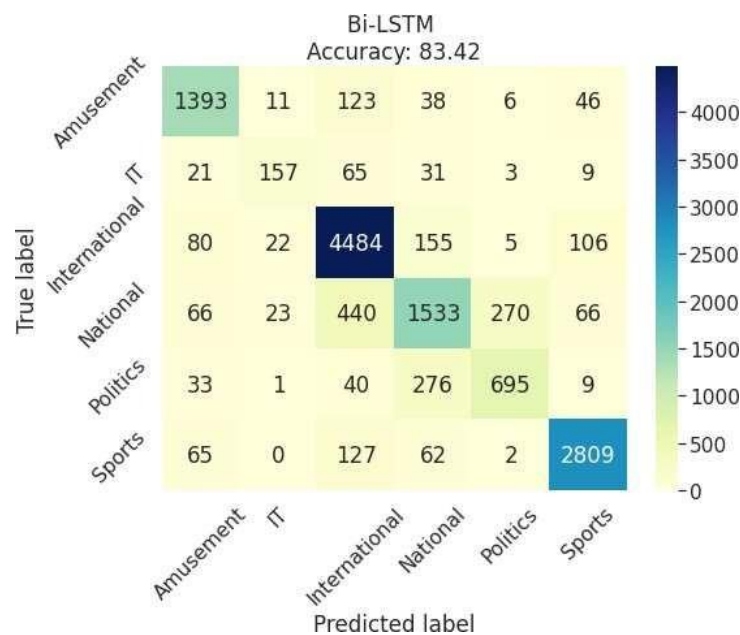


Figure 4.2.1 confusion matrix

The confusion matrix, on the other hand, is in charge of calculating the overall summary of findings based on classification process predictions. The count values, which are split down by distinct classes, may be determined.

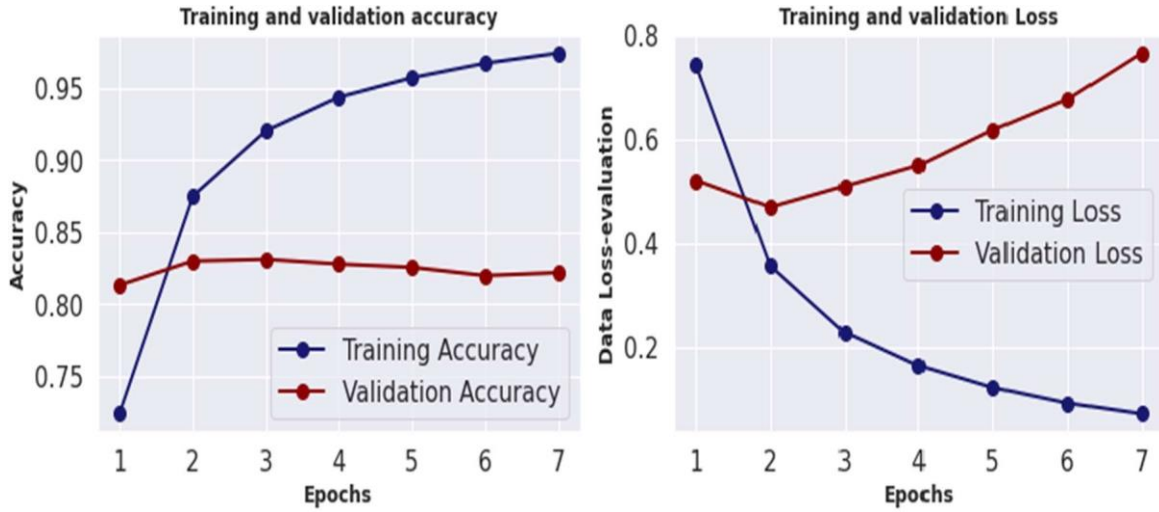


Figure 4.2.2 Training and Validation Accuracy and Loss

Table-1: Model accuracy

Model	Accuracy
LSTM	84%

4.3 Summary

There are a number of additions in the comments, including emoji, suggested data, and the Bengali language, which is more sophisticated than English due to its differing punctuation and symbols. As a result, I'll have to work harder to concentrate on those specific issues.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

More and more individuals are becoming involved in the world of the internet these days. The globe is rapidly becoming into a global village. Humans may communicate and share their lifestyles and other information over the internet in these modern times. That is the primary reason why the number of internet users is rapidly expanding. The advantages and disadvantages of using the internet are both there. Document classification remarks are a major concern on the negative side. It has also become one of the world's most hotly debated topics. To address this issue, I made the initiative to remove it from the Bengali language. This evaluation will continue to have a beneficial influence on society.

5.2 Impact on Environment

It's the widespread dissemination of false information through social media that creates harm. If you're an individual or corporation that provides a lot of content, perhaps through the use of web-based life programming, you need be particularly cautious. It only stops for a second or two to check whatever you see on social media sites. Consider the story's starting point. Whether you've never heard of it, look it up on the internet to see if it's credible. If you don't have that much time, it's best to ignore it, especially if it appears to be satire, false information, or intentional publicity. By not spreading skewed information, you may help to reduce the spread of deceit and fake news.

One of the key advantages of a web-based existence, as mentioned above, is the ability to transmit info to a large number of people in a very short period of time. Even though this might be perceived as a huge benefit in a crisis, it can also be a huge hindrance because info with no validity can be transmitted in a split second. This can lead to a lot of deception and hysteria. This was demonstrated when rumors that the Queen had died circulated because to the Queen missing a Christmas administration due to a common sickness. This, combined with several hoaxers fabricating fake news items, led many to believe the Queen had died.

5.3 Ethical Aspects

Online networking allows you to find meetings that are focused on your advantages and pastimes, making it one of the best ways to meet and interact with new people who have similar interests as you. This is fantastic for meeting new people, as well as for finding love interests and web dating, which has become more popular than a traditional face-to-face meeting as a result of online life and any resemblance to Tinder.

The internet is a fantastic way to convey news quickly throughout the world, with "breaking news" tweets receiving thousands of retweets in minutes. This may be quite useful for updating people on important information, such as climate updates and missing children.

As previously said, internet networking has transformed society in a variety of good ways, yet at no additional cost, as all key web-based life phases are free. Consider another object or service that has ever impacted your life as much as the internet, and then think about how much it costs.

5.4 Sustainability

- 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.
- 65% of all people when they can't access their online life accounts, they feel uneasy and uncomfortable.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

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

6.2 Recommendations

Experimental analysis is the most current trend in understanding the needs of the general public; it's a more straightforward and clever way to see how people feel about a particular issue and the brand influence of smaller-scale blogging. In this circumstance, we considered how individuals felt about the aircraft industry and how we handled United Airlines' recurring challenges and how the general public felt about it. The investigation confirmed our suspicions about how effective a method for dealing with twitter assumption investigation is. The Logistic Regression and Random Forest classifiers used in the calculation, as well as two programming for better results, clearly depict the mass group's assumption, and thus the aircraft could easily decipher the

data and profit from it by attempting to improve the angles that appear to be negative or disliked by the directed crowd. There are various suggestions for this project, like,

- Remove bias from dataset.
- Need the ratio of data categories is equally distributed.
- Use more machine learning classification algorithms.
- Create neural network for better result.
- Need parameter tuning for classification.

6.3 Implication for Further Study

Because of the rapid growth of information 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 literature, Bengali comment inquiry is typically based on little social media data, with only a few days' worth of data. Unless social media material is routinely obtained, this obstacle prevents the acquisition of factually significant and essential consequences. Broad research of tweets to produce a factually massive client assessment must take into account a few factors. These include, at the very least, 1) a sufficiently long period of time during which tweets are gathered to ensure representativeness as opposed to individuals' immediate response following a piece of news about the event, 2) an adequate number of tweets that best speak to each geographic area, 3) a gauge of possible inclination if the tweets originate from a specific geographic area, and 4) if the conclusions drawn match the prevalent attitude gleaned from other market sources.

REFERENCES

- [1] Meparlad, Understanding Text Classification in NLP with Movie Review, [https://www.analyticsvidhya.com/blog/2020/12/understanding-text-classification-in-nlp-with-movie-review-example-example\(2020\).](https://www.analyticsvidhya.com/blog/2020/12/understanding-text-classification-in-nlp-with-movie-review-example-example(2020).)
- [2] Shahin, M. M. H., Ahmmed, T., Piyal, S. H., & Shopon, M. (2020, June). Classification of bangla news articles using bidirectional long short term memory. In *2020 IEEE Region 10 Symposium (TENSYMP)* (pp. 1547-1551). IEEE.
- [3] Yang, Y., & Joachims, T. (2008). Text categorization. *Scholarpedia*, <https://ieeexplore.ieee.org/abstract/document/9230737> 3(5), 4242.
- [4] Hu, Y., Li, Y., Yang, T., & Pan, Q. (2018, November). Short text classification with a convolutional neural networks based method. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)* (pp. 1432-1435). IEEE, <https://ieeexplore.ieee.org/abstract/document/8581332>
- [5] Stein, R. A., Jaques, P. A., & Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. *Information Sciences*, 471, 216 <https://www.sciencedirect.com/science/article/abs/pii/S002002551830693/>
- [6] Omidvar, A., Jiang, H., & An, A. (2018, September). Using neural network for identifying clickbaits in online news media. In *Annual International Symposium on Information Management and Big Data* (pp. 220-232). Springer, Cham.
- [7] Cai, J., Li, J., Li, W., & Wang, J. (2018, December). Deep learning model used in text classification. In *2018 15th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)* (pp. 123-126). IEEE. <https://ieeexplore.ieee.org/document/8632592>
- [8] Shahi, T. B., & Pant, A. K. (2018, February). Nepali news classification using naïve bayes, support vector machines and neural networks. In *2018 International Conference on Communication Information and Computing Technology (ICCICT)*(pp.15).IEEE.https://www.researchgate.net/publication/324098346_Nepali_news_classification_using_Naive_Bayes_Support_Vector_Machines_and_Neural_Network,
- [9] Dhar, P., & Abedin, M. (2021). Bengali News Headline Categorization Using Optimized Machine Learning Pipeline. *International Journal of Information Engineering & Electronic Business*, 13(1). https://scholar.google.com/scholar?q=Bengali+News+Headline+Categorization+Using+Optimized+Machine+Learning+Pipeline&hl=en&as_sdt=0&as_vis=1&oi=scholart.
- [10] Khushbu, S. A., Masum, A. K. M., Abujar, S., & Hossain, S. A. (2020, July). Neural network based bengali news headline multi classification system: Selection of features describes comparative performance. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&as_vis=1&q=Neural+network+based+bengali+news+headline+multi+classification+system%3A+Selection+of+features+describes+comparative+performance&btnG=
- [11] Al-Tahrawi, M. M. (2015). Arabic text categorization using logistic regression. *International Journal of Intelligent Systems and Applications*,
- [12] 7(6), 71. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&as_vis=1&q=Arabic+text+categorization+using+logistic+regression.+International+Journal+of+Intelligent+Systems+and+Applications&btnG=
- [13] Zia, T., Abbas, Q., & Akhtar, M. P. (2015). Evaluation of Feature Selection Approaches for Urdu Text Categorization. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&as_vis=1&q=Evaluation+of+Feature+Se

lection+Approaches+for+Urdu+Text+Categorization.&btnG=

[14] Gambäck, B., & Sikdar, U. K. (2017, August). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online* (pp. 85-90). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&as_vis=1&q=Using+convolutional+neural+networks+to+classify+hate-speech&btnG=

[15] Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*(pp.45804584).IEEE.https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&as_vis=1&q=Convolutional%2C+long+short-term+memory%2C+fully+connected+deep+neural+networks&btnG=

[16] Kostadinov, S. (2017). Understanding GRU networks. Towards Data Science. *Towards Data Science, Towards Data Science, 16*.

[17] Bangladesh protidin, <https://www.bd-protidin.com> (2021).

[18] Doinik Jugantor, <https://www.jugantor.com> (2021).

[19] Daily Inqilab, <https://www.dailyinqilab.com> (2021).



Long Short-Term Memory deliberation

ORIGINALITY REPORT

4%	2%	1%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Shinas College of Technology Student Paper	1%
2	Submitted to Daffodil International University Student Paper	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	Submitted to Coventry University Student Paper	1%
5	Submitted to TechKnowledge Student Paper	<1%
6	kth.diva-portal.org Internet Source	<1%

Exclude quotes Off
Exclude bibliography Off

Exclude matches Off