

Cyberbullying Detection from Romanized Bangla Text using Deep Learning Techniques

BY

MD. MOJNU MIA
ID: 221-25-097

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

Supervised By

Ms. Naznin Sultana
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Fizar Ahmed
Associate Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project/Thesis titled “Cyberbullying Detection from Romanized Bangla Text using Deep Learning Technique”, submitted by Md. Mojnu Mia, ID No: 221-25-097 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 M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17-01-2023.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan, PhD

Professor and Head

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

Chairman

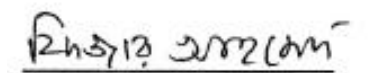


Ms. Nazmun Nessa Moon

Associate Professor

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

Internal Examiner

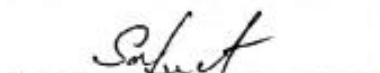


Dr. Fizar Ahmed

Associate Professor

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

Internal Examiner



Md. Safaet Hossain

Associate Professor & Head

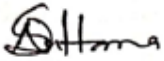
Department of Computer Science and Engineering
City University

External Examiner

DECLARATION

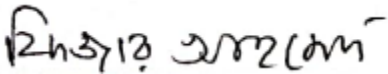
I hereby declare that this thesis has been done by us under the supervision of **Ms. Naznin Sultana, Associate Professor, and Department of CSE** Daffodil International University. I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:



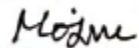
Ms. Naznin Sultana
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By:



Dr. Fizar Ahmed
Associate Professor
Department of CSE
Daffodil International University

Submitted by:



Md. Mojnu Mia
ID: - 221-25-097
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

I really grateful and wish our profound our indebtedness to **Ms. Naznin Sultana, Associate Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Machine Learning and Deep Learning” to carry out this thesis. 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 thesis.

I 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 thesis and to other faculty member and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

People can freely express their ideas and opinions on a wide range of topics directly through a variety of social media outlets thanks to advancements in the internet and social media. In Bangladesh, people use both English and Romanized Bangla to write comments on social media sites like Facebook and Instagram. These remarks can be both positive and negative. Writing in our mother tongue simply stands out more than writing in other languages. Researchers have put a lot of effort into detecting cyberbullying in English and Bangla among other languages. In this thesis, I present my work on identifying offensive language from comments taken from celebrities on various social sites. What usually happens on social sites is that users post comments, some of which are good and some of which are offensive. The social site user follows all comments except those filtered by the poster's profile. I created a data set of about 1200 comments, including good comments and offensive comments. Where I have leveled the good and offensive comments. I have leveled good comments as 0 and offensive comments as 1. When celebrities post on social media my algorithm will check with the data set and users will try to share their comments and show only positive comments. In real-time, all abusive comments will come to spam and social media harassment will decrease.

TABLE OF CONTENTS

| CONTENTS | PAGE |
|------------------------------------|-------------|
| Approval | i |
| Board of examiners | i |
| Declaration | ii |
| Acknowledgements | iii |
| Abstract | iv |
| Table of Contents | v |
| List of Tables | viii |
| List of Figures | ix |
| | |
| CHAPTER | |
| CHAPTER 1: Introduction | 1-4 |
| 1.1 Introduction | 1 |
| 1.2 Motivation | 3 |
| 1.3 Objectives | 3 |
| 1.4 Expected Output | 3 |
| 1.5 Report Layout | 4 |
| | |
| CHAPTER 2: Background Study | 6-13 |

| | |
|---|--------------|
| 2.1 Preliminaries | 6 |
| 2.1.1 Natural Language Processing | 6 |
| 2.1.2 Tokenization | 7 |
| 2.1.2.1 Word Tokenization | 8 |
| 2.1.2.2 Sentence Tokenization | 8 |
| 2.1.3 Natural Language Toolkit (NLTK) | 9 |
| 2.2 Related Works | 9 |
| 2.3 Comparative Studies | 12 |
| 2.4 Scope of the Problem | 13 |
| 2.5 Challenges | 13 |
| CHAPTER 3: Proposed Methodology | 14-16 |
| 3.1 Data Collection Procedure | 14 |
| 3.2 Preprocessing | 14 |
| 3.3 Proposed Methodology | 15 |
| 3.4 Representation Test Raw Comments | 16 |
| 3.5 Implementation Requirements | 16 |
| CHAPTER 4: Experimental Results and Discussion | 18-30 |
| 4.1 Experimental Result and Analysis | 18 |
| 4.1.1 Sentence Tokenization | 29 |
| 4.1.2 Test Results and Report | 29 |
| 4.2 Discussion | 30 |

| | |
|--|--------------|
| CHAPTER 5: Impact on Society, Environment, Sustainability | 33-35 |
| 5.1 Impact on Society, Environment and Psychological state | 33 |
| 5.2 Ethical Aspects | 35 |
| CHAPTER 6: Summary, Conclusions and Implication for Future Research | 36-37 |
| 6.1 Summary of the Study | 36 |
| 6.2 Conclusions | 36 |
| 6.3 Implication for Further Study | 37 |
| REFERENCES | 38-39 |

LIST OF TABLES

| TABLES | PAGE |
|---|-------------|
| Table 4.1: Confusion Matrix for MNB | 19 |
| Table 4.2: Classification Report for MNB | 19 |
| Table 4.3: Confusion Matrix for RF | 20 |
| Table 4.4: Classification Report for RF | 21 |
| Table 4.5: Confusion Matrix for DT | 22 |
| Table 4.6: Classification Report for DT | 22 |
| Table 4.7: Confusion Matrix for SVM | 23 |
| Table 4.8: Classification Report for SVM | 24 |
| Table 4.9: Confusion Matrix for KNN | 25 |
| Table 4.10: Classification Report for KNN | 25 |
| Table 4.11: Confusion Matrix for LR | 26 |
| Table 4.12: Classification Report for LR | 27 |
| Table 4.13: Performance Metrics of Classifier | 30 |

LIST OF FIGURES

| FIGURES | PAGE |
|--|-------------|
| Figure 2.1: Tokenization | 7 |
| Figure 3.1: Proposed Methodology | 15 |
| Figure 3.2: Representation Test Raw Comments | 16 |
| Figure 4.1: Visualization Between Actual and Predicted Value of MNB | 20 |
| Figure 4.2: Visualization Between Actual and Predicted Value of RF | 21 |
| Figure 4.3: Visualization Between Actual and Predicted Value of DT | 23 |
| Figure 4.4: Visualization Between Actual and Predicted Value of SVM | 24 |
| Figure 4.5: Visualization Between Actual and Predicted Value of KNN | 26 |
| Figure 4.6: Visualization Between Actual and Predicted Value of LR | 27 |
| Figure 4.7: Visualization Between Actual and Predicted Value of LSTM | 28 |
| Figure 4.8: Confusion Matrix of RNN | 28 |
| Figure 4.9: Sentence Tokenization | 29 |

CHAPTER 1

Introduction

1.1 Introduction

Social media is a common platform for dialogue, sharing of content, expressing opinions, and promoting concepts and goods. Nowadays, one of the most common digital pastimes is spending time on social networking sites like Facebook, Instagram, Twitter, etc. However, social media's application for expanding reach is truly fantastic. Particularly in Asian nations like Bangladesh, where social media is used by the general public more than anything else. Since everyone has the right to free speech, people are free to write and say whatever they want without worrying about what is popular. Due to the lack of supervision, using harsh language has become increasingly popular on social media. It gets serious, especially when people can write and speak in their own tongue. In Bangladesh, people use both English and Romanized Bangla to write comments on social media sites like Facebook and Instagram, both for positive and negative things. Stopping cyberbullying on social media sites has grown more difficult. The detection of texts including cyberbullying and the subsequent blocking of their appearance is one potential solution. With 234 million native speakers worldwide, Bangla is ranked fifth among the top 10 most spoken first languages in 2022 [1]. As indicated by the Bangladesh Telecommunication Regulatory Commission (BTRC) information, there were 113.73 million mobile internet subscribers in Bangladesh at the end of December, 2021 [2]. In our country, there were 49.55 million online media users in January 2022, and between 2021 and 2022, that number increased by 4.6 million (+10.1%) [3]. At least 32% of children between the ages of 10 and 17 in Bangladesh, are vulnerable to online violence, cyberbullying, and digital harassment, said a latest study of United Nations International Children's Emergency Fund (UNICEF), in Dhaka on Tuesday, February 5, 2019 [4]. Computer languages like C++ and C are not considered to be natural languages, unlike human languages. For instance, natural languages include English, Chinese, French, Bangla, and Romanized Bangla. Making a machine understand natural languages is possibly the most difficult problem in computer

science. One area of computer science where we research intercommunication between machines and people is natural language processing. The amount of research being done in this area is growing daily. Nearly every industry, including customer service, the automotive and healthcare industries, benefits from NLP. Natural language processing uses a variety of methods, including sentiment analysis, named entity identification, text classification, and text summarization. The goal of text classification is to group sentences or paragraphs belonging to different categories together into a single text genre. The ability to read, decode, grasp, and make sense of human dialects is a clear objective of NLP. The process of lexical analysis entails recognizing and examining word structures. A language's lexicon refers to its collection of words and expressions. The complete collection of texts is broken down into paragraphs, phrases, and words by the lexical analysis. In syntactic analysis, or parsing, the words in the phrase are examined for grammar and arranged to demonstrate how they relate to one another. Cyberbullying may be effectively detected using machine learning-based categorization algorithms. Machine learning models' effectiveness in prediction and detection has been demonstrated over time. Machine learning-based prediction and detection have been the subject of a significant amount of study. Training data, also known as inputs and predictions, are used to build data models in supervised machine learning. These labeled data are used by algorithms like Multinomial Naive Bayes, Decision Trees, Support Vector Machine, Logistic Regression, etc. to categorize based on some properties of the labeled data.

This study tries to filter offensive remarks on social media by identifying their nature and displaying supportive ones. This essay is absolutely reliable in order to prevent environmental disruption and emotional pain. But because each region of Bangladesh has its own distinctive dialect, I am currently working on the Romanized Bangla language's standard format and intend to eventually include the regional tongues as well. I have completed my work using a data collection of 1200 texts that have been romanized in Bangla. However, I plan to expand the size of the data set in the future. I want to learn about various syntactic levels, characters, tokenization of words and languages according to regions, and use of characters in sentences. In theory, better recollection may result from

this processing.

1.2 Motivation

People are now able to directly share their ideas and opinions on a variety of topics through numerous social media platforms, thanks to the development of the internet and social media. Particularly in Asian countries like Bangladesh, where social media use is at an all-time high among the populace. Since everyone has the freedom to free speech, anyone can write or say whatever they want without having to worry about it being accepted. People utilize these social media sites, such as Facebook, Twitter, Instagram, etc., to exchange information, thoughts, and sentiments. As the usage of social networking grows every day, so does the use of these platforms for extortion, cyberbullying, and torture. The main causes of these circumstances are the dissemination of insulting comments and messages, as well as bad image sharing. Now that people are allowed to express themselves freely, they write and say whatever they want. Our daily lives are hampered by various spam communications that are occasionally not filtered. Now, occasionally, some people post abusive remarks, and it can be quite challenging to filter and read the post comments, making the poster uncomfortable as they read abusive comments. Therefore, doing this will save a lot of time spent looking for positive remarks and reduce online harassment from social media.

1.3 Objective

Cyberbullying Detection from Romanized Bangla Text is the title of my envisioned research work. I designed the suggested approach so that it can recognize any abusive messages in conversations. The goal of this project is to benefit the public by preventing them from wasting their time and energy on harmful, troublesome, and unimportant things, people, and messages. I want to organize all the data into tokenization using the Natural Language Toolkit, and then I want to address this issue.

1.4 Expected Outcome

People may easily use internet media platforms like Facebook, Twitter, Instagram, and

others. On social media, they may readily exchange facts, opinions, and feelings. Since it can be difficult to filter comments on social media and read every message, it happens that some individuals will comment on abusive comments, which can be quite upsetting for the host. The majority of the beneficiaries will be women and girls. They will typically be afraid of being harassed if they express their own opinions on internet media sites. They should express whatever opinions they may have on social media. There will be an improvement in the use of Romanized Bangla in online media platforms. My study therefore seeks to automatically recognize abusive message

1.5 Report Layout

Chapter 1: Introduction

In chapter one we examined objective, motivation, expected outcome in my thesis work, and the report layout.

Chapter 2: Background

Chapter two is background. We talk about the foundation conditions of my task. In this chapter, I also talk about related works. Scope of problems, comparatives studies, and challenges in my thesis work.

Chapter 3: Research Methodology

In this section I discuss data collection procedure, data preprocessing, proposed Methodology, representation test raw comments, and implementation requirements.

Chapter 4: Experimental Result and Discussion

In this chapter, I talk about the result of my study and analysis of the result of what I found by experiment.

Chapter 5: Impact on Society, Environment and Sustainability

In this section, I talk about the Impact on Society, Environment and Psychological state, and ethical aspect.

Chapter 6: Summary, Conclusion, and Implication for Future Research

Chapter six is Summary, Conclusion, and Implication for Future Research. Here I discuss the summary of my thesis work, conclusions, and implication for future study.

CHAPTER 2

Background

2.1 Preliminaries

Social media is a popular forum for discussion, material sharing, opinion sharing, and product and concept promotion. However, social media has some really amazing applications for increasing reach. Particularly in Asian countries like Bangladesh, where social media use is at an all-time high among the populace. Since everyone has the freedom to free speech, anyone can write or say whatever they want without having to worry about it being accepted. Using foul language has risen in popularity on social media as a result of the absence of monitoring. When people are able to write and communicate in their own tongue, the situation becomes more serious. When posting comments on social media platforms like Facebook and Instagram, people in Bangladesh utilize both English and Romanized Bangla, both for positive and negative things. It has become more challenging to stop cyberbullying on social media platforms. Spam communications of many kinds, sometimes not screened, interfere with our daily lives. Now, periodically, some people will write offensive statements, and it might be difficult to filter and read them. This causes discomfort for the poster when they read the offensive comments. As a result, adopting this will cut down on online trolling from social media and save a lot of time wasted hunting for compliments.

2.1.1 Natural Language Processing

Contrary to popular belief, computers cannot speak in dialects. They communicate using machine language or machine code, whereas we use human languages like Bangla, English, Hindi, or French. A subject of semantics, software engineering, and computerized reasoning called "common language handling" is concerned with how computers and human language may work together, specifically how to program computers to measure and analyze a lot of regular language data. The vast majority of us have no idea what the many zeros and ones that computers transmit mean. Furthermore, unless they are adjusted

to do so, computers cannot understand human language. Natural language processing can help with it. The basic tasks of NLP include tokenization, parsing, lemmatization, language disclosure, marking of syntactic components, and semantic association recognition and confirmation. Generally speaking, NLP divides spoken language into more constrained, regular chunks, tries to appreciate associations within the portions, and looks at how the portions work together to make sense.

I'll use Google Collab for my work. It is an open-source online application that includes information representation, information cleaning and update, and AI.

Tokenization methods, NLTK, and other tools have all been used by me.

2.1.2 Tokenization

Tokenization involves dividing raw materials into little lanes. Tokenization divides the raw text into tokens, which are words and sentences. A sentence is a token in a passage, just as a word is a token in a sentence. These tokens aid in comprehending a particular circumstance or developing a model for natural language processing (NLP). By looking at how the words are arranged, tokenization aids in determining the content's importance.

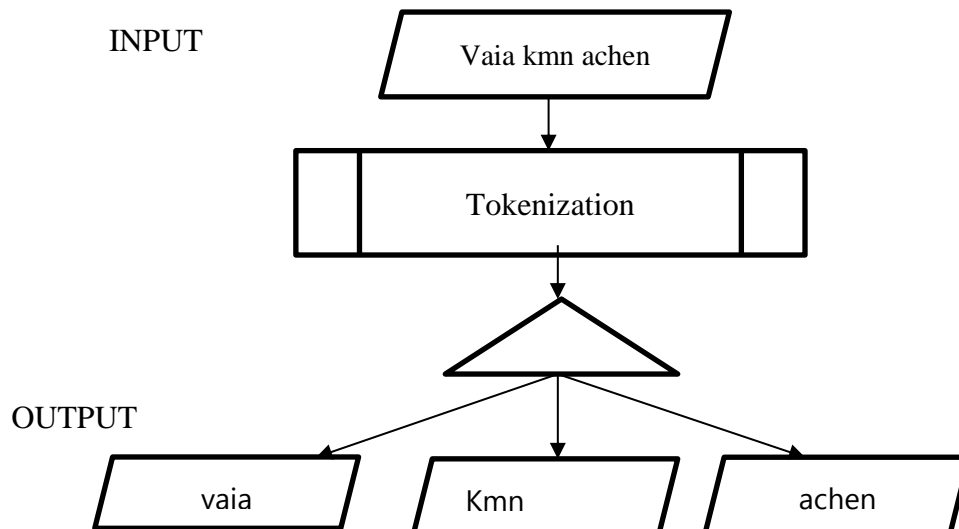


Fig 2.1: Tokenization

2.1.2.1 Word Tokenization

The most used tokenization computation is word tokenization. We use the word tokenize () technique to break a statement up into words. For better content comprehension in AI applications, word tokenization output might be switched to Data Frame. It can also be used as a contribution to other content cleaning processes like accent removal, numeric character removal, or stemming. In order to prepare themselves and generate expectations, AI models require numerical data. Word tokenization becomes a crucial component of the transition of content (strings) to numeric information.

2.1.2.2 Sentence Tokenization

The method for separating text into individual sentences is sentence tokenization. The spacy tokenization calculations perform flawlessly for writing, reporting, and formal archiving, and the tokenizer is built on a corpus of methodical English text. For electronic health records that contain truncations, medical words, spatial estimates, and other structures that are not found in typical written English, the sentence tokenizer performs less effectively. The guilty content, in substitution structure, is saved flawlessly during the tokenization cycle because it won't split a single word. The converse replacements are created once the particular sentences are created, reestablishing distinct content in a number of enhanced sentences.

The reason sentence tokenization is necessary while word tokenization is an option should be a prominent question in our minds. To determine the proportion, we need to tokenize sentences as well as the words. Such a yield serves as a key component for machine preparation because a numeric response is the proper one.

2.1.3 Natural Language Toolkit (NLTK)

The most popular package for managing tasks related to common language processing in Python is the natural language toolkit (NLTK). It is a collection of open-source tools that

were developed in 2001 at the University of Pennsylvania to simplify the creation of Python NLP measurements. Since this bundle's unusual turn of events, it has been expanded thanks to the extensive efforts of open-source users. By providing the necessary tools, NLTK essentially provides the foundation for starting any NLP cycle. Instead of creating each of those devices from scratch, we might combine these core tools to accomplish the final goal.

2.2 Related Works

Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts [5], In this paper they use Natural Language Processing (NLP) technique and various machine learning models and present their model for cyberbullying detection based on comments from YouTube video which are Bangla and Bangla Romanized text. Take 5000 Bangla comment and 7000 Romanized Bangla text from different social media persons. These two datasets are combined to create a third dataset of 12000 texts. After train the machine learning classifier model, it can be seen that Support Vector Machine (SVM) has 76% accuracy in the first dataset and Multinomial Naive Bayes (MNB) has 84% and 80% accuracy for the second and third dataset respectively.

Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla text: A Comparative Study [6], In this paper they use Natural Language Processing (NLP) technique and various machine learning models and present their model for cyberbullying detection based on comments from YouTube video which are Bangla and Bangla Romanized text. Take 5000 Bangla comments and 7000 Romanized Bangla text from different social media persons. These two datasets are combined to create a third dataset of 12000 texts. The preprocessed datasets were trained using the Machine Learning and Deep Learning algorithms. CNN performed best for Bangla dataset and achieved 84% accuracy. Machine learning algorithm Multinomial Naive Bayes (MNB) has 84% and 80% accuracy for the second and third dataset respectively.

An Approach to Detect Abusive Bangla Text [7], They have proposed a root level algorithm for Bengali abusive text recognition in this regard and also proposed a unigram string feature to get better results. They collected data from various social sites. They used Natural Language Processing (NLP) technique and their proposed algorithm they got a good result. In the future they will use machine learning algorithms in combination with this paper's algorithm for more accurate results.

Bangla Hate Speech Detection in Videos Using Machine Learning [8], Their approach is to categorize videos using machine learning classification techniques. Using several sets of keyword searches, they constructed a collection of hate speech videos from YouTube and manually labelled them. They used their dataset, extracted the audio, and transformed it into textual format before applying NLP methods to several machine learning models. The deep learning hyper-parameter tuning method was also used to apply the LSTM and GRU deep learning techniques. The outputs of several models were then compared and contrasted, and they were then examined using various assessment metrics. According to the results of their tests, the deep learning GRU model and logistic regression produce the best accuracy on the available dataset.

BD-SHS: A Benchmark Dataset for Learning to Detect Online Bangla Hate Speech in Different Social Contexts [9], In this study, a manually labeled HS dataset for the Bangla language is presented. It was gathered from comment areas on social media and video streaming websites. The generated dataset consists of 50,281 comments that have been methodically labeled to assist in training machine learning classifiers. They report the benchmark result of their dataset by training several NLP models culminating in the top one achieving an F1-score of 91.0%. In their tests, they discovered that, compared to other pre-trained embeddings, a word embedding trained exclusively utilizing 1.47 million comments from social media and streaming sites consistently produced better modeling of HS detection.

Comparative Analysis on Joint Modeling of Emotion and Abuse Detection in Bangla language [10], This work presents a new approach for both abuse identification and emotion detection based on a comparative analysis in Bangla Language from the findings of other researchers. The examined studies produced accuracy levels of up to 82.2%, 78%, 77.16%, 77.78%, and 78.40%. In order to collect more data using an MTL framework, joint modeling which is covered in this study uses both emotive traits and abusive contents. The

study will be fully implemented in the future, and the paper anticipates that it will be helpful in a variety of ways.

Identifying and Categorizing Opinions Expressed in Bangla Sentences using Deep Learning Technique [11], In their study, Bengali sports news comments from various newspapers were used to train a deep learning model that can classify remarks based on their sentiment. Based on immanent mood, comments are collected and separated. Here, the classification was of a different kind. Authors were more explicit about emotions that were subtly expressed in a sentence. Additionally, a promising model has been developed that can categorize sentences based on their hidden emotional content. Two datasets in the Hindi language are also used to test the model's consistency. It was discovered that the model performed well on those datasets as well.

Sentiment Analysis on Reviews of E-commerce Sites Using Machine Learning Algorithms [12], In this study, sentiment analysis is carried out using machine learning on three datasets (Bangla, English, and Romanized Bangla). They decided to collect data via daraz.com.bd. All of the review data has been classified as either negative, positive, neutral, slightly negative, or slightly positive. They have to oversample on an unbalanced dataset to gain better accuracy. Six algorithms for machine learning have been used. Support Vector Machine (SVM) algorithm, which achieved 94% accuracy for the Bangla dataset, and Random Forest algorithm, which achieved 93% and 94% accuracy for the English and Romanized Bangla datasets, respectively, both performed best.

Bangla Abusive Language Detection using Machine Learning on Radio Message Gateway [13], In this research, they offer a method for identifying abusive language in Bangla coming from a real-time radio message gateway. More than 45000 Bangla sentences that have been classified as abusive and non-abusive have been compiled into a dataset for this study. Their suggested approach involves checking live messages using datasets, separating positive comments, and filtering the offensive remarks. To find the one that

performs the best at predicting abusive language, they have used three different machine learning algorithms. According to their research, the random forest classifiers among the machine learning algorithms they utilized provided the highest accuracy compared to the others.

A Deep Learning Approach to Detect Abusive Bengali Text [14], Several machine learning and deep learning-based algorithms have been tested in this work to identify different types of abusive Bengali text. Additionally, they followed Bangla spelling conventions by using the streaming rule. They achieved the highest level of accuracy in this work using the RNN algorithm of deep learning. To improve their learning in the future, they will employ further deep learning techniques.

Cyberbullying Detection Using Deep Neural Network from Social Media Comments in Bangla Language [15], In this study, it was suggested that a hybrid neural network binary and multiclass classifier model be used to detect Bengali cyberbullying. From the profiles of cricket players, celebrities, and other users on Facebook, they gathered more than 44000 comments. Data was organized into five classes. The dataset employed for this study is current and contains a sizable amount of data, enabling them to obtain results with a high degree of precision. They experimented and discovered that their models performed differently, with the binary classification model providing the highest level of accuracy.

2.3 Comparative Studies

For NLP, RNN is a popular neural network design. Building language models and performing speech recognition tasks have both shown it to be relatively accurate and effective. Sequential data is processed using RNNs. In conventional neural network models, the input layer, the hidden layer, and the output layer are where computation takes place. There is no connection between the nodes of any of these levels, which are all completely connected. In order to make predictions, it maintains data for the present feature as well as nearby features. A RNN keeps track of past data in a memory, allowing the model to forecast the present output based on distant attributes. Despite the fact that an

RNN can learn dependencies, it can only learn about current data. This issue can be resolved with the aid of LSTM because it can comprehend context and recent reliance. As a result, LSTM are a particular type of RNN where being aware of context might be helpful. The main distinction between LSTM networks and RNNs is the replacement of hidden layer updates with memory cells. Because of this, they become more adept at identifying and highlighting long-range relationships in data, which is crucial for sentence constructions.

2.4 Scope of Problem

My job is to identify offensive language on social media. By using this model, I can stop cyberbullying on social media and create a secure and positive environment for women and our people. Additionally, my system can stop violence towards women in various media. I have experience with Bangla that has been romanized. Will not be able to tell if someone spells something incorrectly.

2.5 Challenges

At every stage of handling this study endeavor, we encounter several difficulties. Romanized Bengali text shares structural similarities with English text, but because it uses single words more frequently than Bengali and English, Romanized Bengali is a more sophisticated language. Text written in Romanized Bangla is challenging to read. The primary difficulty in this approach is dealing with datasets. One of the challenges for us is detecting that everyone's writing style is different. We also have to deal with the regional residents' disagreements, which is one of our challenges.

CHAPTER 3

Research Methodology

3.1 Data Collection Procedure

To advance our endeavor, we first required a data set. Bangla data must be gathered in order to detect abusive language in Bangla. Our investigation revealed that popular pages of many classes, including those of VIPs, celebrities, ideological groups, politicians, and athletes, are hotbeds for haters to spew hate speech. Despite our best efforts, we were unable to locate the Bangla data set online. In order to develop it, we gathered information from several radio live shows as well as other sources. These came from various celebrities' social media profiles and from content on radio station social media channels. I saved our entire data set as a CSV file. And that is how we intend to go. We recently came to the conclusion that this endeavor is worthwhile when we were gathering these.

3.2 Preprocessing

This dataset contains both excellent and terrible data. A significant task and advancement in text mining and natural language processing is preprocessing. The crude remarks use unusual characters, like and so on, accents, and feelings. I physically removed these unusual symbolic letters, accents, and Unicode sentiments during the pre-processing hour. Conjunctions are also removed because they are unnecessary for harsh content detection. Following the processing of the dataset and the completion of sheet 1, we began classifying the data by type. Since computers cannot grasp our language, they must be taught to do so. The entire data set has been leveled into portions of 0 and 1. Section 0 was deemed a positive comment, whereas Section 1 was deemed an offensive comment. A deep learning technique was used to train the data set. Now, if a user's comment falls under comment level 0, it will be shown in the interface. However, if a user's comment falls under comment level 1, it will be considered

offensive and will be moved to spam instead of being shown.

3.3 Proposed Methodology

The two types of abusive text classification techniques are paired arrangement and multi-class order. We can just determine in pairs whether a comment is cyberbullying or not.

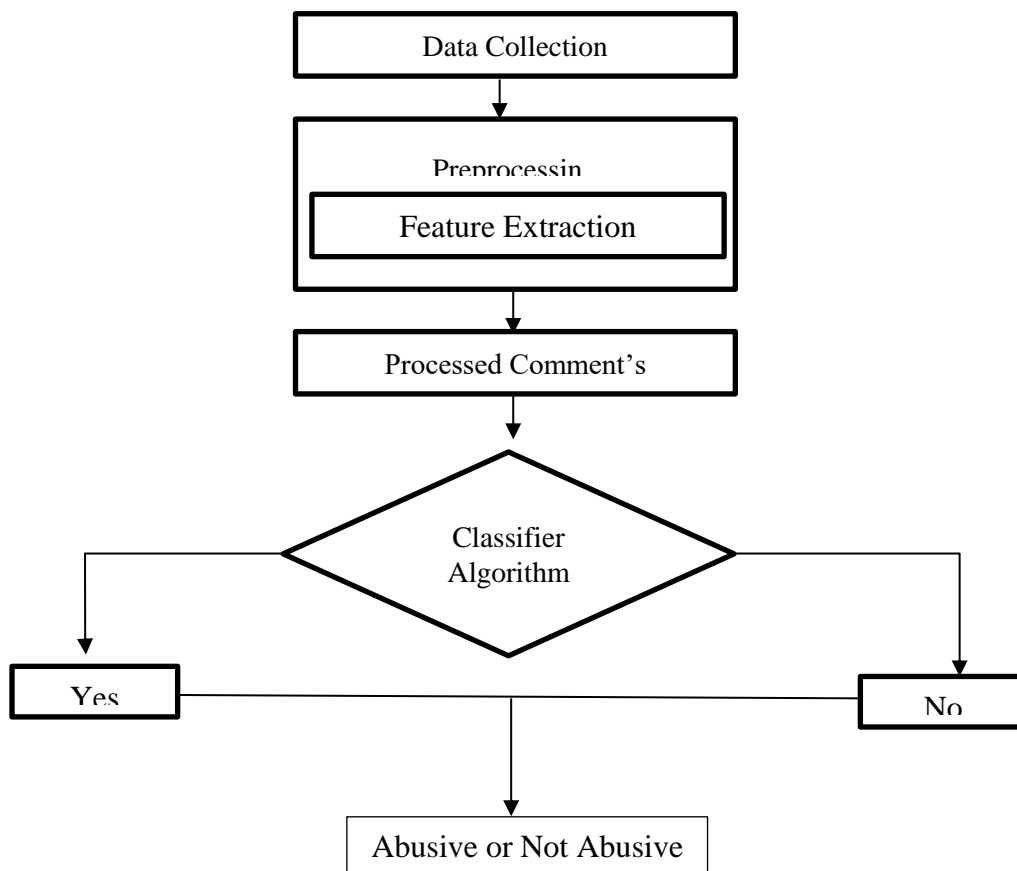


Fig 3.1: Proposed Methodology

3.4 Representation Test Raw Comments

Below is a stage-by-stage explanation of our suggested algorithm using an example comment.

| | Comment | label |
|----|---|-------|
| 0 | sala kamda korlo ki | 1 |
| 1 | Apnara ma meye ta matro janlam.khub valo wish ... | 0 |
| 2 | Ata vua news mehjabin onar meye noi | 0 |
| 3 | Ma tw ma hoibut ak jon adorsho ma ja ka k... | 0 |
| 4 | Khub balo kaj koreco | 0 |
| 5 | Ami to janie na | 0 |
| 6 | Ara 2 jon maa meye ata na jana ami | 0 |
| 7 | Ekhon tomar ammu ke bolo biya daitto ta jani a... | 0 |
| 8 | Tuio kore fel akhon bondhu | 0 |
| 9 | Kire ata ke nuton Manus naki Tor | 0 |
| 10 | Akhon koi chakri koren | 0 |
| 11 | Vai babi naki | 0 |
| 12 | Ronaldho sob club jaiye nijer sera ta diyece..... | 0 |

Figure 3.2: Representation Test Raw Comments

3.5 Implementation Requirements

3.5.1 Local System

- CPU: Intel® Core™ i5
- RAM: 4 GB
- Clock Speed: 2600 MHz
- Video Memory: 2 GB
- L3 Cache: 2 MB
- Operating System: Windows 10 Pro 20H2

3.5.2 Implementation Requirements

- Language: Python (3.8+)
- IDE: Google Collab

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Result and Analysis

To determine the accuracy of my work, I applied Deep Learning and Machine Learning algorithms to the pre-processed data. I used several machine learning algorithms like Multi Nominal Naive Bayes, Decision tree, Random forest, Support vector Machine, K-Nearest Neighbors, Logistic Regression out of which I found that Decision tree, Random forest, Support vector Machine, K-Nearest Neighbors, Logistic Regression gave us 64.41% accuracy. Figure 8 shows that, compared to other algorithms, the Multi Nominal Naive Bayes algorithm produces slightly superior results. This algorithm's accuracy rate is 64.97%, which is 0.56% greater than that of the other methods. I used the Long Short-Term Memory (LSTM) approach from Figure 9 to apply a deep learning model. This algorithm has an accuracy rate of 67.90%, which is higher than the accuracy rate of the machine learning algorithms mentioned before. From Figure 2 we can see that using another deep learning algorithm recurrent neural network (RNN) model I got 69.33% accuracy which is better than my other deep learning algorithm Long Short-Term Memory (LSTM) model with 1.43% accuracy.

Multinomial Naive Bayes Algorithm:

This graph illustrates various scores for recall, precision, accuracy, F1 score, Support, Macro Average and Weighted Average. It demonstrates how the multinomial algorithm compares each score.

Here, Test data was 15% and train data was 85% among 1175 data. Then determined confusion matrix.

TABLE 4.1: Confusion Matrix for MNB

| | True | False |
|-----------------|-------------|--------------|
| Positive | 114 | 0 |
| Negative | 62 | 1 |

In this case, TP = 114, FP = 0, TN = 62 and FN = 1

Now, the classification report

TABLE 4.2: Classification Report for MNB

| | Precision | Recall | F1-score | Support |
|---------------|-----------|--------|----------|---------|
| 0 | 0.65 | 1.00 | 0.79 | 114 |
| 1 | 1.00 | 0.02 | 0.03 | 63 |
| Accuracy | 0.65 | | | 177 |
| Macro Avg. | 0.82 | 0.51 | 0.41 | 177 |
| Weighted Avg. | 0.77 | 0.65 | 0.52 | 177 |

Now, Visualization between actual value and predicted value.

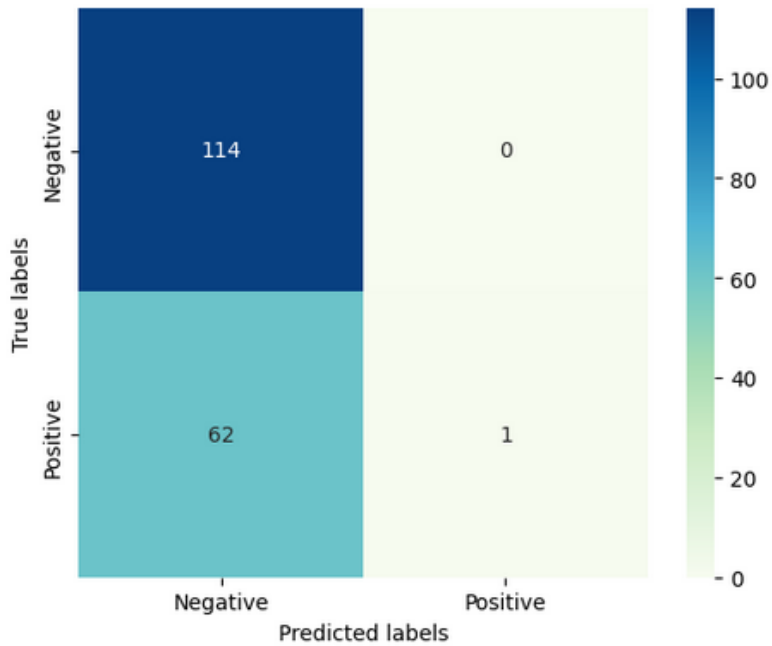


Fig 4.1: Visualization between actual and predicted value of MNB

Random forest classifier:

This graph illustrates various scores for recall, precision, accuracy, F1 score, and several code portions. It demonstrates how the multinomial algorithm compares each score.

TABLE 4.3: Confusion Matrix for RF

| | True | False |
|----------|------|-------|
| Positive | 114 | 0 |
| Negative | 63 | 0 |

In this case, TP = 114, FP = 0, TN = 63 and FN = 0

Now, the classification report

TABLE 4.4: Classification Report for RF

| | Precision | Recall | F1-score | Support |
|---------------|-----------|--------|----------|---------|
| 0 | 0.64 | 1.00 | 0.78 | 114 |
| 1 | 0.00 | 0.00 | 0.00 | 63 |
| Accuracy | 0.64 | | | 177 |
| Macro Avg. | 0.32 | 0.50 | 0.39 | 177 |
| Weighted Avg. | 0.41 | 0.64 | 0.50 | 177 |

Now, Visualization between actual value and predicted value.

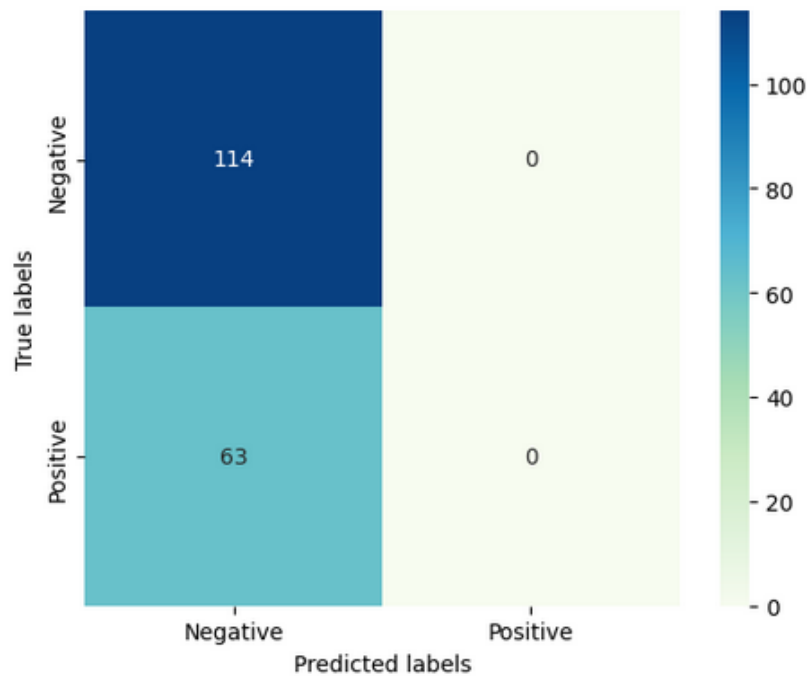


Fig 4.2: Visualization between actual and predicted value of RF

Decision Tree Classifier:

This graph illustrates various scores for recall, precision, accuracy, F1 score, and several code portions. It demonstrates how the multinomial algorithm compares each score.

TABLE 4.5: Confusion Matrix for DT

| | True | False |
|-----------------|-------------|--------------|
| Positive | 114 | 0 |
| Negative | 63 | 0 |

In this case, TP = 114, FP = 0, TN = 63 and FN = 0

Now, the classification report

TABLE 4.6: Classification Report for DT

| | Precision | Recall | F1-score | Support |
|---------------|-----------|--------|----------|---------|
| 0 | 0.64 | 1.00 | 0.78 | 114 |
| 1 | 0.00 | 0.00 | 0.00 | 63 |
| Accuracy | 0.64 | | | 177 |
| Macro Avg. | 0.32 | 0.50 | 0.39 | 177 |
| Weighted Avg. | 0.41 | 0.64 | 0.50 | 177 |

Now, Visualization between actual value and predicted value.

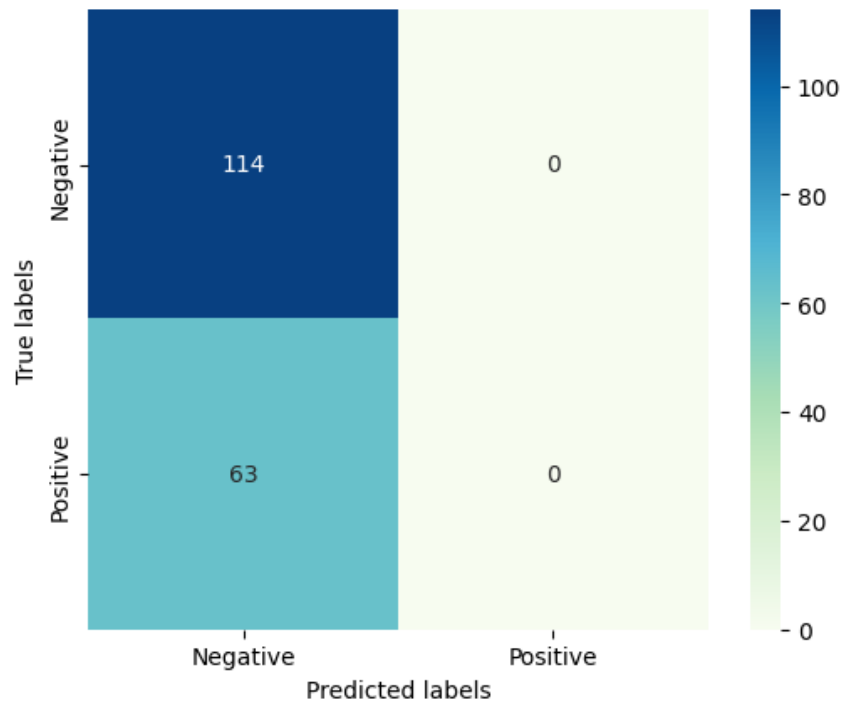


Fig 4.3: Visualization between actual and predicted value of DT

Support Vector Machine Classifier:

This graph illustrates various scores for recall, precision, accuracy, F1 score, and several code portions. It demonstrates how the multinomial algorithm compares each score.

TABLE 4.7: Confusion Matrix for SVM

| | True | False |
|----------|------|-------|
| Positive | 114 | 0 |
| Negative | 63 | 0 |

In this case, TP = 114, FP = 0, TN = 63 and FN = 0

Now, the classification report

TABLE 4.8: Classification Report for SVM

| | Precision | Recall | F1-score | Support |
|---------------|-----------|--------|----------|---------|
| 0 | 0.64 | 1.00 | 0.78 | 114 |
| 1 | 0.00 | 0.00 | 0.00 | 63 |
| Accuracy | 0.64 | | | 177 |
| Macro Avg. | 0.32 | 0.50 | 0.39 | 177 |
| Weighted Avg. | 0.41 | 0.64 | 0.50 | 177 |

Now, Visualization between actual value and predicted value.

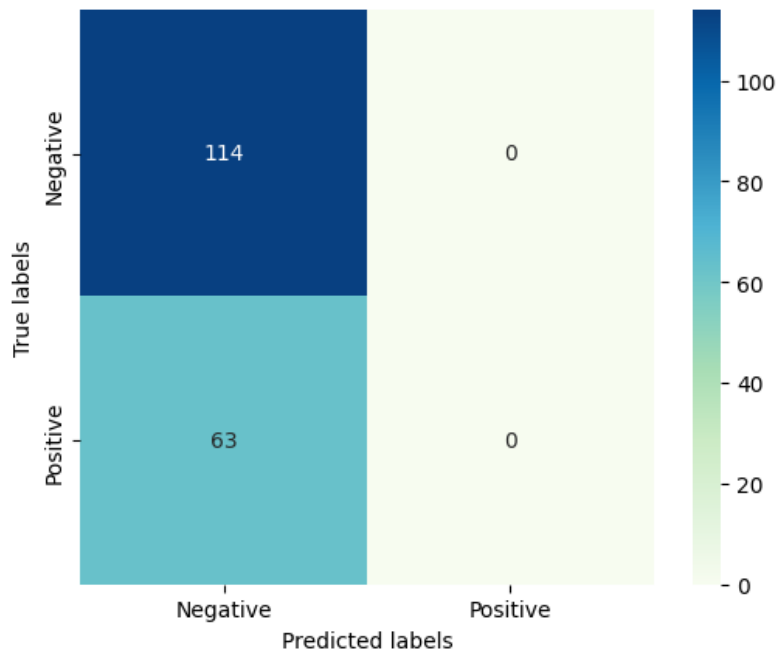


Fig 4.4: Visualization between actual and predicted value of SVM

K Neighbors Classifier:

This graph illustrates various scores for recall, precision, accuracy, F1 score, and several code portions. It demonstrates how the multinomial algorithm compares each score.

TABLE 4.9: Confusion Matrix for KNN

| | True | False |
|-----------------|-------------|--------------|
| Positive | 114 | 0 |
| Negative | 63 | 0 |

In this case, TP = 114, FP = 0, TN = 63 and FN = 0

Now, the classification report

TABLE 4.10: Classification Report for KNN

| | Precision | Recall | F1-score | Support |
|---------------|-----------|--------|----------|---------|
| 0 | 0.64 | 1.00 | 0.78 | 114 |
| 1 | 0.00 | 0.00 | 0.00 | 63 |
| Accuracy | 0.64 | | | 177 |
| Macro Avg. | 0.32 | 0.50 | 0.39 | 177 |
| Weighted Avg. | 0.41 | 0.64 | 0.50 | 177 |

Now, Visualization between actual value and predicted value.

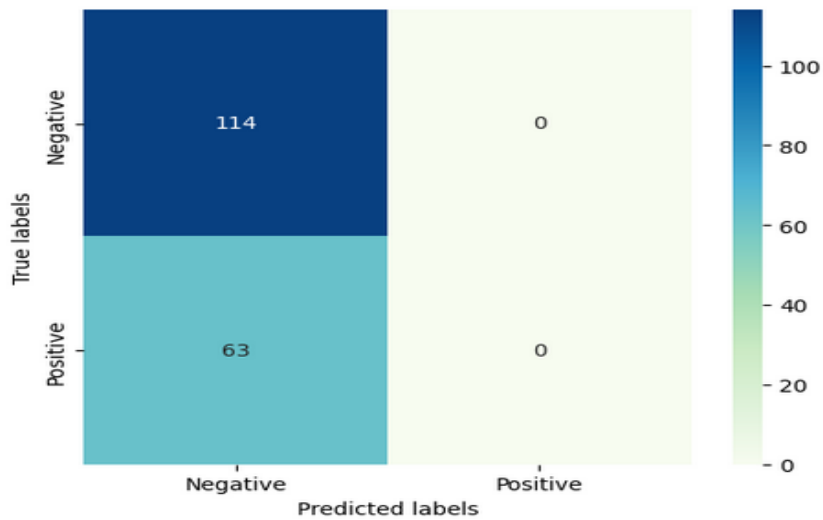


Fig 4.5: Visualization between actual and predicted value of KNN

Logistic regression:

This graph illustrates various scores for recall, precision, accuracy, F1 score, and several code portions. It demonstrates how the multinomial algorithm compares each score.

TABLE 4.11: Confusion Matrix for LR

| | True | False |
|----------|------|-------|
| Positive | 114 | 0 |
| Negative | 63 | 0 |

In this case, TP = 114, FP = 0, TN = 63 and FN = 0

Now, the classification report

TABLE 4.12: Classification Report for LR

| | Precision | Recall | F1-score | Support |
|---------------|-----------|--------|----------|---------|
| 0 | 0.64 | 1.00 | 0.78 | 114 |
| 1 | 0.00 | 0.00 | 0.00 | 63 |
| Accuracy | 0.64 | | | 177 |
| Macro Avg. | 0.32 | 0.50 | 0.39 | 177 |
| Weighted Avg. | 0.41 | 0.64 | 0.50 | 177 |

Now, Visualization between actual value and predicted value.

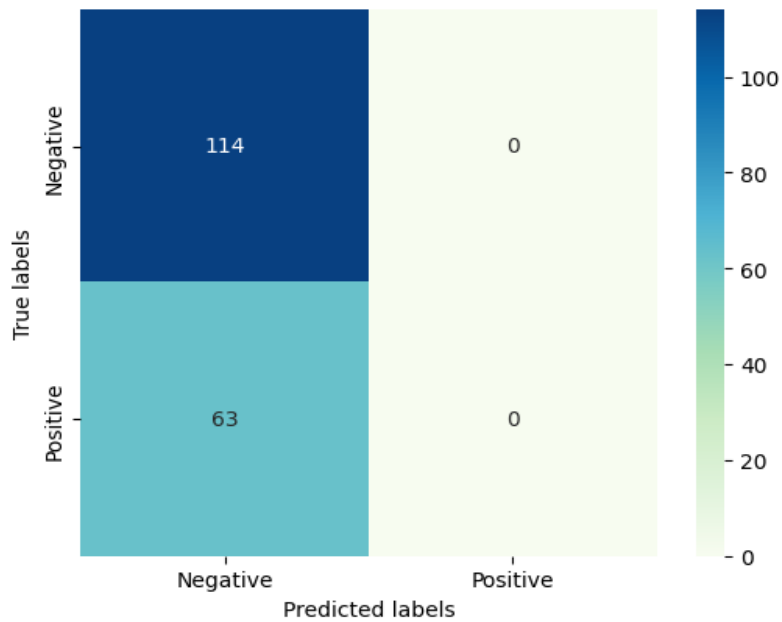


Fig 4.6: Visualization between actual and predicted value of LR

Long Short-Term Memory:

Now, Visualization between actual value and predicted value,

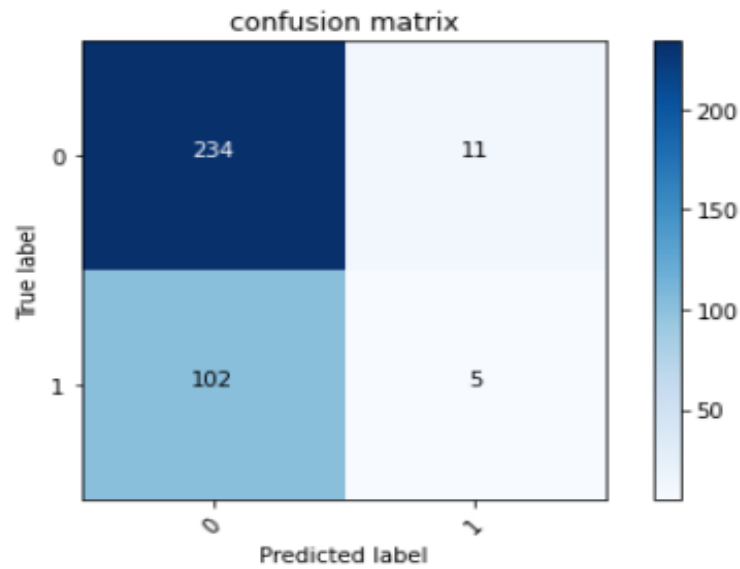


Fig 4.7: Visualization between actual and predicted value of LSTM

Recurrent Neural Network:

Now, Visualization between actual value and predicted value,

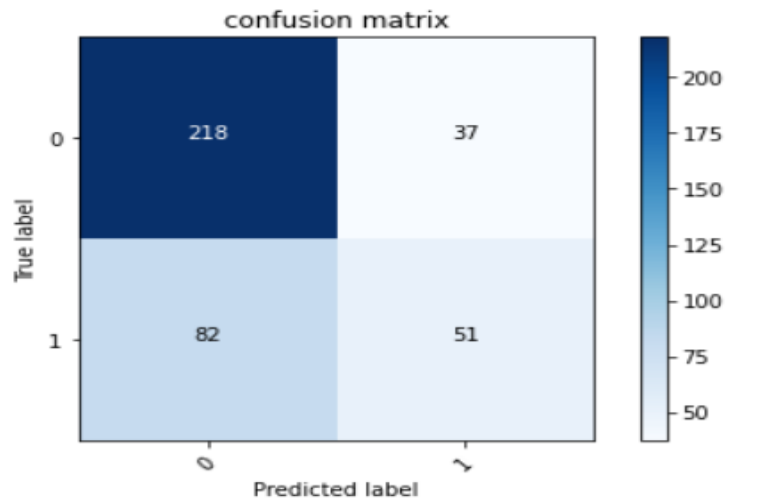


Fig 4.8.: Confusion metrics of RNN

4.1.1 Sentence Tokenization

Sentence tokenization is a technique for decomposing text into individual sentences. For writing, reporting, and formal archiving, the spacy tokenization calculations work wonderfully, and the tokenizer is based on a corpus of systematic English text. The sentence tokenizer performs less well for electronic health records that include truncations, medical terminology, spatial estimates, and other structures that are not common in written English. Because it won't split a single word, the guilty content in substitution structure is faultless saved during the tokenization cycle. Once the specific sentences are written, the opposite changes are made, restoring different information in a number of improved sentences. We should be asking ourselves why sentence tokenization is required but word tokenization is a choice. We must tokenize both the words and the phrases in order to calculate the proportion. Given that the right response is a numeric one, such a yield is essential for machine preparation.

In my analyzed remarks, sentence tokenization.

```
[['sala', 'kamda', 'korlo', 'ki'],  
 ['apnara', 'meye', 'ta', 'matro', 'janlamkhub', 'valo', 'wish', 'best'],  
 ['ata', 'vua', 'news', 'mehjabin', 'onar', 'meye', 'noi'],
```

Fig 4.9: Sentence Tokenization

4.1.2 Test Results and Report

Unusual characters, dialects, and moods are used in the crass statements, such as and others. During the pre-processing hour, I physically eliminated these strange symbolic letters, accents, and Unicode feelings. Additionally, conjunctions are eliminated because they are not required for strict content identification. We have only taken into consideration those classifiers that have been demonstrated to perform better in situations where the prediction challenge can be formulated as a binary classification issue. I used various classification algorithms, and Google Collab was used for the applications. Machine learning aficionados frequently utilize Google Collab, a platform for machine learning, to

run classification models and other essential processes like data filtering, sampling, attribute selection, etc.

My thesis result is shown in table.

Table 4.13: Performance Metrics of Classifier

| Classifier | Accuracy |
|--------------------------|----------|
| Decision Tree | 64.41% |
| Random Forest | 64.41% |
| Support Vector Machine | 64.41% |
| K-Nearest Neighbor | 64.41% |
| Logistic Regression | 64.41% |
| Multinomial Naïve Bayes | 64.97% |
| Long Short-Term Memory | 67.90% |
| Recurrent Neural Network | 69.33% |

From table 4.13 we can see that decision tree, random forest, support vector machine, k-nearest neighbor, and logistic regression gave 64.4% same accuracy which made me very worried. Later for better results I used another model of machine learning multinomial naive bayes algorithm and this algorithm gave 64.97% accuracy. Which is 0.56% higher than the above algorithms. Then I used Long Short-Term Memory (LSTM) model of deep learning algorithm and got 67.90% accuracy which is quite significant. Then I use the Recurrent Neural Network (RNN) model and get 69.33% accuracy which gives more accuracy than all the above models.

4.2 Discussion

As a result of social media's explosive growth in popularity, communicating with people online has never been easier. People can freely express their ideas and opinions on a wide range of topics directly through a variety of social media outlets thanks to advancements in the internet and social media. These social media platforms, including Facebook, Twitter, Instagram, etc., are used by people to exchange information, feelings, and emotions. The

use of social media sites for blackmail, online harassment, and tyranny are all growing quickly as social networking usage rises daily. Negative image sharing, offensive comments, and messages are the main causes of these kinds of situations. This freedom of expression, however, is abused to spread hatred toward specific people or groups of people because of things like race, religion, gender, and so on. It becomes passionate, especially when people have the ability to express themselves verbally and in writing in their native tongue. Spending time on social networking sites like Facebook, Instagram, Twitter, etc. is one of the most popular digital past times today. However, social media has some really amazing applications for increasing reach. Particularly in Asian countries like Bangladesh, where social media use is at an all-time high among the populace. Since everyone has the freedom to free speech, anyone can write or say whatever they want without having to worry about it being accepted. Using foul language has risen in popularity on social media as a result of the absence of monitoring. When people are able to write and communicate in their own tongue, the situation becomes more serious. The spread of disrespectful remarks and messages, as well as sharing of negative images, are the main causes of these circumstances. Spam communications of many kinds, sometimes not screened, interfere with our daily lives. Now, periodically, some people will write offensive statements, and it might be difficult to filter and read them. This causes discomfort for the poster when they read the offensive comments. As a result, adopting this will cut down on online trolling from social media and save a lot of time wasted hunting for compliments. Romanized Data on Cyberbullying My intended research project's title is Bangla Text. I created the suggested method to be able to identify any offensive comments made during talks. By avoiding wasting their time and effort on dangerous, bothersome, and inconsequential things, people, and messages, the public will gain from this endeavor. After organizing all the data into tokenization with the help of the Natural Language Toolkit, I wish to address this problem. Since it can be challenging to weed through social media comments and read every message, it occasionally happens that people will comment on abusive remarks, which can be quite unpleasant for the host. Most of the recipients will be women and young girls. They frequently worry about being bullied if they voice their own thoughts on social media

platforms. On social media, they ought to air any opinions they may have. Romanized Bangla will be used more frequently in internet media outlets. Therefore, the goal of my study is to automatically identify harmful texts. I used Deep Learning and Machine Learning algorithms on the pre-processed data to assess the accuracy of my work. A number of machine learning methods, including Multi Nominal Naive Bayes, Decision Trees, Random Forests, Support Vector Machines, K-Nearest Neighbors, and Logistic Regression, were employed by me. I applied a deep learning model utilizing the Long Short-Term Memory (LSTM) method and a recurrent neural network (RNN) model, which is yet another deep learning methodology. This outperforms my machine learning techniques' accuracy and other deep learning algorithms.

CHAPTER 5

Impact on Society, Environment, Sustainability

5.1 Impact on Society, Environment and Psychological state

Irreverence, which is also known as reviling, swearing, revile words, adjuration words, awful words, and exclamations, is a socially unfriendly vernacular. Obscenity, as it is used in this consciousness, is language that is occasionally seen to be blatantly indecent, unkind, or antagonistic by particular segments of a society. This may be interpreted as the corruption of someone or something, or it may be planned as a solid outpouring from anything. Negative words can have long-lasting effects that extend well beyond the person to whom they were directed.

Words are powerful. Children who grew up in homes where incisive criticism and callous insults were the norm may attest that words can hurt considerably more than physical wounds. This is due to the fact that they ingrain themselves in young personalities and the misery that someone who should respect and admire them—a mother or a father—clearly does not have a favorable impression at all.

Such agonizing harms keep repeating themselves for a very long period. Few young children can escape the scimitar that a cruel parent or other authority figure has named on them.

In 2011, research by Jeffrey Bowers asserted that railing usage has to persuade and modify human behavior. This investigation inspired us to think about adjuration terms and doublespeak in terms of phonetic comparison. A 20-minute trial with 24 individuals ranging in age from 18 to 26 by way of 21 included their reactions, adjuration terms verbally proclaimed resoundingly, and their reactions marked. Additionally, due to the adjuration terms, their electro dermal effect was assessed using household devices that intentionally alter skin blockage. A second Stanford study conducted in 2016 found a direct correlation between irreverence and sincerity. The greatest two US states for irreverence were also the most notable for respectability, according to this investigation of 307

participants. While knowledge believes that kids learn how adoration thus the conduct of consequence, previous theories about children taking up swearing from adult behavior have been discovered incorrectly.

There are numerous educations that emphasize adult emotions, and they are becoming more prevalent as a result of people listening exclusively to unforgiving and callous accessories. Many young people engage in awful behaviors that damage them even more severely, such as sexual immorality, persistent drug use, and even crimes like robbery. These behaviors often include hopeless mental self-views. These young adults have a tendency to snap, which is understandable given that children who grow up in an unstable environment are more likely to become violent than children raised in a more secure environment. Cruelty is bred by callousness. A child who has only experienced jokes, verbal abuse, and mockery will grow up to harass his parents. It is the major route he is aware of for enduring. If any child knows, the effects of using derogatory language are long-lasting and wide-ranging. They will treat others as they have been treated, and this might negatively impact those who never knew the victim as a child but may run into him as an adult in a dark alley.

Singularity has the power to influence how traits that control physical and emotional pressure are expressed. Verifiable terms, like "tranquility" and "amour," might change how certain attributes come across, strengthening certain areas of our frontal projections and enhancing the psychological functioning of the mind. As instructed by the founders, they immediately impel the inspiring focuses of the mind and build strength. However, unpleasant speech can disturb explicit features that are vital to the production of the neurochemicals that protect us from stress. People are made to put parts of our fundamental brains under stress in order to protect us from threats to our endurance, thus this is where our musings usually travel.

In any case, even a single contradicting word can cause our amygdala to become active. It distributes several hormones and synapses that increase pressure, interfering with the function of our brains. (That applies in particular to logic, causation, and dialect.) According to Newberg and Waldman, angry words send warning signals through the brain

and partially loosen the rational and cognitive focuses located in the front projections.

Everything is done through media, such as television, the internet, music, and so forth. The use of awful phrases seems to be inescapable. Obscenity has lost its shocking value as a result of the web's consideration that it will be everywhere and appear unedited. Today, in order to attract attention, one must use a large number of words creatively strung together to convey the most repulsive message.

Without our knowledge, insulting words slowly seep into our lexicon and eventually come to represent a typical reaction to a variety of situations. Swearing can be incredibly disruptive, demonstrate a lack of self-control, and reveal a bad attitude or youth. When people of our age interact with those who are older, this is far more accurate.

5.2 Ethical Aspect

We must refrain from using aggressive, obscene, or curse language in our daily interactions in light of the problems they cause for society and interpersonal relationships. We have created a program that will recognize and identify curse words used in social media or online social platforms specifically for this reason. Not only will this identify, but it will also stop information from being captured on these platforms. Because abusive language is widespread, it is important to stop it from being used. Our curriculum will help create a place that is secure and peaceful.

CHAPTER 6

Summary, Conclusions and Implication for Future Research

6.1 Summary of the Study

Cyberbullying Detection from Romanized Bangla Text is the name of my proposed works. I gather data for my thesis from various celebrities' social media accounts and radio station programs on social media platforms. Using parts of 0 and 1, we leveled the entire dataset. I divided the entire set of data into two sections using the Natural Language Toolkit (NLTK): sentence tokenization and word tokenization. I designed the project so that it can recognize any offensive language in sentences or words used in discussions. This study tries to filter offensive remarks on social media by identifying their nature and displaying supportive ones. This essay is absolutely reliable in order to prevent environmental disruption and emotional pain. But because each region of Bangladesh has its own distinctive dialect, I am currently working on the Romanized Bangla language's standard format and intend to eventually include the regional tongues as well. I have completed my work using a data collection of 1200 texts that have been Romanized in Bangla. However, I plan to expand the size of the data set in the future. I want to learn about various syntactic levels, characters, tokenization of words and languages according to regions, and use of characters in sentences. In theory, better recollection may result from this processing.

6.2 Conclusions

The use of social media sites for blackmail, online harassment, and tyranny are all growing quickly as social networking usage rises daily. Negative image sharing, offensive comments, and messages are the main causes of these kinds of situations. This freedom of expression, however, is abused to spread hatred toward specific people or groups of people because of things like race, religion, gender, and so on. It becomes passionate, especially when people have the ability to express themselves verbally and

in writing in their native tongue. In Bangladesh, people use both English and Romanized Bangla to write comments on social media sites like Facebook and Instagram. These remarks can be both positive and negative. Writing in our mother tongue simply stands out more than writing in other languages. Cyberbullying against individuals on social media has grown as a result. Cyberbullying should be identified and stopped because it can significantly cause emotional and psychological pain. Researchers have put a lot of effort into detecting cyberbullying in English and Bangla among other languages. Since we performed this research to identify cyberbullying in texts written in romanized Bangla. For the greatest results, I have employed both deep learning techniques and several machine learning algorithms. Deep learning models perform better for me than machine learning models. Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) are the two deep learning models I use (LSTM). LSTM provided accuracy of 64.76 %, and RNN provided accuracy of 69.3 %. RNN performs better than any of the aforementioned models. I've made the decision to aim for 99% precision in local languages. In order to tokenize them in word- and region-based language, I intend to do research employing various degrees of syntactic analysis and characters.

6.3 Implication for Future Research

Our plan has several obstacles, thus the suggested technique can be expanded with the supplementary future works.

- I'll try to implement the complete plan to make it more efficient and computerized.
- To acquire more accurate results, new highlights can be added.
- To obtain a more precise result, the informative index's number of comments must be increased. That means the size of the data set should be increased
- When people browse various social websites using programs or mobile applications, an application can be made to detect harmful texts.
- I'll attempt to improve regional language accuracy

REFERENCES

- [1] "Translate Day," 10 01 2023. [Online]. Available: <https://www.translateday.com/most-spoken-languages-in-the-world/>.
- [2] BTRC, "Bangladesh Telecommunication Regulatory Commission," 12 2021. [Online]. Available: <http://old.btrc.gov.bd/content/internet-subscribers-bangladesh-december-2021>
- [3] Datareportal, "Datareportal," 2022. [Online]. Available: <https://datareportal.com/reports/digital-2022-bangladesh>.
- [4] Dhakatribune, "Dhakatribune," 2022. [Online]. Available: <https://archive.dhakatribune.com/bangladesh/dhaka/2019/02/05/unicef-prevent-online-bullying-harassment-of-children-in-bangladesh>
- [5] Md. Tofael Ahmed, Maqsur Rahman, Shafayet Nur, Abu Zafar Muhammad Touhidul Islam, & Dipankar Das. (2022). Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts. TELKOMNIKA (telecommunication, Computing, Electronics and Control), 20(1), 89–97. <https://doi.org/10.12928/telkomnika.v20i1.18630>
- [6] M. T. Ahmed, M. Rahman, S. Nur, A. Islam and D. Das, "Deployment of Machine Learning and Deep Learning Algorithms in Detecting Cyberbullying in Bangla and Romanized Bangla text: A Comparative Study," *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2021, pp. 1-10, doi: 10.1109/ICAECT49130.2021.9392608
- [7] M. G. Hussain, T. A. Mahmud and W. Akthar, "An Approach to Detect Abusive Bangla Text," *2018 International Conference on Innovation in Engineering and Technology (ICIET)*, 2018, pp. 1-5, doi: 10.1109/CIET.2018.8660863.
- [8] M. I. Hossain Junaid, F. Hossain and R. M. Rahman, "Bangla Hate Speech Detection in Videos Using Machine Learning," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, pp. 0347-0351, doi: 10.1109/UEMCON53757.2021.9666550.
- [9] Romim, N., Ahmed, M., Islam, M., Sharma, A. S., Talukder, H., & Amin, M. R. (2022). BD-SHS: A Benchmark Dataset for Learning to Detect Online Bangla Hate Speech in Different Social Contexts. *arXiv preprint arXiv:2206.00372*.
- [10] Rahman, A.I. *et al.* (2022). Comparative Analysis on Joint Modeling of Emotion and Abuse Detection in Bangla Language. In: Singh, M., Tyagi, V., Gupta, P.K.,

Flusser, J., Ören, T. (eds) *Advances in Computing and Data Sciences. ICACDS 2022. Communications in Computer and Information Science*, vol 1614. Springer, Cham. https://doi.org/10.1007/978-3-031-12641-3_17

- [11] Rahman, Moqsadur & Haque, Summit & Saurav, Zillur. (2020). Identifying and Categorizing Opinions Expressed in Bangla Sentences using Deep Learning Technique. *International Journal of Computer Applications*. 176. 13-17. 10.5120/ijca2020920119.
- [12] M. J. Hossain, D. D. Joy, S. Das and R. Mustafa, "Sentiment Analysis on Reviews of E-commerce Sites Using Machine Learning Algorithms," *2022 International Conference on Innovations in Science, Engineering and Technology (ICISSET)*, 2022, pp. 522-527, doi: 10.1109/ICISSET54810.2022.9775846.
- [13] S. S. Ritu, J. Mondal, M. M. Mia and A. A. Marouf, "Bangla Abusive Language Detection using Machine Learning on Radio Message Gateway," *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, 2021, pp. 1725-1729, doi: 10.1109/ICCES51350.2021.9489131.
- [14] E. A. Emon, S. Rahman, J. Banarjee, A. K. Das and T. Mitra, "A Deep Learning Approach to Detect Abusive Bengali Text," *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 2019, pp. 1-5, doi: 10.1109/ICSCC.2019.8843606.
- [15] Ahmed, Md Faisal & Mahmud, Zalish & Biash, Zarin & Ryen, Ahmed & Hossain, Arman & Ashraf, Faisal. (2021). Cyberbullying Detection Using Deep Neural Network from Social Media Comments in Bangla Language.

Mojnu_421

ORIGINALITY REPORT

14%

SIMILARITY INDEX

12%

INTERNET SOURCES

7%

PUBLICATIONS

4%

STUDENT PAPERS

PRIMARY SOURCES

1

dspace.daffodilvarsity.edu.bd:8080

Internet Source

5%

2

Submitted to Daffodil International University

Student Paper

3%

3

Sumaiya Salim Ritu, Joysurya Mondal, Md. Moinu Mia, Ahmed Al Marouf. "Bangla Abusive Language Detection using Machine Learning on Radio Message Gateway", 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021

Publication

1%

4

istina.ipmnet.ru

Internet Source

1%

5

Submitted to University of Melbourne

Student Paper

1%

6

claritynlp.readthedocs.io

Internet Source

1%

7

"Advances in Computing and Data Sciences", Springer Science and Business Media LLC,

1%