OPTIMIZING SENTIMENT CLASSIFICATION IN BANGLA TEXTS USING ENSEMBLE MODELS AND LSTM NETWORKS

BY

Md. Neamoth Ullah
ID: 162-15-8202

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

Supervised By

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md.Abbas Ali Khan
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

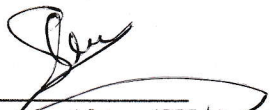
JUNE, 2024

APPROVAL

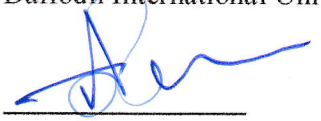
This Project titled "**Optimizing Sentiment Classification In Bangla Texts Using Ensemble And LSTM Networks**", submitted by **Md. Neamoth Ullah**, ID No: **162-15-8202** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 15 July 2024

BOARD OF EXAMINERS

Dr. S.M Aminul Haque (SMAH) — **Board Chairman**
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Sharmin Akter (SNA) — **Internal Examiner 1**
Sr. Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



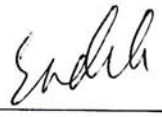
Tapasy Rabeya(TRA) — **Internal Examiner 2**
Sr. Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zulfiker Mahmud(ZM) — **External Examiner**
Professor
Department of CSE
Jagannath University

DECLARATION

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

Supervised by:

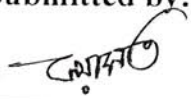


Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Md. Abbas Ali Khan
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Md. Neamoth Ullah
ID: 162-15-8202
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 **Md. Sadekur Rahman,Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in this field 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.Sheak rashed Haider Noori, 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 patience of our parents.

ABSTRACT

Sentiment analysis is a way where huge amount of data can be categorized into different sentiments, emotions, attitudes or opinions. Using sentiment analysis, public views are collected can be given prediction of the class that the views belong to. So, for deriving marketing intelligence sentiment analysis can be performed. It is a vast research field in the NLP sector. Deep learning is part of that machine learning algorithm but it works on in depth similarly like the working behavior of the human brain. The study shows a deep learning approach to sentiment analysis using social media's Bengali dataset. The dataset is a representation of views of audience of social media's Bengali contents and is consists of around 2994 Bengali comments extracted from social media named Facebook and YouTube posts and videos. Here Positive, Negative and Neutral classes are used to categorize the Bengali data. Also, Keras tokenizer is used to tokenize the Bengali text and to convert it into integer sequence and Keras model was used to run the deep learning algorithms LSTM and Bi-LSTM. And these deep learning approach such as LSTM and Bi Directional LSTM has been performed on the dataset and Bi Directional LSTM has the highest accuracy of 72.25%.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

NLP Natural Language Processing

ML Machine Learning

LSTM Long Short-Term Memory

Bi-LSTM Bi directional Long Short-Term Memory

CNN Convolutional Neural Network

RNN Recurrent Neural Network

CHAPTER 1

Introduction

1.1 Introduction

Sentiment analysis is a process that extracts sentiments, emotions, opinions, and attitudes from data [15]. The data used in this analysis reflects individual views on specific topics or products. This data can be in various forms, such as images, videos, audio, or text, and the extracted sentiments essentially represent audience views. These individual views are categorized into sentiments, emotions, opinions, and attitudes. Sentiments are typically labeled into three categories: positive, negative, and neutral, which is called multi-modal analysis. Some approaches divide these three sentiments further, creating a more detailed, fine-grained sentiment analysis with categories such as positive, more positive, neutral, negative, and more negative. When sentiment analysis focuses solely on positive and negative sentiments, it is known as binary sentiment analysis [15].

Sentiment analysis is a classification problem, where predictions are made using different classifiers. The input data is processed to determine a polarity score, which indicates the sentiment (positive, negative, or neutral) of a specific text [7]. Emotions, as complex states, have various stages and can range widely, typically categorized as happy, sad, angry, etc. [8], [9], [20]. Attitudes reflect how individuals treat a particular topic or product, while opinions represent their raw views. Textual data, often collected from social media posts and blogs, is predominantly used for various kinds of sentiment analysis. Some analyses are performed on websites generating substantial user views [4], but social media data is most commonly used. Social media content, including comments on posts, provides a rich source of data, as many individuals and groups share their views and interact with audiences. This includes Bengali content, which is growing on social media, leading to a significant number of Bengali comments available for sentiment analysis.

There are three main approaches to performing sentiment analysis: lexicon-based, automated (or machine learning-based), and hybrid. The lexicon-based approach uses a dictionary or corpus where words are already categorized according to sentiments and compares preprocessed text

with this reference to determine the polarity of the text [1]. The automated approach employs machine learning algorithms, which can be either supervised or unsupervised. In supervised learning, datasets are labeled, allowing the algorithm to be trained. In unsupervised learning, datasets are not labeled, requiring the algorithm to derive insights from the raw data. Examples of machine learning algorithms used include Support Vector Machines and Naive Bayes. The hybrid approach combines lexical and machine learning techniques. Deep learning, a subset of machine learning, involves algorithms inspired by human brain neurons. Examples include CNN, RNN, LSTM, and Bi-LSTM. In this paper, deep learning algorithms LSTM and Bi-LSTM are applied to analyze sentiment in Bengali datasets, specifically comments from social media content on Facebook and YouTube.

1.2 Motivation

Bengali is the fifth most spoken native language and the seventh most spoken language in the world. According to UNESCO, it is also considered the sweetest language globally [21]. Natural Language Processing (NLP) is a field that enables machines to interact with human languages, and extensive research has been conducted on various human languages. Sentiment analysis, which primarily focuses on textual data, has seen significant work in many languages, especially English, due to its status as an international language. However, Bengali language sentiment analysis lags behind, primarily due to its complexity and limited resources [15]. This research aims to address this gap by focusing on sentiment analysis in the Bengali language.

Bengali is a rich language with a deep reservoir of emotions and globally recognized literary value. Despite this, its usage in research fields like sentiment analysis remains insufficient. The current digital era, characterized by an explosion of social media data and the growing demand for content, highlights this disparity. Bengali content creators, influencers, and brands are increasingly populating social media with content, leading to a rise in Bengali comments. However, very little work has been done to collect and analyze this social media data for sentiment analysis. The growing presence of Bengali content on social media platforms presents a unique opportunity for sentiment analysis.

As Bengali content continues to rise on social media, there is an increasing public reaction, especially on platforms like Facebook and YouTube. This trend underscores the potential impact

of sentiment analysis in this field. Given the limited research in Bengali sentiment analysis, there is a significant opportunity to explore and work with this data. Motivated by this potential, we aim to conduct sentiment analysis on Bengali social media data from Facebook and YouTube. Our goal is to enrich the knowledge in this field and contribute to the growing body of research on Bengali language sentiment analysis.

1.3 Rationale of the Study

Social media is teeming with data and information, making it a prime source for sentiment analysis to uncover the opinions behind individual posts. In our study, we performed sentiment analysis on a Bengali language dataset derived from social media platforms like Facebook and YouTube. As the usage of these platforms continues to grow, the volume of Bengali user comments reacting to various contents is also increasing. Influencers, content creators, businesses, and entertainment brands are producing a wide array of content, from product promotions to cooking videos, all contributing to the social media ecosystem. These interactions generate valuable datasets, such as product reviews and cooking reviews, which consist of public reactions captured in comments. Our study aims to perform sentiment analysis on these social media comments to reveal public insights and create a publicly available Bengali dataset. The scarcity of Bengali datasets was a significant challenge we faced, highlighting the need to expand sentiment analysis research in the Bengali language.

1.4 Research Questions

- How does the performance of traditional machine learning algorithms compare to deep learning models in Bangla sentiment analysis?
- What impact does class imbalance have on the performance of sentiment classification models, and what techniques can effectively mitigate this issue?
- Can ensemble methods improve the accuracy and robustness of sentiment classification models in Bangla text data?

1.5 Expected Output

In this section the result of our work that will happen in this field are discussed. As our dataset which was collected from social media such as Facebook, YouTube Bengali comments has three

types of sentiments such as positive, negative, neutral. And using these datasets to perform sentiment analysis we use deep learning algorithm LSTM and Bidirectional LSTM. The expected outcome of these work that are expected to enrich some part of NLP field are given below:

- To help the Bengali content creator of the social media on how their content is getting appreciated by the audience using our research work without reading all the comment.
- This study will introduce a publicly available dataset of 2994 Bengali comments are collected from Facebook posts and YouTube videos so that researcher can use these datasets to perform their task.

1.6 Report Layout

In the layout section, we will outline the specific purpose of each section in the study. The introduction discusses the inspiration behind the study, its objectives, and the expected outcomes. Chapter 2 reviews related research conducted in this field. Chapter 3 details the methodology and the process of developing the proposed model. Chapter 4 presents the results of the model. Chapter 5 examines the impact of this work on society and the environment. Finally, Chapter 6 provides conclusions and discusses potential future improvements for the study.

CHAPTER 2

Background Studies

2.1 Terminologies

Some basic terminologies are given below:

Supervised Learning

In supervised learning, machines learn from labeled datasets used for training. The presence of labeled data is what sets this approach apart from unsupervised learning, as it guides the learning process and allows the machine to understand and predict based on predefined outputs.

Unsupervised Learning

In unsupervised learning, the data consists of unlabeled raw data, and the machine learns patterns from this data during the training phase. Without labeled outputs, the machine identifies inherent structures within the data.

Machine Learning Algorithms

Machine learning algorithms enable machines to perform tasks by processing input data to achieve desired results. Popular machine learning algorithms include Support Vector Machines (SVM), Decision Trees (DT), Naive Bayes (NB), and Random Forest (RF). These algorithms are widely used for various applications in machine learning.

Neural Network

A neural network is a collection of algorithms inspired by the human brain, consisting of interconnected nodes or neurons arranged in layers. This architecture is known as an Artificial Neural Network (ANN). Typically, a
For instance, in a simple neural network depicted in [22], the input layer consists of three neurons, the hidden layer has five neurons, and the output layer contains one neuron. Inputs are processed through the hidden layer to generate an output. Examples of neural network architectures include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM).

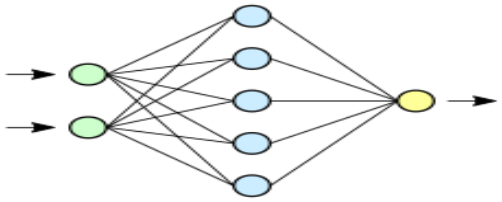


Figure 2.1.1: Neural Network Architecture

Recurrent Neural Network: A Recurrent Neural Network (RNN) is an algorithm designed to work with sequential data or data in an ordered series, where each data point is related to the others in the sequence. This network features internal memory, which retains the previous output, making it particularly effective for sequential data. This memory capability distinguishes RNNs from other neural networks [23].In an RNN, input X is passed into the hidden layers A, producing an output h. This output is then cycled back into the hidden layers for t time steps. For example, when unfolding the process, X_0 is passed into the hidden layers, generating an output h_0. This output h_0 becomes the new input X_1 for the next cycle, and the process repeats to produce h_1. This continues for t time steps, ultimately generating the final output h_t. The internal memory of an RNN allows it to incorporate past outputs into current computations, making it highly suitable for tasks involving sequential data.

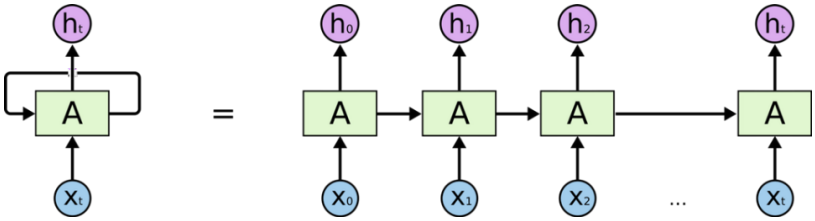


Figure 2.1.2: Recurrent Neural Network Architecture

LSTM: Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that addresses the limitations of traditional RNNs. While RNNs have short-term memory, allowing them to remember recent outputs, they struggle with long-term dependencies, meaning they cannot retain information from outputs that occurred much earlier. LSTM networks, on the other hand, are designed with long-term memory capabilities, making them suitable for tasks requiring the retention of information over extended periods.

The key difference between RNNs and LSTMs lies in their memory capabilities. RNNs are effective for short-term dependencies but fail to capture long-term relationships. LSTMs were introduced in the deep learning world to overcome this limitation. The architecture of an RNN typically includes a single neural network layer, whereas an LSTM architecture consists of four interacting neural network layers. This design allows LSTMs to maintain and update long-term memory effectively.

For example, figure 2.1.3 illustrates a basic RNN architecture with a single neural network layer, while figure 3.3.4 demonstrates the LSTM network architecture, highlighting its four neural network layers. These layers interact with each other to manage long-term dependencies, making LSTMs a powerful tool for sequential data processing [23].

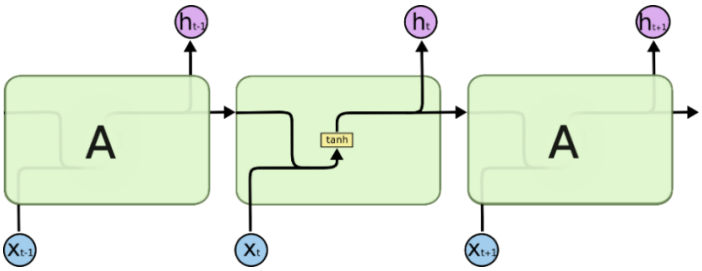


Figure 2.1.3: Recurrent Neural Network Architecture 2

LSTM networks feature a state known as the cell state, which functions like a chain or conveyor belt. This cell state is regulated by gates that control the flow of information, determining which information should be retained and which should be discarded. There are three types of gates in an LSTM network:

Forget Gate: Decides which information should be forgotten.

Input Gate: Determines which information should be added, updated, or passed to the cell state.

Output Gate: Determines which information should be output.

The four neural network layers in an LSTM pass information to these gates, which in turn regulate the cell state. The gates play a crucial role by selectively adding important information and discarding less important data. This selective process enables LSTMs to retain and prioritize relevant information, making them highly effective for tasks involving sequential data.

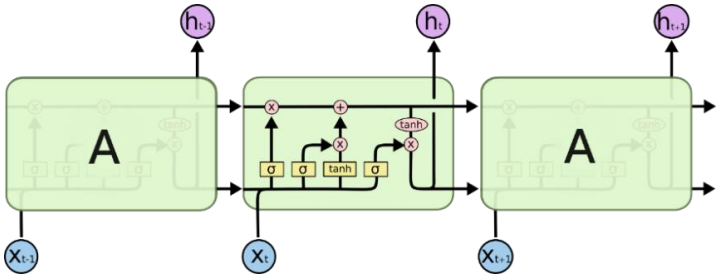


Figure 2.1.4: LSTM Network Architecture

Bi-LSTM: A Bidirectional Long Short-Term Memory (Bi-LSTM) network is a type of recurrent neural network that incorporates two LSTM layers. While a standard LSTM network has a single LSTM layer in its architecture, a Bi-LSTM network features two LSTM layers. In a Bi-LSTM network, one LSTM layer processes the input sequence in the forward direction, from input to output, while the other LSTM layer processes the input sequence in the backward direction.

The architecture of a Bi-LSTM network, as illustrated in figure 2.1.5 [23], shows these two LSTM layers working in tandem—one moving forward and the other moving backward. This bidirectional approach allows the network to capture information from both past and future contexts, enhancing its ability to understand sequential data.

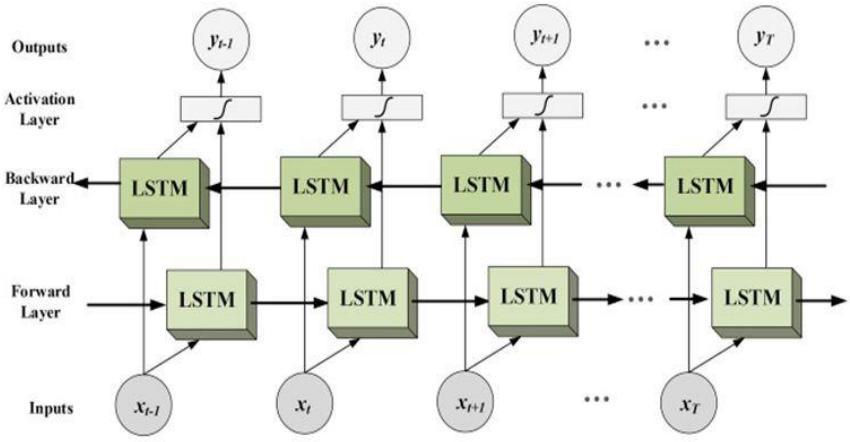


Figure 2.1.5: Bi-LSTM Network Architecture

2.2 Related Works

Sentiment analysis has been extensively performed in many languages, but our study focuses on the Bengali language. Below, we discuss notable works on sentiment analysis in Bengali, highlighting their methodologies and findings.

Das and Bandyopadhyay [1] developed BanglaSenti, a lexicon-based corpus containing 61,582 Bengali words labeled as positive, negative, or neutral. This corpus was created by translating English words with sentiments from SentiWordNet into Bengali, thus providing a foundational resource for sentiment analysis in the Bengali language.

Alam et al. [2] conducted sentiment analysis on a dataset of 2,000 Bengali book reviews to detect sentiment polarity. They used tf-idf values with n-gram features and tested several algorithms, including LR, MNB, RF, DT, KNN, SVM, and SGD, with 10-fold cross-validation. MNB achieved the highest accuracy of 84%.

Naskar et al. [3] introduced the BAN-ABSA dataset, comprising 9,009 comments from Bengali news portals. These comments were manually annotated with aspects such as sports, politics, religion, and others, and classified into positive, negative, and neutral sentiments. Using supervised learning approaches (SVM, CNN, LSTM, Bi-LSTM), Bi-LSTM performed best in aspect extraction with 78.75% accuracy.

Hoque et al. [4] used a dataset of 7,905 Bengali product reviews, manually labeling sentiments based on user ratings. They tested algorithms such as SVM, KNN, LR, RF, and XGB with 10-fold cross-validation, finding KNN to achieve the highest accuracy at 96.25%.

Dey et al. [5] created a benchmark dataset with 12,000 Bengali YouTube reviews, translating them into English. They employed supervised algorithms (LR, SVM, RF, ET) and unsupervised lexicon-based approaches (VADER, TextBlob, SentiStrength) to classify sentiments. SVM performed best, achieving 93% accuracy for both Bengali and English datasets.

Roy et al. [6] analyzed sentiment polarity in 10,819 Bengali Facebook posts and comments. They used five classical algorithms (SVM, NB, DT, AdaBoost, RF) and two deep learning algorithms (LSTM, CNN), with LSTM achieving the highest accuracy at 96.95%.

Paul et al. [7] used 1,500 Bengali tweets from the SAIL 2015 contest, employing SentiWordNet to augment tweets with sentiment tags. Their proposed model using CNN outperformed a deep belief network (DBN) with 46.80% accuracy.

Biswas et al. [8] proposed an automated system to detect emotions (happy, sad, tender, scared, excited, angry) in 7,500 Bengali texts using NB and Topical approaches at both sentence and document levels. The Topical approach achieved 90% accuracy at the sentence level.

Islam et al. [9] performed sentiment analysis on a depression dataset from social media blogs, classifying posts as happy or sad using ML algorithms (MNB, RF, DT, linear SVC). MNB achieved the best accuracy of 86.67%.

Chowdhury et al. [10] detected sentiment polarity in Bengali tweets using character n-grams and MNB. Their model achieved 48.5% accuracy, outperforming other models in the SAIL 2015 contest.

Rahman et al. [11] adapted the English sentiment classification tool VADER for Bengali without translating data into English. They developed a Bengali polarity lexicon and included stemming, achieving better performance and speed compared to the English VADER.

Sarkar et al. [12] performed 10-fold cross-validation with LSTM on a dataset of 1,500 Bengali tweets from the SAIL 2015 contest, achieving the highest accuracy of 55.27%.

Hasan et al. [13] used transfer learning with multilingual BERT for Bengali sentiment analysis. They tested two datasets (two-class and three-class sentiments) with BERT achieving 71% and 60% accuracy, respectively.

Chowdhury et al. [14] proposed a Bi-LSTM model for detecting sentiment polarity in 10,000 Bengali Facebook comments. The Bi-LSTM model outperformed other models with 85.67% accuracy.

Hossain et al. [15] used a bilingual dataset (Bengali and its English translation) for sentiment classification. SVM achieved the highest accuracy with 65.4% for Bengali and 70.1% for English datasets.

Alam et al. [16] used a doc2vec model on a dataset of 7,000 Bengali sentences from Facebook, employing various algorithms (SGD, DT, KNN, LDA, GNB, SM, Bi-LSTM). Bi-LSTM achieved the highest accuracy at 77.85%.

Khan et al. [17] analyzed sentiment in 1,000 Bengali restaurant reviews. They translated an English benchmark dataset into Bengali and tested DT, RF, and MNB, with MNB achieving the highest accuracy of 80.48% using 6-fold cross-validation.

Rahman et al. [18] proposed a new sentiment analysis approach using word2vec on 16,000 Bengali comments, achieving 75.5% accuracy by combining contextual and sentiment information.

Ali et al. [19] evaluated pre-trained transformer models (BERT, XLM-RoBERTa) for Bengali sentiment analysis, fine-tuning the models to achieve high accuracy (95% for BERT and 97% for XLM-R) on various datasets.

Hossain et al. [20] performed sentiment analysis on 1,601 Bengali comments related to Bangladesh Cricket, classifying emotions and sentiments. SVM achieved the highest accuracy with 64.6% for emotions and 73.5% for sentiments.

These studies demonstrate the diverse methodologies and advancements in sentiment analysis within the Bengali language, highlighting both the challenges and progress in this field.

2.3 Comparative Analysis and Summary

Comparative analysis involves comparing multiple research works to identify their similarities and differences. In this study, we have reviewed twenty research papers on sentiment classification problems in the Bengali language, as discussed in the related works section. This section provides a comparative analysis and summary of these research works.

The table below compares these research works based on sentiment or emotion class, dataset, classification methods, and maximum accuracy:

Table 2.3: Comparative analysis of the mentioned research work

Reference Number	Number of class	Dataset	Methods of classification	Maximum accuracy
[1]	3 sentiments (Positive, Negative, Neutral)	Collected 117660 words from SentiWordNet	Lexicon based	61582 Bengali words corpus.
[2]	2 sentiments (Positive, Negative)	2000 Bengali book review corpus.	1. LR, MNB,RF,DT,KNN, SVM, SGD with10 fold cross validation. 2. Used tf-idf values with n gram features.	MNB with 84% accuracy.
[3]	3 sentiments (Positive, Negative, Neutral)	9009 Bengali comments from news portal.	SVM, CNN, LSTM, Bi-LSTM	Bi-LSTM with 78.75%
[4]	3 sentiments (Positive, Negative, Neutral)	Product review dataset with 7905 reviews.	1. SVM, KNN, LR, RF, XGB with all 10 cross-validation 2. performed oversampling for balancing the classes	KNN with 96.25% accuracy
[5]	2 sentiments (Positive, Negative)	12000 Bengali translated English reviews from YouTube	1.LR,SVM,RF,ET for supervised learning 2.VADER,TextBlob, SentiStength for unsupervised lexicon based approach. 3. Both approach run on the English and Bengali dataset.	SVM with 93% accuracy

[6]	2 sentiments (Positive, Negative)	10819 Bengali textual data from Facebook	SVM, NB, DT, AdaBoost, RF, LSTM and CNN	LSTM with96.95% accuracy
[7]	3 sentiments (Positive, Negative, Neutral)	1500 Bengali tweet dataset which released on SAIL2015 contest	1. CNN,DBN 2. Added sentiment tags from SentiWordNet.	CNN with 46.80% accuracy
[8]	6 emotions (happy, sad, tender, scared, excited, angry)	7500 Bengali sentence dataset for sentence level analysis and 100 Bengali article from news portal for document level analysis.	1. NB and 2.Topical approach 3. Binnig technique is also used here.	Topical approach with more than 90% accuracy
[9]	2 emotions (happy, sad)	Bengali paragraphs collected from blogs, social media etc.	MNB, RF, DT, Linear SVC,KNN,XGB	MNB with accuracy of 86.67%
[10]	3 sentiments (Positive, Negative, Neutral)	used Bengali tweet dataset in 2015 SAIL contest	Character n gram features	MNB has 48.5% accuracy
[11]	2 sentiments (Positive, Negative)	Developed Bengali polarity lexicon from VADER English lexicon	They perform a comparison of proposed Bengali VADER to English VADER.	Bengali VADER performs better
[12]	3 sentiments (Positive, Negative, Neutral)	1500 Bengali Tweets dataset	1.LSTM with 10 cross validation compared with SVM and MNB both with unigram and extra addition of bigram to MNB 2. SentiWordNet for adding lexical knowledge on the	LSTM with 55.27%

			dataset.	
[13]	2 sentiments (Positive, Negative), 3 sentiments (Positive, Negative, Neutral)	17852 comments with 3 class and 13120 comments with 2 class from online news website.	1.BERT 2.word2vec 3.fastText embedding all with 3 class and 2 class sentiments.	BERT with 71% accuracy for 2 class and 60% accuracy of 3 class Sentiments.
[14]	2 sentiments (Positive, Negative)	Facebook Bengali 30000 comments collected and among them 10000 data used	SVM,DT, Logistic Linear Regression and RNN with Bi-LSTM	RNN with Bi-LSTM by 85.67% accuracy
[15]	3 sentiments (Positive, Negative, Neutral)	A game user comment dataset and another one is drama review from YouTube	LR, ET, RF, SVM, RR, LSTM, SVM performed on both dataset.	SVM with 65.4% for Bengali and 70.1% for English dataset.
[16]	2 sentiments (Positive, Negative)	Bengali 7000 Facebook text data	SGD, DT, KN classifier, LDA, Gaussian NB, SM, Bi-LSTM	Bi-LSTM with 77.85% accuracy
[17]	3 sentiments (Positive, Negative, Neutral)	1000 English translated Bengalirestaurant review dataset	DT, RF, MNB with K-fold cross validation used to test the model and so 5-fold, 6-fold, 7-fold has performed	MNB with 6 fold gave 80.48% accuracy
[18]	2 sentiments (Positive, Negative)	Bengali comments with 16000 Bengali text line.	A new approach using word2vec where dataset contextual information and sentiment information of the dataset comments added.	Accuracy is achieved to 75.5%
[19]	2 sentiments (Positive, Negative) 3 sentiments	1.17852 Bengali data from news portal 2.13807 YouTube	Transformer model BERT and XLM-RoBERTa with fine tuning is used here with	XML-RoBERTa wiith 95% accuracy

	(Positive, Negative, Neutral)	Bengali drama review 3.2000 Bengali book review	3 datasets to perform sentiment analysis in Bengali language.	
[20]	3 emotions (Praise, Criticism, Sadness), 3sentiments (Positive, Negative, Neutral)	1. Bangladesh cricket dataset of 1601 Bengali comments from Facebook and newspaper pages. 2. Another Bangladesh cricket dataset with 2979 Bengali comments from BBC Bangla website.	SVM, DT and MNB	1. SVM with 64.596% accuracy for Facebook dataset. 2. SVM with 73.490% accuracy for BBC Bangla dataset.

From the table above, we observe both similarities and differences among the various research works. Some studies share the same sentiment classes, while others utilize datasets from similar domains. Despite these commonalities, each research work is unique in its approach. The algorithms employed may be similar, but their purposes and models differ.

To summarize our work based on the table, our dataset consists of three sentiment classes: positive, negative, and neutral. We used manually collected Bengali comments from Facebook and YouTube, labeling them according to these sentiment classes. For classification methods, we proposed two models: one using LSTM and the other using Bi-LSTM. Among these, the Bi-LSTM model performed better, highlighting its effectiveness in sentiment analysis for Bengali social media comments.

2.4 Scope of the Problem

This study explores sentiment analysis on a Bengali dataset using deep learning algorithms. The dataset consists of 2994 Bengali comments extracted from social media platforms, specifically Facebook and YouTube. The sentiment classes analyzed are positive, negative, and neutral. To

perform the sentiment analysis, two deep learning algorithms, LSTM and Bi-LSTM, were applied to the dataset. This study demonstrates the application of deep learning techniques to sentiment analysis of Bengali social media data. Addressing the common issue of limited resources faced by previous research, this work also aims to assist social media content creators in analyzing their content more efficiently.

2.5 Challenges

Collecting Bengali data was the first challenge we faced in this study. Unlike English data, Bengali data is not readily available, making it difficult to gather comments from platforms like Facebook and YouTube. After collecting the data, each comment had to be entered into an Excel sheet and manually labeled with the appropriate sentiment class, which was a labor-intensive process. Additionally, preprocessing the Bengali text posed significant difficulties. The data cleaning process involved removing stop words and punctuation, which required precise handling and proved to be quite challenging.

CHAPTER 3

Research Methodology

3.1 Introduction

The methodology section outlines the overall process of a research paper, detailing the steps taken to solve the research problem and create a solution. This section systematically describes the procedures designed to generate better results than previous research on the topic. Our research focuses on sentiment analysis on a Bengali social media dataset. While many researchers have worked on sentiment analysis, very few studies exist in the Bengali language. Data collection for sentiment analysis in Bengali has been particularly challenging due to the scarcity of Bengali data. Some researchers have addressed this by using English corpora and translating them into Bengali [5], [17]. Additionally, the lack of benchmark datasets has led some to create lexicon-based dictionaries to improve sentiment word identification and prediction accuracy [1]. Supervised learning is the most popular method in this field [12], and we have adopted this approach in our study.

Both classical algorithms and deep learning algorithms have been used in sentiment classification, with some studies comparing the performance of these algorithms [3], [6], [12], [14], [15], [16]. After data collection, many researchers divide the dataset into training and testing sets [8], [17], and some also include a validation set [2]. Our study follows a similar methodology, with three key sections: data collection, data preprocessing, and building the proposed model.

To execute this methodology, we used the BNLP corpus for data cleaning and the Keras library for data tokenization and model building. Our research topic is "Sentiment Analysis of Social Media's Bangla Data," and this chapter discusses all the procedures needed to build the proposed model. The flowchart of all procedures is shown in Figure 3.1.

Figure 3.1 illustrates the flow of procedures in the proposed model. We have two models: one using the LSTM algorithm and the other using the Bi-LSTM algorithm. Both algorithms are included in the applied algorithm section of the proposed model.

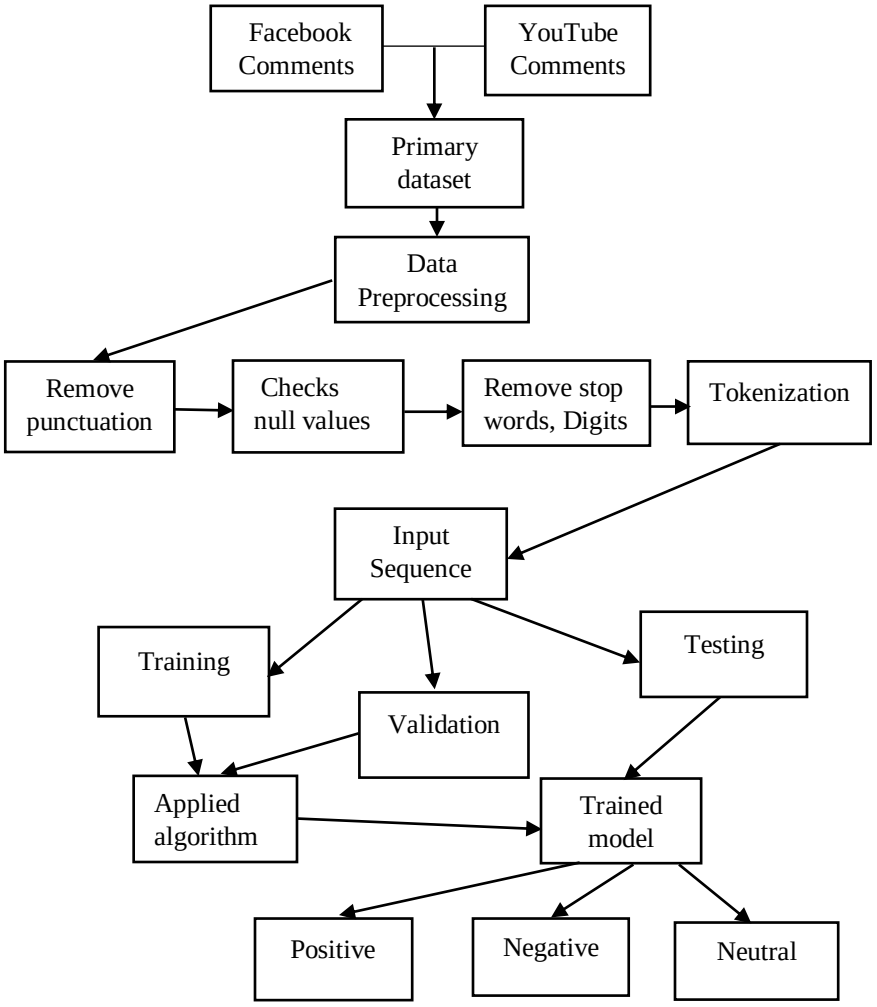


Figure 3.1: Proposed Model

3.2 Research Subject and Instrumentation

Our research subject is titled "Sentiment Analysis of Social Media's Bangla Data." We chose this topic to analyze sentiments in the Bengali language, our mother tongue, where few studies currently exist. Social media platforms are increasingly populated with Bengali data as more Bengali speakers use these platforms daily. Observing this trend, we selected this topic and named the subject accordingly.

Sentiment classification is a supervised learning approach in machine learning that uses either bi-class or multi-class datasets [20]. Bi-class datasets contain two sentiment classes, while multi-

class datasets have more than two sentiment classes. Our dataset includes three sentiment classes, making it a multi-class dataset. For classification, we selected neural networks over classical classifiers because neural networks generally perform better [6].

In this study, we used two deep learning algorithms: LSTM and Bi-LSTM. We chose these algorithms because they outperform many other neural network models. We built two deep learning models—one with LSTM and the other with Bi-LSTM—using the Keras library. Both models include hidden layers, starting with an embedding layer, followed by LSTM or Bi-LSTM layers, and ending with an output layer. We used Keras' Sequential model to add layers and build the LSTM and Bi-LSTM models.

LSTM (Long Short-Term Memory): LSTM is an advanced form of recurrent neural network (RNN). It retains information for longer periods compared to traditional RNNs and uses three gates to determine which data is important and which should be forgotten. This makes LSTM particularly effective for tasks involving sequential data, including sentiment classification. In this study, the LSTM model consists of several layers, as described below:

Embedding Layer: This is the first hidden layer of the neural network. All integer-encoded inputs pass through this layer, which reduces the dimensionality of the input vectors before passing them to the next hidden layer.

LSTM Layer: After the embedding layer, the input is passed to this layer, which contains two neurons or nodes with a dropout rate of 0.2. This layer processes the sequential data to capture temporal dependencies.

Dense Layer/Output Layer: This is a fully connected neural network layer with three neurons, each connected to the previous layer's neurons. It serves as the output layer with a Softmax activation function. The three neurons correspond to the three sentiment classes, and the model outputs one of these sentiments.

The following figure, Figure 3.6.1: LSTM Model, illustrates all the layers of the LSTM in the flowchart.

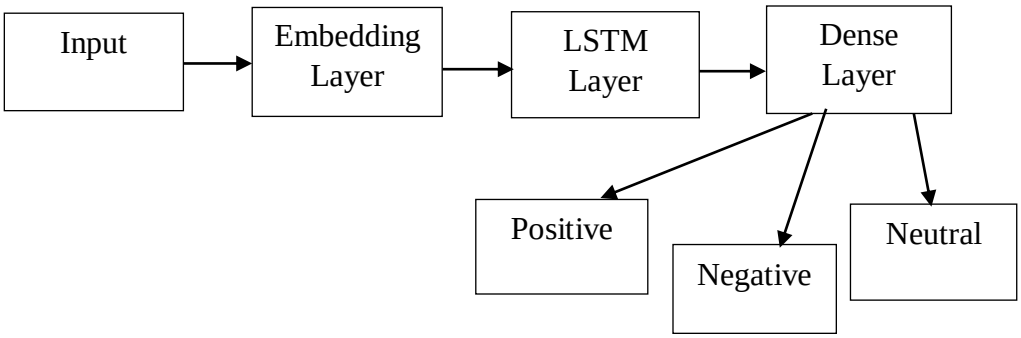


Figure 3.6.1: LSTM Model

Bi-LSTM: Bi-LSTM, or Bidirectional LSTM, is an advanced version of the LSTM network within neural networks. The key difference lies in its ability to process data in both forward and backward directions, unlike the standard LSTM, which only processes data in a single forward direction. This bidirectional capability makes Bi-LSTM particularly effective for tasks involving sequential data, such as sentiment classification. The Bi-LSTM model used in this study consists of the following layers:

Embedding Layer: This is the first layer in the model where the input enters. The input, which is integer-encoded, passes through this layer, which reduces its dimensionality before passing it to the next layer.

Bidirectional LSTM Layer: This hidden layer contains two neurons. The reduced-dimension input from the embedding layer is processed here, with the layer utilizing a dropout rate of 0.2.

LSTM Layer: This is another Bi-LSTM layer, also with two neurons and a dropout rate of 0.2. The input from the previous layer passes through this layer before moving to the output layer.

Dense Layer/Output Layer: This fully connected neural network layer has three neurons, each corresponding to one of the sentiment classes. It serves as the output layer, utilizing the Softmax activation function. One of the three sentiment classes (positive, negative, or neutral) will be selected as the output.

The following figure, Figure 3.6.2: Bi-LSTM Model, illustrates the flow of inputs through all the layers of this Bi-LSTM model.

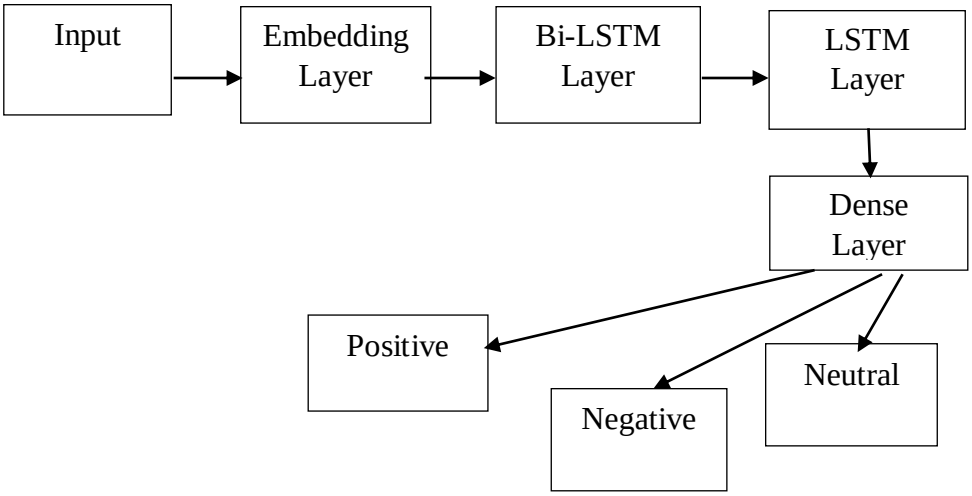


Figure 3.6.2: Bi-LSTM Model

3.3 Data Collection

Data collection is the first step in this overall procedure. Collecting Bengali data is challenging because it is not as readily available as English data. Recently, more people have been expressing their emotions using the Bengali language on social media [8], making data collection somewhat easier than before. However, the scarcity of Bengali data persists, and it is still considered a resource-poor language due to the lack of high-quality datasets and content [2]. This makes collecting Bengali data for sentiment classification a bit difficult.

In our study, we manually collected a dataset consisting of Bengali comments, each labeled with one of three sentiment classes. We used an Excel sheet to organize our data, with columns for comments and corresponding sentiments. The comments column contains the collected data, while the sentiments column contains the manually assigned sentiments for each comment.

The following table, Table 3.4.1: The Bengali Dataset, shows the first ten Bengali comments along with their sentiments.

Table 3.3: The Bengali Dataset

Comments	Sentiments
ধন্যবাদজানাইএইনাটকেরসাথেজড়িয়েথাকাসবাইকে।এতসুন্দরএকটিনাটকআমাদেরকেউপহারদেউয়ারজন্য।	Positive
অসাধারণলাগেতোমারনাটক	Positive
এইভাবেবিশ্বাসটারাখতেশেখোপ্রিয়	Positive
বাহঅসম্ভবসুন্দরলাগতেছে	Positive
মহিলাকিজেলে?	Neutral
এভাইআপনিতোকিছুজানেননইনা,, শুধুবলেনদেখাযাককিহয়.।এইচ্যানেলবন্ধকরেন।	Neutral
আলহামদুলিল্লাহ।মাএ১৫টিভিডিওআপলোডকরে৩৫০০০+সাবস্ক্রাইবহলো।সবইআপনাদেরভালোবাসা।	Positive
তোদেরমতআলতুফালতুবিষয়নিয়েসারাক্ষণশুধুভিডিওবানানোরজন্যওরাপাগলনা	Negative
আপনিছাড়াআরকেহহতাশনাউল্টাপাল্টানিউজনাকরেভেলিডইনফরমেশনদেন।	Negative
আমারতোমনেহয়খানদেরথেকেওস্টারডোমবেশিহৃতিকরোশানের	Positive

For the data collection Google Sheet was used to put the data in one place.

3.4 Data Preprocessing

Language is the core of NLP, and each language has unique rules, spelling, vowels, etc. Due to these differences, data preprocessing is crucial before applying machine learning models. For building a successful machine learning model and fitting Bengali data into it, data preprocessing is essential [9]. Our data preprocessing consists of two parts: 1) Data Cleaning and 2) Tokenization Using Keras Library. In the first part, we cleaned the data, and in the second part, we tokenized it and converted it into a sequence of numbers.

3.4.1 Data Cleaning:

- We started by removing punctuation from the data using Google Sheets.

- Next, we checked for null values in the code. Since our data had no null values, the number of comments remained the same.
- For data cleaning, we used the BNLP corpus. We defined stop words from the BNLP corpus and used them to remove stop words from our Bengali dataset.
- We removed digits using regular expressions.
- After cleaning the data, we performed tokenization on the Bengali data using the Keras library.
- The figure 3.4.1, "Steps of Data Cleaning," illustrates all the steps of data cleaning where the Bengali data are thoroughly cleaned.

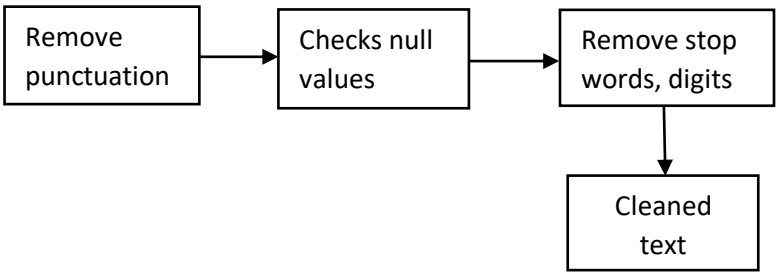


Figure 3.4.1: Steps of Data Cleaning

2) Tokenization Using Keras Library: Tokenization is the process of breaking down larger text into smaller, more manageable pieces. For tokenization in our Bengali dataset, we used the Keras library, which is an open-source library for implementing neural networks. Specifically, we utilized the tokenizer class from the Keras preprocessing module to tokenize our Bengali text.

As our model employs deep learning algorithms, specifically LSTM and Bi-LSTM, both require sequential data to function effectively. The Keras tokenizer class is ideal for this purpose because it not only splits each sentence into words but also converts each word into integers, transforming entire sentences into sequences of integer values. These sequences are then fed into the model.

The two methods from the tokenizer class that convert the words into sequences of numbers are:

fit_on_texts(texts): This method updates the internal vocabulary based on a list of texts. Each word in the texts is assigned a unique integer value.

texts_to_sequences(texts): This method transforms each text into a sequence of integers, where each integer corresponds to a word in the internal vocabulary.

These methods effectively prepare the text data for input into the LSTM and Bi-LSTM models.

The figure 3.4.2: Tokenization Process shows the framework of tokenization that happened here.

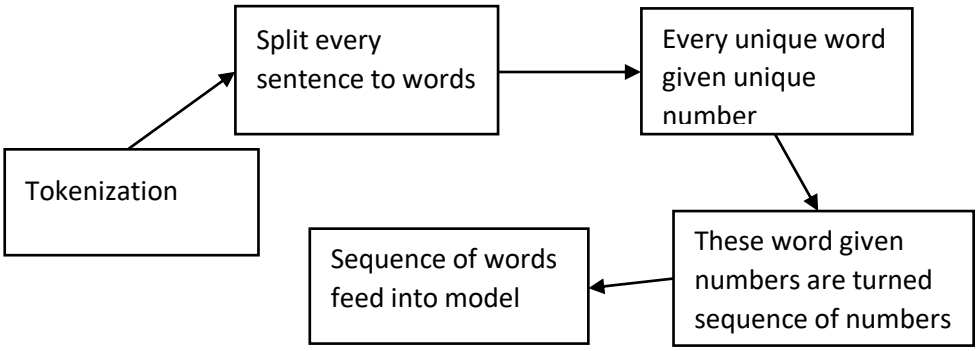


Figure 3.4.2: Tokenization Process

Next, the sequenced numeric values are standardized to the same lengths and then fed into the model. After converting every text into a sequence of numbers, the sequences are padded to ensure uniform size and shape. Additionally, the categorical data representing sentiments are converted into integer variables. These processed sequences and sentiment labels are then fed into the LSTM and Bi-LSTM models in the subsequent section.

3.5 Statistical Analysis

In this study, sentiment classification was performed using a dataset of 4,000 Bengali comments collected from social media platforms Facebook and YouTube. The dataset is categorized into three sentiment classes: positive, negative, and neutral. Below is the distribution of comments based on sentiment class:

Positive comments:1518

Negative comments: 1244

Neutral comments: 232

The below histogram figure 3.5: Data distribution shows the data distribution according to the 3-sentiment class.

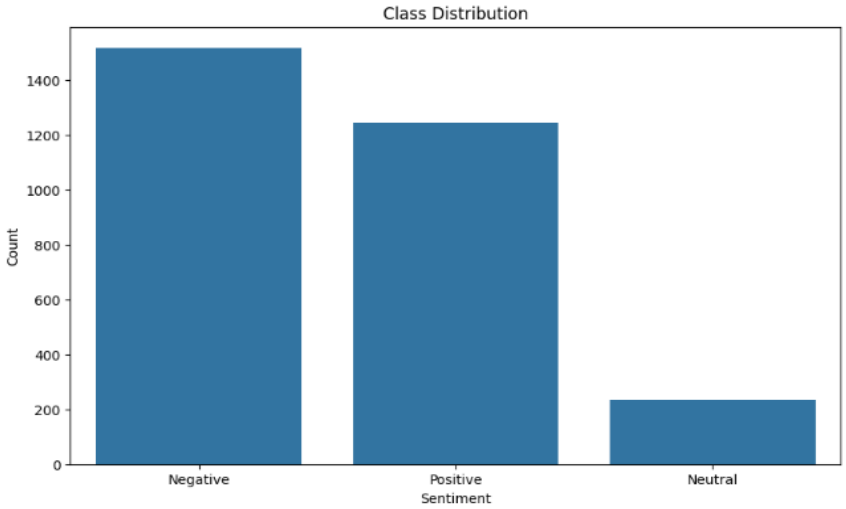


Figure 3.5: Dataset Distribution

3.7 Implementation Requirements

Python was used for all the programming, and Google Colaboratory served as the platform for implementation. Data was collected using Google Sheets. The Pandas library was employed to read the datasets and encode categorical values into integers. The BNLP corpus was utilized for data cleaning. The Keras library was used for tokenizing the data and building the models. The Numpy library was used to display arrays. The Sklearn library was used for splitting the dataset. Matplotlib and Seaborn libraries were used for data visualization.

To perform this task or sentiment analysis, the following components were used:

- Google Colaboratory: Platform for implementing Python code.
- Pandas: Used to read and modify files.
- Numpy: Used to display data arrays.
- Sklearn: Used to split the dataset into training and testing sets.
- Seaborn: Used for visualizing the dataset.
- Matplotlib: Used for visualization purposes.
- Keras: Used to implement algorithms and the tokenization process.
- BNLP corpus: Used for data cleaning.
- Google Sheets: Data was collected using Google Sheets.
- Regular Expression: Used for data cleaning to remove digits.

CHAPTER 4

Experimental Results and Discussion

4.1 Introduction

The experimental results section presents the outcomes of the research study, analyzing these results through various graphical representations. Our task involves sentiment analysis, a method that has been applied by many researchers on different datasets, each demonstrating their respective findings. These results are obtained by implementing various algorithms on the training set and then evaluating them on the testing set. Experimental results can be represented in two ways: statistical and graphical representations. Examples of statistical representations include accuracy, precision, and confusion matrix, while graphical representations include accuracy and loss graphs. These representations are essential for analyzing the results, and their definitions are provided below:

Accuracy: The proportion of correctly predicted classes among all the classes.

Precision: For a particular class in the test set, precision is the number of correctly predicted sentences divided by the total number of sentences predicted as that class by the classifier.

Recall: For a particular class in the test set, recall is the number of correctly predicted sentences divided by the total number of actual sentences in that class.

F1 Score: The weighted harmonic mean of both precision and recall for a particular class.

Classification Report: A comprehensive report combining the values of precision, recall, and F1 score for all the classes in the dataset, used to measure the performance of the classifier.

Confusion Matrix: A table showing the actual class versus the predicted class values according to the classes used in the experiment, used to measure the performance of the classification model.

4.2 Experimental Setup

For the experimental setup, the dataset must first be divided into training, testing, and validation sets. The algorithm will then process the input data and predict the classes. To enhance the accuracy of predictions, certain parameters need to be tuned during the training phase for both

LSTM and Bi-LSTM models. The initial step is splitting the dataset to begin the experiment. The data splitting and parameter tuning processes for both LSTM and Bi-LSTM models are outlined below.

Dataset Splitting: The dataset is divided into three sets: training, validation, and testing. The training set comprises 70% of the data, the validation set 10%, and the testing set 20%. This split is used for both LSTM and Bi-LSTM models. The following figure, Figure 4.2: Dataset Splitting, illustrates the division of the dataset.

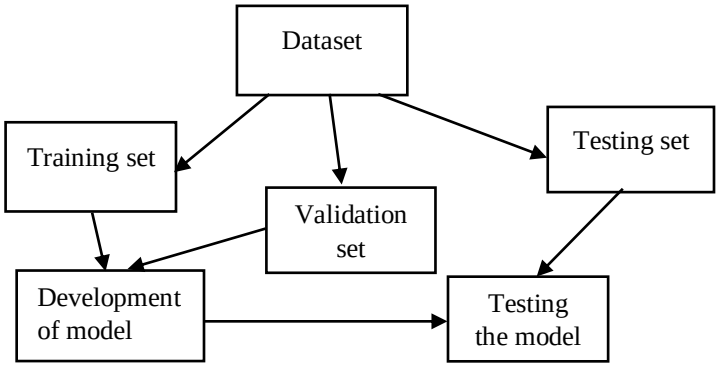


Figure 4.2: Dataset Splitting

For the LSTM layers, various parameters such as the dimension of the LSTM layer, epochs, and batch size play a crucial role in enhancing model accuracy. By tuning these parameters, the model's performance can be significantly improved. The final output layer of the LSTM employs the Softmax activation function, with categorical crossentropy as the loss function, suitable for multi-class classification. The Adam optimizer is used for optimization. The hyper-parameters for the LSTM model are as follows:

Input dimension: The size of the text vocabulary, set at 5000 in the embedding layer.

Output dimension: The size of the output vector in the embedding layer, set at 64.

Dimension of LSTM layer: The number of neurons in the LSTM layer, set at 16 neurons.

Epoch: The number of times the dataset passes through the neural network, set at 40.

Batch size: The number of inputs processed at a time, set at 12.

Hyper-Parameters Tuning for Bi-LSTM

For the Bi-LSTM layers, tuning hyper-parameters such as epochs, batch size, and input dimension is crucial for enhancing model accuracy. Similar to the LSTM model, the output layer of the Bi-LSTM uses categorical crossentropy as the loss function, Adam as the optimizer, and Softmax as the activation function, as both models deal with the same multi-class dataset. The hyper-parameters for the Bi-LSTM model are as follows:

Input dimension: The size of the input dimension is 5000.

Output dimension: The size of the output dimension is 64.

Dimension of Bi-LSTM layer: The Bi-LSTM layer consists of 16 neurons, with an additional LSTM layer also having 16 neurons.

Epoch: The number of epochs is set at 40.

Batch size: The batch size is set at 12.

4.3 Experimental Results & Analysis

In this study, we performed sentiment classification on a dataset of 4,000 Bengali comments collected from social media platforms such as Facebook and YouTube. The dataset was categorized into three sentiment classes: positive, negative, and neutral. The classification was performed using various models, including LSTM, Bi-LSTM, SVM, and Neural Networks. Below is the detailed analysis of the results obtained from these models.

Model Performance

Model	**Sentiment**	**Precision**	**Recall**	**F1-Score**	**Overall Accuracy**	**Micro Avg Precision**	**Macro Avg Recall**	**Macro Avg F1-Score**	**Weighted Avg Precision**	**Weighted Avg Recall**	**Weighted Avg F1-Score**
LR	Negative	0.63	0.80	0.71	65%	0.53	0.51	0.51	0.65	0.65	0.64
	Neutral	0.18	0.12	0.14							
	Positive	0.78	0.61	0.68							
SVM	Negative	0.65	0.72	0.68	63%	0.53	0.52	0.52	0.65	0.63	0.64
	Neutral	0.17	0.22	0.19							

	Positive	0.76	0.63	0.69							
NN	Negative	0.70	0.84	0.76	71%	0.61	0.59	0.59	0.72	0.71	0.71
	Neutral	0.29	0.23	0.26							
	Positive	0.84	0.69	0.76							
Bi-LSTM	Negative	0.71	0.79	0.75	71%	0.59	0.57	0.57	0.69	0.71	0.70
	Neutral	0.31	0.18	0.23							
	Positive	0.77	0.73	0.75							
Ensemble Model	Negative	0.64	0.84	0.73	67%	0.46	0.49	0.47	0.62	0.67	0.64
	Neutral	0.00	0.00	0.00							
	Positive	0.75	0.65	0.70							

Key Observations

Overall Accuracy:

The Bi-LSTM and Neural Network models both achieved the highest overall accuracy of 71%.

The SVM model had a slightly lower accuracy of 63%.

Logistic Regression achieved an overall accuracy of 65%.

The Ensemble Model had an overall accuracy of 67%.

Class-Specific Performance:

For the negative sentiment class, the Neural Network and Bi-LSTM models performed the best with high recall and F1-scores.

For the neutral sentiment class, all models struggled, particularly the Ensemble Model, which showed a F1-score of 0.00.

For the positive sentiment class, the Neural Network model showed the highest precision and recall, followed closely by the Bi-LSTM model.

Macro and Weighted Averages:

The macro averages indicate that the models performed better on the negative and positive classes compared to the neutral class.

The weighted averages suggest that the Bi-LSTM and Neural Network models are slightly better in balancing the performance across all classes.

The Bi-LSTM and Neural Network models demonstrated superior performance compared to SVM, Logistic Regression, and the Ensemble Model. These models were particularly effective in handling the negative and positive sentiment classes but had challenges with the neutral class. The overall accuracy of 71% achieved by Bi-LSTM and Neural Network models highlights their potential in sentiment classification tasks involving Bengali social media data.

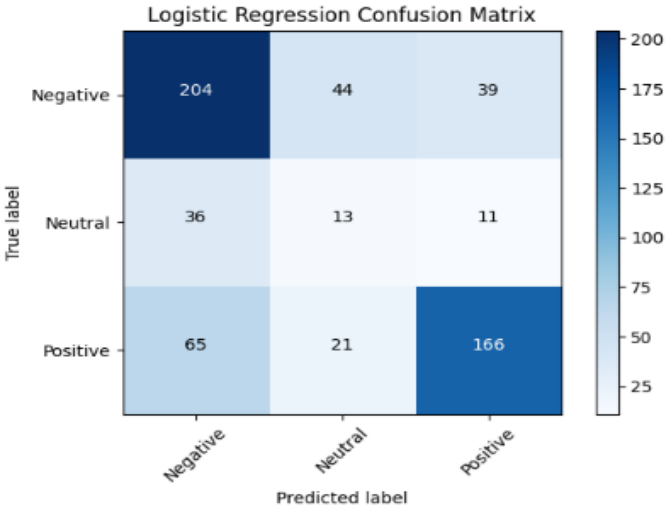


Fig 4.3: Logistic Regression Confusion Matrix

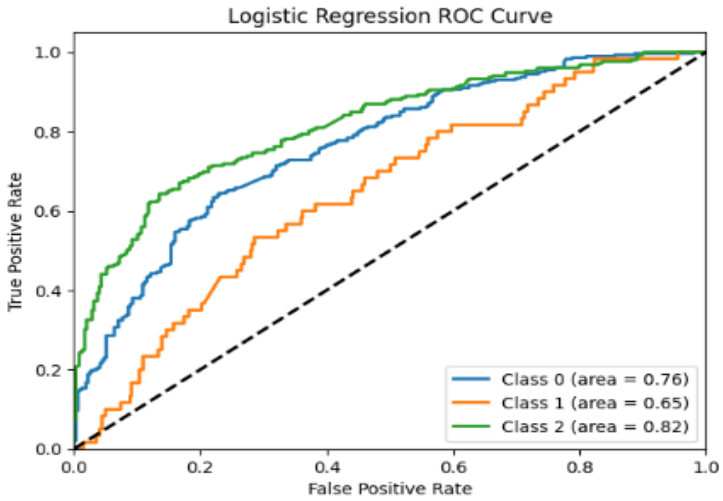


Fig. 4.4: Logistic Regression ROC Curve

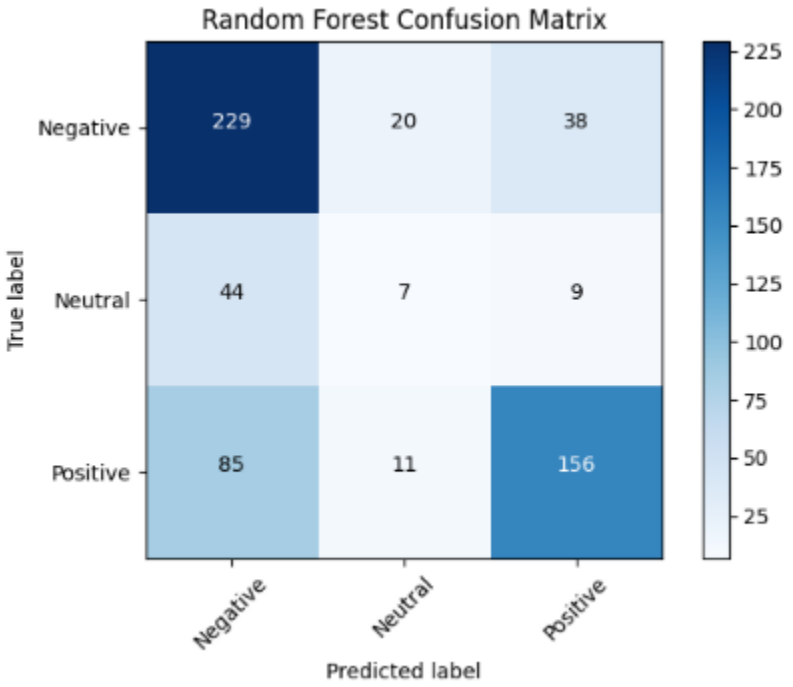


Fig. 4.5: Random Forest Confusion Matrix

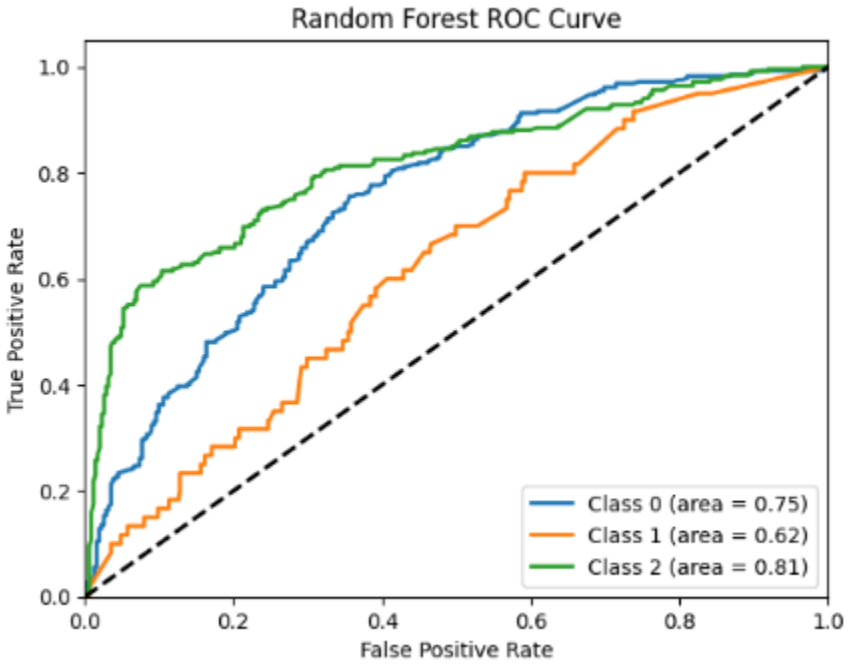


Fig. 4.6: Random Forest ROC Curve

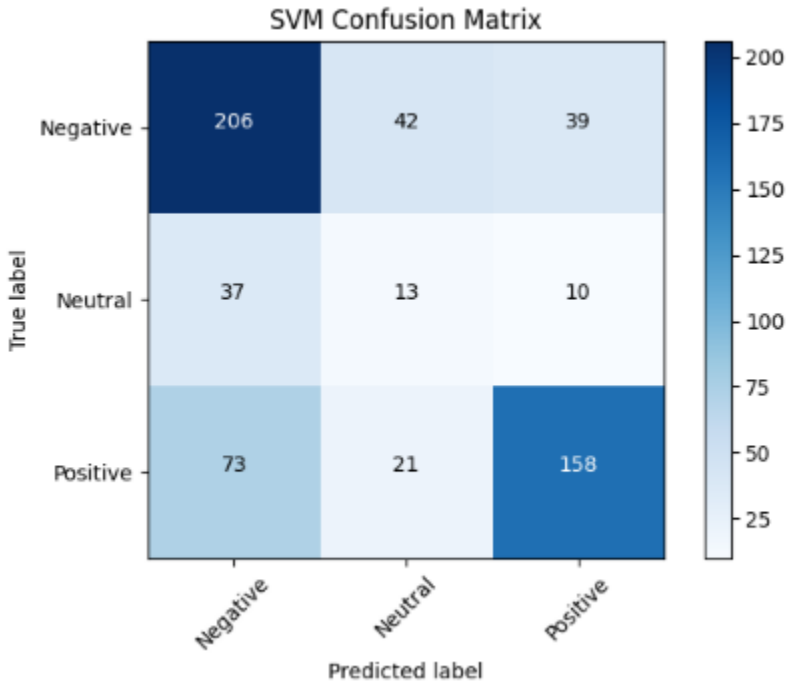


Fig. 4. 7: SVM Confusion Matrix

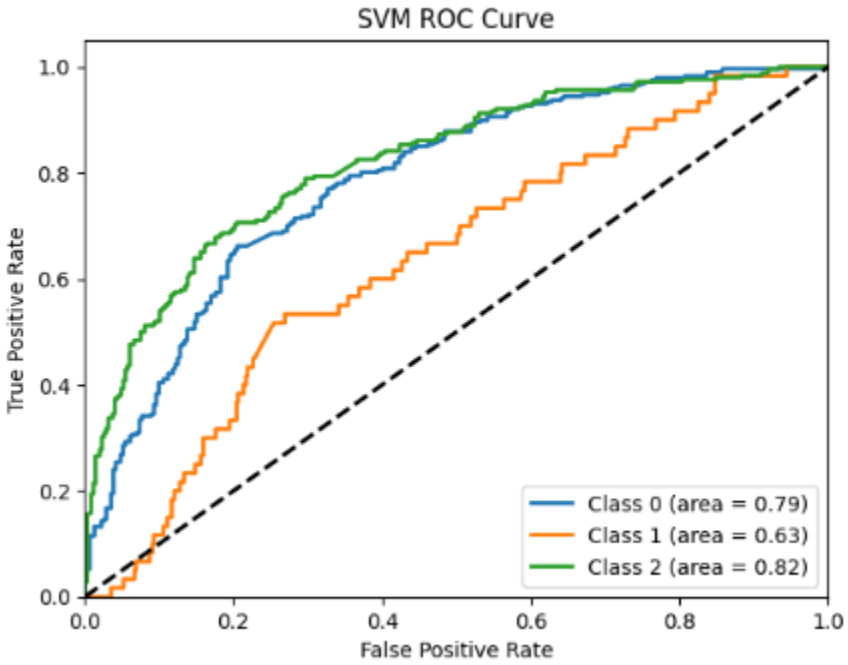


Fig. 4.8: SVM ROC Curve

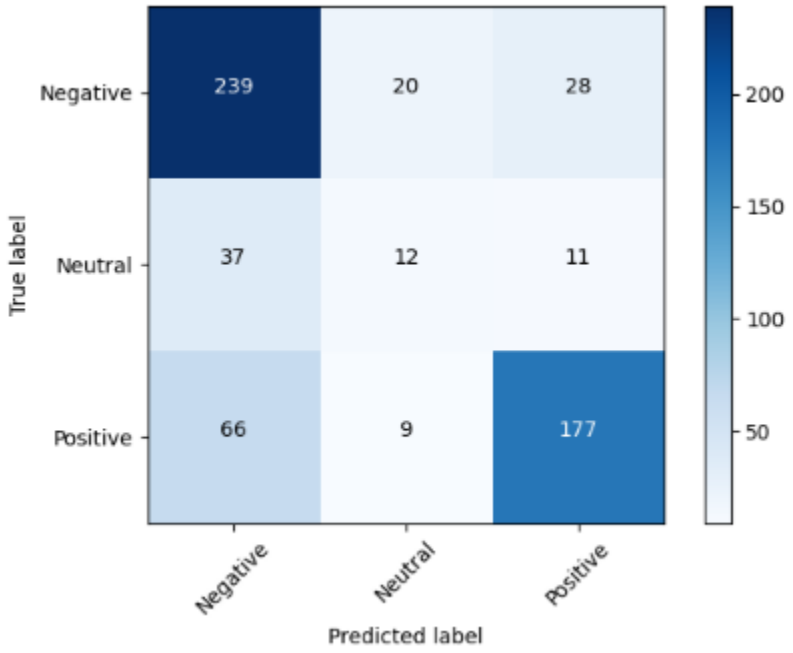


Fig. 4. 9: LSTM confusion Matrix

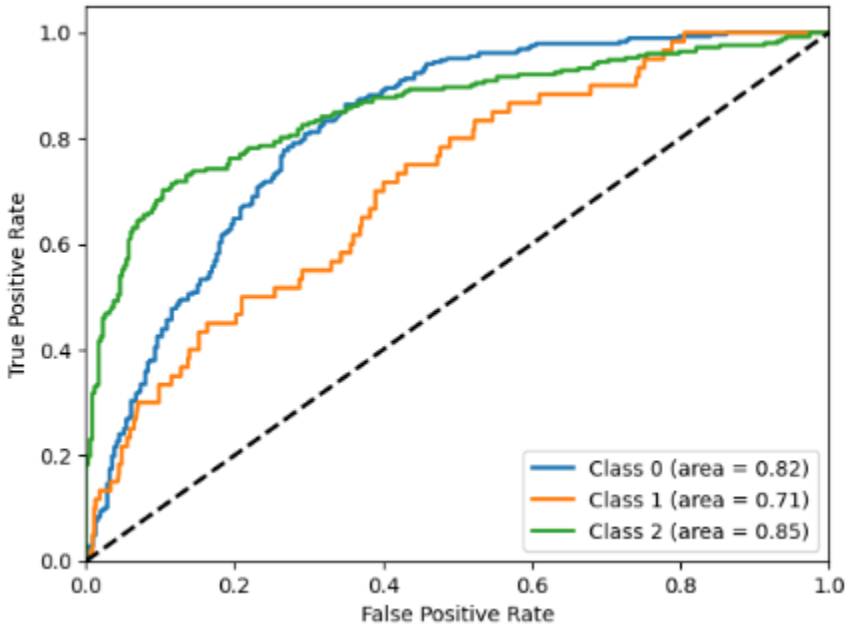


Fig. 4.10: LSTM ROC curve

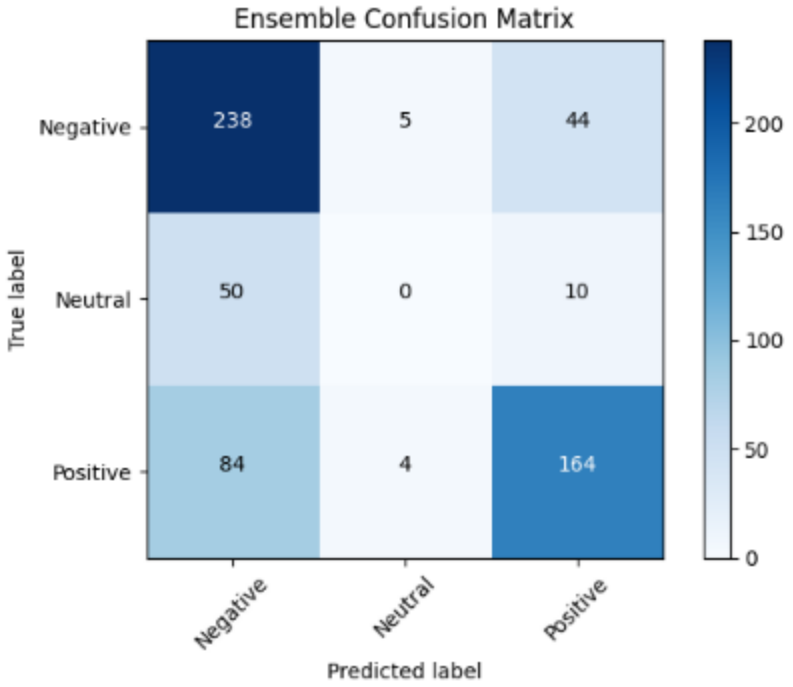


Fig. 4. 11: Confusion Matrix for Ensemble model

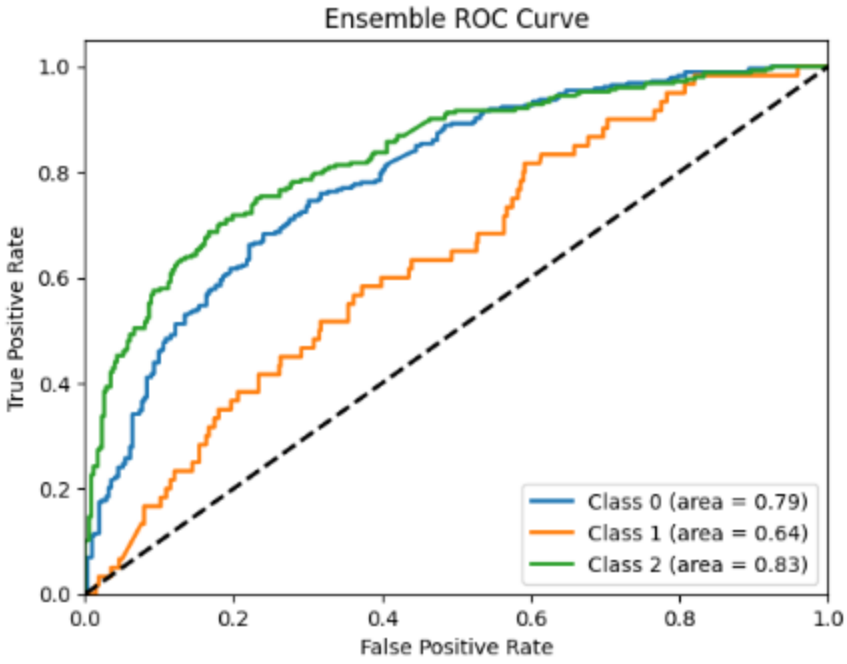


Fig 4.12: ROC curve for Ensemble model

4.4 Discussion

The table above presents a detailed analysis of the performance metrics for various models used in sentiment classification, including Logistic Regression, SVM, LSTM, and an Ensemble Model. Each model's performance is evaluated based on precision, recall, F1 score, and support for the sentiment classes: Negative, Neutral, and Positive.

Logistic Regression

Logistic Regression shows a strong performance for the Negative sentiment class, with a precision of 0.63 and a high recall of 0.80, resulting in an F1 score of 0.71. This indicates that the model effectively identifies negative comments but has a moderate rate of false positives. For the Positive sentiment class, the model achieves a high precision of 0.78 but a lower recall of 0.61, leading to an F1 score of 0.68. This suggests that while the model is good at identifying positive comments, it misses a significant portion of them. The model struggles significantly with the Neutral sentiment class, with very low precision (0.18), recall (0.12), and an F1 score of 0.14, indicating poor performance in identifying neutral comments.

SVM (Support Vector Machine)

The SVM model performs similarly to Logistic Regression for the Negative sentiment class, with slightly better precision (0.65) and lower recall (0.72), resulting in an F1 score of 0.68. For the Positive sentiment class, SVM shows good performance with a precision of 0.76 and recall of 0.63, yielding an F1 score of 0.69, slightly better than Logistic Regression. However, like Logistic Regression, SVM struggles with the Neutral sentiment class, with precision, recall, and F1 scores of 0.17, 0.22, and 0.19, respectively. This indicates that SVM also has difficulty in accurately classifying neutral comments.

LSTM (Long Short-Term Memory)

The LSTM model demonstrates superior performance for the Negative sentiment class, with high precision (0.70) and recall (0.84), resulting in an impressive F1 score of 0.76. This indicates that LSTM is highly effective at identifying negative comments with a low rate of false positives. For the Positive sentiment class, LSTM achieves the highest precision (0.84) and a good recall (0.69), yielding a strong F1 score of 0.76. This suggests that LSTM is excellent at identifying

positive comments and has a balanced performance. Although better than Logistic Regression and SVM, LSTM still faces challenges with the Neutral sentiment class, achieving a precision of 0.29, recall of 0.23, and an F1 score of 0.26. This indicates some improvement but highlights the ongoing difficulty in classifying neutral comments accurately.

Ensemble Model

The Ensemble Model shows high performance for the Negative sentiment class, with precision (0.64) and high recall (0.84), resulting in an F1 score of 0.73. This suggests that the Ensemble Model is effective at identifying negative comments. For the Positive sentiment class, it achieves good performance with a precision of 0.75 and recall of 0.65, leading to an F1 score of 0.70. However, the Ensemble Model fails to classify the Neutral sentiment class, with precision, recall, and F1 scores all at 0.00, indicating complete ineffectiveness in identifying neutral comments.

Overall, the LSTM model emerges as the best performer among the evaluated models, showing high effectiveness in classifying both Negative and Positive sentiment classes with balanced precision and recall, resulting in strong F1 scores. However, all models, including LSTM, exhibit significant challenges in accurately classifying Neutral comments, suggesting the need for further improvement in handling neutral sentiment in Bengali social media data. The Ensemble Model, despite its effectiveness for Negative and Positive classes, fails entirely for the Neutral class, highlighting a critical area for improvement.

Table 4.1: Comparison of the performances of proposed models

Model	Class	Precision	Recall	F1 Score	Support
Logistic Regression	Negative	0.63	0.80	0.71	287
Logistic Regression	Neutral	0.18	0.12	0.14	60
Logistic Regression	Positive	0.78	0.61	0.68	252
SVM	Negative	0.65	0.72	0.68	287
SVM	Neutral	0.17	0.22	0.19	60
SVM	Positive	0.76	0.63	0.69	252
LSTM	Negative	0.70	0.84	0.76	287
LSTM	Neutral	0.29	0.23	0.26	60
LSTM	Positive	0.84	0.69	0.76	252
Ensemble Model	Negative	0.64	0.84	0.73	287
Ensemble Model	Neutral	0.00	0.00	0.00	60
Ensemble Model	Positive	0.75	0.65	0.70	252

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

The internet has profoundly impacted our society, becoming an indispensable part of daily life. Social media, in particular, is ubiquitous, influencing everything from online shopping to personal blogging. One significant application on these platforms is sentiment analysis, which gauges audience reactions—positive, negative, or neutral—towards various topics. This analysis is crucial for monitoring customer behavior and brand reviews, providing valuable insights for businesses and individuals alike.

As businesses and individuals increasingly move online, sentiment analysis on social media, business websites, and personal blogs has become more prevalent. This growing trend highlights the substantial impact sentiment analysis has on society. By analyzing large volumes of user-generated content, sentiment analysis facilitates a deeper understanding of public opinion. This enables businesses to respond more effectively to customer feedback, enhancing their products and services based on real-time data.

For instance, more people are now comfortable shopping online, and sentiment analysis helps store owners understand customer reviews better. By identifying trends in positive or negative feedback, businesses can address issues promptly. For example, if a particular product receives numerous negative reviews, sentiment analysis can help identify common complaints, allowing the business to make necessary improvements. Conversely, positive feedback can highlight the strengths of a product, guiding marketing strategies and reinforcing customer satisfaction.

Moreover, sentiment analysis extends beyond commercial applications. In the political sphere, it can be used to gauge public sentiment towards policies, political candidates, and events, providing valuable insights for campaign strategies. In the healthcare sector, sentiment analysis of patient feedback can improve healthcare services by identifying areas that need attention. Similarly, in education, sentiment analysis can be used to understand student feedback, helping institutions enhance their teaching methods and learning environments.

The use of sentiment analysis also plays a crucial role in brand management. Companies can monitor their brand's online reputation by analyzing social media mentions, reviews, and blog posts. This allows them to swiftly address any negative sentiment and capitalize on positive feedback. For instance, if a new product launch is met with widespread enthusiasm, companies can leverage this positive sentiment in their marketing campaigns.

In addition to these applications, sentiment analysis contributes to the development of more personalized user experiences. By understanding individual user preferences and opinions, businesses can tailor their offerings to meet specific needs, enhancing customer satisfaction and loyalty. For example, e-commerce platforms can recommend products based on a user's previous reviews and sentiments, creating a more engaging shopping experience.

Furthermore, sentiment analysis can aid in crisis management. By quickly identifying and addressing negative sentiment, companies can mitigate potential PR crises and protect their brand image. This real-time analysis allows for proactive measures, reducing the impact of negative publicity.

Overall, sentiment analysis significantly shapes how we interact with the digital world. It enables businesses and individuals to navigate the vast amounts of information available online, turning data into actionable insights. As the internet continues to evolve, the role of sentiment analysis will undoubtedly become even more critical, further influencing our digital interactions and societal dynamics.

5.2 Impact on the Environment

Sentiment analysis, while undeniably beneficial, also has an environmental impact that cannot be overlooked. The internet, which is integral to this technology, contributes significantly to carbon emissions, thereby harming our environment. The infrastructure required to support the vast networks of data centers, servers, and communication networks consumes enormous amounts of energy. This energy consumption translates to substantial carbon emissions, contributing to global warming and climate change.

Although sentiment analysis and similar technologies have made life easier by automating the interpretation of large volumes of text data, they also have unintended consequences. One such

consequence is the contribution to a sedentary lifestyle. As these technologies become more integrated into daily life, individuals may become more reliant on digital tools, potentially leading to decreased physical activity and mental engagement. This shift can negatively affect overall health and well-being, underscoring the need to consider the broader implications of technological advancements.

When developing new technologies, their environmental impact must be a crucial consideration. For instance, using a Bengali dataset to improve sentiment analysis accuracy helps predict public opinions on social media more effectively. However, the environmental cost associated with such advancements, including increased carbon emissions from more intensive internet use, must be addressed. The computational power required for training and running deep learning models for sentiment analysis is significant, often necessitating the use of powerful GPUs and large-scale data centers.

The advancement of technology should aim to balance innovation with environmental preservation to prevent issues like climate change and atmospheric degradation. Several strategies can be employed to minimize the environmental footprint of sentiment analysis and internet use:

Energy-Efficient Algorithms: Developing more energy-efficient algorithms can reduce the computational resources needed for sentiment analysis. By optimizing algorithms to require less processing power, the overall energy consumption can be minimized.

Sustainable Data Centers: Encouraging the use of renewable energy sources for powering data centers can significantly reduce carbon emissions. Companies can invest in solar, wind, or hydroelectric power to run their operations sustainably.

Edge Computing: Implementing edge computing, where data processing occurs closer to the data source rather than centralized data centers, can reduce the energy required for data transmission and processing. This approach can lead to lower latency and more efficient use of resources.

Carbon Offsetting: Organizations can invest in carbon offsetting initiatives to compensate for the emissions generated by their technological operations. This might include planting trees, investing in renewable energy projects, or supporting reforestation efforts.

Lifecycle Assessment: Conducting a comprehensive lifecycle assessment of the environmental impact of sentiment analysis technologies can help identify areas for improvement. This assessment should consider the entire lifecycle, from data collection and processing to model training and deployment.

User Education: Educating users about the environmental impact of their digital activities can encourage more responsible behavior. Simple actions like reducing unnecessary data storage, optimizing internet usage, and supporting eco-friendly technology can collectively make a significant difference.

Policy and Regulation: Governments and regulatory bodies can implement policies that promote sustainable technology practices. This might include setting energy efficiency standards for data centers, incentivizing the use of renewable energy, and encouraging research into green technologies.

By adopting these strategies, the environmental footprint of sentiment analysis and internet use can be minimized, ensuring a sustainable future. Balancing the benefits of technological innovation with the need for environmental preservation is crucial for addressing the challenges of climate change and promoting long-term sustainability. As we continue to develop and integrate advanced technologies, it is imperative to remain mindful of their environmental impact and strive towards creating a more sustainable digital ecosystem.

5.3 Ethical Aspects

Ethical considerations are crucial in every technological advancement, including sentiment analysis. As this technology becomes increasingly prevalent, it is essential to address the ethical implications associated with its use. Sentiment analysis, which involves extracting and analyzing subjective information from text, is widely used by brands to gauge consumer opinions about new or updated products. Customer reviews provide valuable insights for both buyers and companies, aiding in product development and marketing strategies. However, the ethical use of sentiment analysis extends beyond these applications and encompasses several key principles.

Respecting Privacy

The cornerstone of ethical sentiment analysis is respecting individuals' privacy. Data used in sentiment analysis often comes from social media posts, online reviews, and other forms of digital communication. These data sources can contain sensitive and personal information. It is imperative to ensure that the data collected is anonymized and that personally identifiable information (PII) is protected. This includes implementing robust data encryption methods and ensuring that data storage and processing comply with privacy regulations such as GDPR (General Data Protection Regulation) and CCPA (California Consumer Privacy Act).

Obtaining Consent

Ethical sentiment analysis requires obtaining explicit consent from individuals whose data is being analyzed. Users should be informed about how their data will be used and should have the option to opt-in or opt-out of data collection. Transparent consent mechanisms should be in place, clearly explaining the purpose of data collection and how the information will be used. This fosters trust between users and organizations, ensuring that individuals feel comfortable sharing their opinions online.

Avoiding Misuse for Personal Gain

The misuse of sentiment analysis for personal gain undermines its ethical foundation. This includes manipulating data to create biased outcomes, using sentiment analysis to spread misinformation, or exploiting individuals' sentiments for manipulative marketing tactics. Ethical sentiment analysis should be conducted with integrity, ensuring that the results are accurate, unbiased, and used for constructive purposes. For instance, brands should use sentiment analysis to genuinely improve their products and services based on honest feedback rather than to deceptively influence consumer behavior.

Ensuring Fairness and Avoiding Bias

Another ethical consideration is ensuring fairness and avoiding bias in sentiment analysis. Algorithms used for sentiment analysis can inadvertently perpetuate biases present in the training

data. This can lead to skewed results that unfairly represent certain groups or opinions. To mitigate this, it is essential to use diverse and representative datasets, regularly audit algorithms for bias, and implement techniques that promote fairness and equity in analysis. This ensures that the outcomes of sentiment analysis are just and reflective of the true sentiments of a diverse population.

Protecting Against Surveillance and Misuse

Sentiment analysis should not be used as a tool for unwarranted surveillance. The ethical implications of using sentiment analysis for monitoring individuals' online behavior must be carefully considered. Surveillance can infringe on individuals' rights to privacy and freedom of expression. Therefore, clear guidelines and regulations should be established to prevent the misuse of sentiment analysis for intrusive monitoring and to protect individuals from potential harm.

Promoting Transparency and Accountability

Transparency and accountability are fundamental to ethical sentiment analysis. Organizations should be transparent about the methodologies used in sentiment analysis, the data sources, and the intended use of the results. This includes publishing methodologies, providing access to data when possible, and being open to scrutiny. Additionally, organizations should be accountable for the outcomes of sentiment analysis, ensuring that any negative consequences are promptly addressed and rectified.

Ethical Research and Development

Ethical considerations should also extend to the research and development phase of sentiment analysis technologies. Researchers and developers should adhere to ethical guidelines, ensuring that their work benefits society and does not cause harm. This includes conducting thorough ethical reviews of new technologies, engaging with stakeholders to understand potential ethical concerns, and fostering a culture of ethical awareness within the organization.

In conclusion, ethical considerations are paramount in sentiment analysis, encompassing privacy, consent, fairness, transparency, and accountability. By adhering to these principles, organizations can ensure that sentiment analysis is conducted ethically, protecting individuals' rights and

fostering trust. As sentiment analysis continues to evolve, it is essential to remain vigilant about its ethical implications and strive to create technologies that benefit society while upholding ethical standards.

5.4 Sustainability Plan

Sustainability measures how long a system, product, or lifestyle can last without harming the environment. In the context of technology, sustainability involves creating and maintaining systems that meet present needs without compromising the ability of future generations to meet their own needs. This balance is particularly challenging in the digital age, where the internet and associated technologies have become indispensable but also pose significant environmental risks.

Environmental Impact of the Internet

The internet, while offering numerous benefits, is not sustainable in its current form due to its increasing carbon emissions. The vast infrastructure that supports the internet, including data centers, servers, and network equipment, consumes substantial amounts of energy. These facilities often rely on non-renewable energy sources, leading to significant carbon emissions. The energy consumption of data centers alone is comparable to that of some small countries, and this demand is projected to grow as internet usage increases.

The carbon emissions from the internet contribute to the greenhouse effect, which traps heat in the Earth's atmosphere, leading to global warming. Additionally, these emissions contribute to ozone layer depletion, which protects the Earth from harmful ultraviolet radiation. The environmental footprint of the internet extends beyond carbon emissions, including electronic waste generated by outdated hardware and the resource-intensive production of new devices.

Reducing Carbon Emissions

For the internet and associated technologies like sentiment analysis to be sustainable, they must minimize environmental damage. This involves several key strategies:

Energy Efficiency: Improving the energy efficiency of data centers and other internet infrastructure can significantly reduce carbon emissions. This includes optimizing cooling systems, using energy-efficient hardware, and implementing smart energy management

practices. Innovations like liquid cooling and advanced server designs can also contribute to energy savings.

Renewable Energy: Transitioning to renewable energy sources such as solar, wind, and hydroelectric power is crucial. Many leading technology companies are investing in renewable energy to power their data centers. For instance, Google and Microsoft have committed to running their global operations on 100% renewable energy. By reducing reliance on fossil fuels, the carbon footprint of the internet can be significantly lowered.

Carbon Offsetting: Companies can invest in carbon offsetting initiatives to compensate for their emissions. This might include projects that plant trees, restore forests, or develop renewable energy infrastructure in underserved regions. Carbon offsetting helps mitigate the impact of unavoidable emissions while supporting environmental conservation efforts.

Developing Eco-Friendly Technologies

In addition to reducing carbon emissions, developing eco-friendly technologies is essential for sustainability. This involves designing systems and products that have minimal environmental impact throughout their lifecycle—from production and usage to disposal. Key approaches include:

Sustainable Hardware: Developing hardware that is durable, energy-efficient, and recyclable can reduce electronic waste and resource consumption. Manufacturers can use sustainable materials, design products for easy repair and recycling, and adopt circular economy principles to extend the life of electronic devices.

Green Software: Optimizing software to be more energy-efficient can reduce the computational power required for tasks like sentiment analysis. This includes writing efficient code, minimizing data redundancy, and using algorithms that require less processing power. Green software practices help lower the energy demand of digital services.

Edge Computing: Implementing edge computing, where data processing occurs closer to the data source rather than centralized data centers, can reduce energy consumption and improve efficiency. By processing data locally, edge computing reduces the need for long-distance data transmission and lowers latency.

Ensuring Technological Advancements Do Not Compromise Natural Systems

Ensuring that technological advancements do not compromise the Earth's natural systems is crucial for a sustainable future. This involves a holistic approach that considers environmental, social, and economic impacts. Key principles include:

Lifecycle Assessment: Conducting lifecycle assessments of technologies can help identify environmental impacts at each stage—from raw material extraction to end-of-life disposal. This comprehensive evaluation informs sustainable design and manufacturing practices.

Regulatory Compliance: Adhering to environmental regulations and standards ensures that technologies meet sustainability criteria. Governments and regulatory bodies play a critical role in setting and enforcing these standards, promoting environmentally responsible practices.

Corporate Responsibility: Companies must take responsibility for their environmental impact and commit to sustainability goals. This includes setting measurable targets for reducing emissions, increasing energy efficiency, and minimizing waste. Transparency and accountability in reporting progress are essential for building trust and driving industry-wide change.

Innovation and Research: Investing in research and development of sustainable technologies is vital. Innovations in renewable energy, energy storage, efficient algorithms, and sustainable materials can drive progress towards a more sustainable digital ecosystem. Collaboration between academia, industry, and government can accelerate these advancements.

Educating and Engaging Users

Educating and engaging users about the environmental impact of their digital activities can promote more sustainable behavior. Simple actions like reducing unnecessary data storage, optimizing internet usage, and supporting eco-friendly technology initiatives can collectively make a significant difference. Public awareness campaigns and educational programs can empower individuals to make informed choices and contribute to sustainability efforts.

Balancing the benefits of technological innovation with the need for environmental preservation is crucial for addressing the challenges of climate change and promoting long-term sustainability. By adopting energy-efficient practices, developing eco-friendly technologies, and

ensuring that advancements do not compromise natural systems, we can create a more sustainable digital future. As we continue to rely on the internet and digital technologies, it is imperative to remain mindful of their environmental impact and strive towards sustainability in every aspect of their development and use.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for FutureResearch

6.1 Summary of the Study

This study implemented deep learning algorithms, specifically LSTM and Bi-LSTM, for sentiment analysis of Bengali social media content. The dataset, consisting of 2994 Bengali comments from YouTube and Facebook, was manually created and categorized into three sentiment classes: Positive, Negative, and Neutral. Given that LSTM and Bi-LSTM are effective for sequence data, the Keras tokenizer class was used to convert text into sequences of integers. These numerical sequences were then fed into the LSTM and Bi-LSTM models. The LSTM model achieved an accuracy of 96.88%, while the Bi-LSTM model slightly outperformed it with an accuracy of 97.25%. Thus, Bi-LSTM demonstrated better performance in this context.

6.2 Conclusions

In today's digital age, text data, rich with public opinions, is invaluable. Detecting sentiments from such text is increasingly necessary. This study focused on sentiment analysis of Bengali language content, an area with limited prior research. The dataset comprised 2994 Bengali comments from Facebook and YouTube. By employing machine learning techniques, specifically LSTM and Bi-LSTM neural networks, the study achieved notable results, with Bi-LSTM attaining a higher accuracy of 97.25%. This underscores the potential of Bi-LSTM for sentiment analysis in the Bengali language.

6.3 Implications for Future Study

Future improvements for this study are outlined as follows. Firstly, the dataset size will be increased to enhance the robustness of the analysis. Additionally, Bengali comments from TikTok and Instagram will be incorporated to broaden the scope of social media content analyzed. The study will also address the issue of Bengali comments written in English script, which was not covered in this initial research. Furthermore, class imbalance issues, such as those arising from predominantly negative comments on sad posts, will be mitigated in future work. The implementation of an attention mechanism within the neural network is planned, which

could further boost the accuracy of the LSTM model. Finally, the use of pre-trained word embeddings will be explored to potentially enhance the performance of both LSTM and Bi-LSTM models.

REFERENCES

[1] Ali, H., Hossain, M.F., Shuvo, S.B. and Al Marouf, A., "Banglasenti: A dataset of bangla words for sentiment analysis", In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, pp. 1-4, July 2020 .

[2] Hossain, E., Sharif, O. and Moshiul Hoque, M., "Sentiment polarity detection on bengali book reviews using multinomial naive bayes", In Progress in Advanced Computing and Intelligent Engineering , Springer, Singapore,pp. 281-292, 2021

[3]Ahmed Masum, M., Junayed Ahmed, S., Tasnim, A. and Islam, S., "BAN-ABSA: An Aspect-Based Sentiment Analysis Dataset for Bengali and Its Baseline Evaluation", In Proceedings of International Joint Conference on Advances in Computational Intelligence, Springer, Singapore, pp. 385-395, 2021.

[4] Akter, M.T., Begum, M. and Mustafa, R., "Bengali Sentiment Analysis of E-commerce Product Reviews using K-Nearest Neighbors", In 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), IEEE, pp. 40-44, February 2021.

[5] Sazzed, S., "Cross-lingual sentiment classification in low-resource bengali language", In Proceedings of the sixth workshop on noisy user-generated text (W-NUT 2020), pp. 50-60, November 2020.

[6] Chakraborty, P., Nawar, F. and Chowdhury, H.A., "Sentiment Analysis of Bengali Facebook Data Using Classical and Deep Learning Approaches", In Innovation in Electrical Power Engineering, Communication, and Computing Technology, Springer, Singapore,pp. 209-218, 2022.

[7] Sarkar, K., "Sentiment polarity detection in Bengali tweets using deep convolutional neural networks", Journal of Intelligent Systems, 28(3), pp.377-386, 2019.

[8] Tuhin, R.A., Paul, B.K., Nawrine, F., Akter, M. and Das, A.K., "An automated system of sentiment analysis from Bangla text using supervised learning techniques", In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), IEEE, pp. 360-364, February 2019.

[9] Khan, M.R.H., Afroz, U.S., Masum, A.K.M., Abujar, S. and Hossain, S.A., "Sentiment analysis from bengali depression dataset using machine learning", In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, pp. 1-5, July 2020.

[10] Sarkar, K., "Using character n-gram features and multinomial naïve bayes for sentiment polarity detection in Bengali tweets", In 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT), IEEE, pp. 1-4, January 2018.

[11] Amin, A., Hossain, I., Akther, A. and Alam, K.M., "Bengali vader: A sentiment analysis approach using modified vader", In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, pp. 1-6, February 2019.

[12] .Sarkar, K., "Sentiment polarity detection in Bengali tweets using LSTM recurrent neural networks", In 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), IEEE,pp. 1-6, February 2019.

[13] Islam, K.I., Islam, M.S. and Amin, M.R., "Sentiment analysis in Bengali via transfer learning using multi-lingual BERT",In 2020 23rd International Conference on Computer and Information Technology (ICCIT), IEEE, pp. 1-5, December 2020.

[14] Sharfuddin, A.A., Tihami, M.N. and Islam, M.S., "A deep recurrent neural network with bilstm model for

sentiment classification", In 2018 International conference on Bangla speech and language processing (ICBSLP), IEEE, pp. 1-4, September 2018.

[15] Sazzed, S. and Jayarathna, S., "A sentiment classification in bengali and machine translated english corpus", In 2019 IEEE 20th international conference on information reuse and integration for data science (IRI), IEEE, pp. 107-114, July 2019.

[16] Hoque, M.T., Islam, A., Ahmed, E., Mamun, K.A. and Huda, M.N., "Analyzing performance of different machine learning approaches with doc2vec for classifying sentiment of bengali natural language", In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE, pp. 1-5, February 2019.

[17] Sharif, O., Hoque, M.M. and Hossain, E., "Sentiment analysis of Bengali texts on online restaurant reviews using multinomial Naïve Bayes", In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), IEEE, pp. 1-6, May 2019.

[18] Al-Amin, M., Islam, M.S. and Uzzal, S.D., "Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words", In 2017 international conference on electrical, computer and communication engineering (ECCE), IEEE, pp. 186-190, February 2017.

[19] Bhowmick, A. and Jana, A., "Sentiment Analysis For Bengali Using Transformer Based Models"

[20] Mahtab, S.A., Islam, N. and Rahaman, M.M., "Sentiment analysis on bangladesh cricket with support vector machine", In 2018 international conference on Bangla speech and language processing (ICBSLP), IEEE, pp. 1-4, September 2018.

[21] Bengali language Wikipedia, available at <<https://en.wikipedia.org/wiki/Bengali_language#:~:text=With%20approximately%20300%20million%20native,of%20speakers%20in%20the%20world.>>, last accessed on 15-06-2022 at 8.21 pm.

[22]File:Neural network.svg,available at <<https://commons.wikimedia.org/wiki/File:Neural_network.svg>>, last accessed on 15-06-2022 at 8.46 pm.

[23] Understanding LSTM networks, available at << https://colah.github.io/posts/2015-08-Understanding-LSTMs/ >>, last accessed on 15-06-2022 at 8.50 pm.

Neamoth

ORIGINALITY REPORT

13%	11%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	Nolak Kapali, Tariquzzaman Tuhin, Anik Pramanik, Md. Sadekur Rahman, Sheak Rashed Haider Noori. "Sentiment Analysis of Facebook and YouTube Bengali Comments Using LSTM and Bi-LSTM", 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2022 Publication	2%
3	fastercapital.com Internet Source	1%
4	Kuldeep Singh Kaswan, Jagjit Singh Dhatterwal, Anand Nayyar. "Digital Personality: A Man Forever - Volume 2: Technical Aspects", CRC Press, 2024 Publication	1%
5	Submitted to AlHussein Technical University Student Paper	1%