# SENTIMENT ANALYSIS OF SOCIAL MEDIA'S BANGLA DATA

**BY**

**TARIQUZZAMAN TUHIN**
**ID: 183-15-11912**
**AND**

**NOLAK KAPALI**
**ID: 183-15-12029**

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

Supervised By

**Dr. Sheak Rashed Haider Noori**
Associate Professor and Associate Head
Department of CSE
Daffodil International University

Co-Supervised By

**Mr. Abdus Sattar**
Assistant Professor and Coordinator
Department of CSE
Daffodil International University

**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**12 SEPTEMBER 2022**

# APPROVAL

This Project titled Sentiment Analysis of Social Media’s Bangla Data**,** submitted by Tariquzzaman Tuhin and Nolak Kapali 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 12 September 2022.

## BOARD OF EXAMINERS

14/9/22

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

**Dr. Md. Monzur Morshed** **Internal Examiner**
**Professor**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

**Ms. Samia Nawshin** **Internal Examiner**
**Assistant Professor**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Farid

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

# DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

**Supervised by:**

11.09.2022

**Dr. Sheak Rashed Haider Noori**
Associate Professor and Associate Head
Department of CSE
Daffodil International University

**Co-Supervised by:**

Sattar 11/09/22

**Mr. Abdus Sattar**
Assistant Professor and Coordinator
Department of CSE
Daffodil International University

**Submitted by:**

Tuhin

**Tariquzzaman Tuhin**
ID: -183-15-11912
Department of CSE
Daffodil International University

Nolak Kapali

**Nolak Kapali**
ID: -183-15-12029
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 **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head**, 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.Touhid Bhuiyan, Professor and Head, Department of CSE**, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University. We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work. Finally, we must acknowledge with due respect the constant support and patients of our parents.

# ABSTRACT

Sentiment analysis is a way where huge amount of data can be categorized into different sentiments, emotions, attitudes or opinions. Using sentiment analysis sentiments or emotions can be predicted from individual's viewpoint which means individual, and public are expressing their views and from using these views their sentiments can be detected using sentiment analysis. Recently sentiment analysis has been performed on various information of social media to derive market intelligence. As we know social media is filled with different contents and audiences are interacting there making it a huge opportunity to perform sentiment analysis on this information. In terms of Bengali contents, many audiences there interact with Bengali language which makes it a treasure trove to perform sentiment analysis in Bengali NLP field. Here in this study sentiment analysis has performed on the audience's Bengali comments expressing different views of regarding social media's Bengali contents. The dataset containing these public views are 4000 Bengali comments collected from Facebook and YouTube Bengali contents. Here positive, negative and neutral classes are used to categorize the Bengali data and tokenizer from Keras library is used to tokenize the Bengali text. Deep learning is part of that machine learning algorithm where its structures are inspired by the human brain. In this study deep learning algorithm LSTM and Bi-LSTM are used where Keras library is used to run these algorithms. Deep learning algorithm LSTM and Bi-LSTM are performed, and Bi-Directional LSTM has the highest accuracy of 97.25% than LSTM.

# TABLE OF CONTENTS

**CONTENTS** **PAGE**

# 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 which extracts the sentiments, emotions, opinions, attitudes of a data [1]. The data that has been used in this analysis bears the sentiments of the individual thinking on certain topics or products. For deeper understanding, this data that can be any image or video or audio data or textual data and the sentiments are extracted from this data are basically the audience viewpoint of any individual topics or products. The individual viewpoint is categorized into sentiments, emotions, opinions, attitudes etc. In terms of sentiments, they are labelled into three category positive, negative and neutral calling it multi-modal and some have divided these three sentiments into more depth adding another two sentiments so that people viewpoint can be analyzed more precisely and it called fine-grained sentiment and these sentiments are positive, more positive, neutral, negative, more negative and when sentiment analysis performs on only positive and negative sentiments it is called binary sentiments [1]. Sentiment analysis is a classification problem as prediction are given by using different classifier where data given as input to it and it has polarity score which determines output sentiments which are positive, negative and neutral of a specific text [2]. Emotions are complex states with different stages and has a huge range of variations and can categorized into happy, sad, anger, sad etc. [3] [4] [5]. Attitudes are the way to treat any particular topic or product etc. where opinions are the raw views of it. The data that are used on various kinds of sentiment analysis are mostly textual data collected from social media posts, blogs etc. Also, some sentiment analysis has performed on websites that generated huge amount user views [6] but social media data use is seen mostly. Social media content is the source of collecting the data as huge amount people share their point of views on these contents as comments. Social media content is everything that is shared by any individuals or groups of people to interact with the audience. Many individuals such as content creators or influencer and many groups or pages such as business page or entertainment page are all giving contents on social media and for Bengali language Bengali contents are also growing in social media and huge Bengali comments are also found expressing their views to these contents. Contents can be a blog or advertisements or sharing reviews of foods, movies, clothes etc. and sentiment analysis can be performed on these Bengali comments expressing their views to the contents. To perform sentiment analysis there are three approaches. One is lexicon based another one is automated or machine learning based and another one is hybrid. In the lexicon-based approach it performs on the dictionary or in the corpus where word is already categorized according to the sentiments and compares the preprocessed text with it and give result about the polarity of different text [7]. In automated approach or machine learning approach machine learning algorithm performed sentiment analysis. It can be supervised and unsupervised. In supervised learning the datasets are labelled data and machine learning algorithm can be trained using this dataset. In unsupervised learning the dataset is not labelled and machine learning algorithm has to perform using this dataset. Support Vector Machines, Naive Bayes etc. are the example of these algorithms. There is another approach called hybrid approach which is the combination of lexical and machine learning based approach. Deep learning is a part of machine learning. Deep learning is a sector of machine learning where all the algorithms are built from getting inspiration from the human brain neurons. CNN, RNN, LSTM, Bi-LSTM are the example of the deep learning algorithms. This deep learning algorithm are called neural networks which was inspired by the human brain and consists of neurons or nodes and has layered structures. In these paper deep learning algorithm LSTM and Bi-LSTM has been performed to see how they work on the Bengali dataset of social media content's comments from Facebook and YouTube for sentiment analysis.

## 1.2 Motivation

Bengali is the fifth most spoken native and seventh most spoken language in the world. It is the sweetest language in the world according to UNESO [21]. And NLP stands for Natural Language Processing which is the field of machines connecting to human languages and many researchers have done many works in different human language. As sentiment analysis is mostly perform on textual data many researchers have done work in sentiment analysis in various languages. English is one of the top one among them as it is an international language. In the world filled with so many different languages Bengali language is falling behind in the work of sentiment analysis. The work done in this language is not sufficient because of the complexity and poor resources [1] and this is the reason this research has focused on Bengali language to perform sentiment analysis. Bengali is a rich language filled with emotions and the world recognized its literature values. But in research field such as sentiment analysis the usage of the language is not sufficient. As of in today’s world social media is blasting with huge data and information and demand of contents to gain viewers in social media causing users to increase contents in social media and in terms of Bengali language users such as influencers or Bengali content creators or Bengali brands are filling up social media with content and that’s why number of Bengali comments reacting to these contents are also increasing but very few work has been done gathering the social media Bengali data in sentiment analysis. Also, Bengali contents are rising in social media and people eagerness towards it are rising and many brands seeing this an opportunity and making social media as their advertisement platform to promote their brand. So, there are increasing Bengali content in social media which makes increasing comments of public towards these contents especially in Facebook and YouTube and so performing sentiment analysis will be very impactful in this field and also few researchers have done work in that field so we are motivated by how big opportunity is there in the field to work with all those data.  So mainly we are motivated to work in this field dues to these reasons are given below:

- Lack of Research work in sentiment analysis using Bengali language.
- Lack of Bengali publicly available dataset or resources.
- Lack of work in Bengali contents in social media.

To enrich more in this field, we are motivated to do the sentiment analysis on Bengali language social media data from Facebook and YouTube.

## 1.3 Rationale of the Study

Social media is full of data and information, sentiment analysis has been performing on this vast information to know the opinions behind each and every people. We have performed sentiment analysis on social media’s data and extracted sentiments behind it. The dataset that we have worked on is a Bengali language dataset. As social media’s such as Facebook, YouTube etc. usage is increasing day by day, number of contents on these platforms are growing larger and larger and for that reason Bengali users’ comments filling expressing their reaction towards the contents are increasing. In Bengali contents, performed by influencers and content creators where their main work is to create content and not only influencers but also business and entertainment brands also producing contents to promote their brands through collaborating with social media influencers etc. and these all contents including promotion of products to cooking channel videos all are falling into social media contents and these creating dataset like product review and cooking review  are actually all the comments of public reaction toward the content and in this study, we are performing sentiment analysis using these social media content’s comments. The reason for doing this work is to generate the insider views of public regarding the contents and also to create a publicly available Bengali dataset. The scarcity of Bengali dataset is high as we face this during our study. Also, the reason behind of doing this work is to increase the sentiment analysis work on Bengali language and all of them are given below:

- To generate the insider views of public regarding the contents
- To create a publicly available Bengali dataset.
- To build a better deep learning model for sentiment analysis work on Bengali language

## 1.4 Research Questions

- How sentiments, emotions, attitudes and opinions are different?
- What is sentiment analysis?
- What is social media content?
- Why use tokenization using Keras library?
- How does embedding layer work?

## 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:

- A deep learning model that classifies sentiments from Bengali text.
- A publicly available dataset of 4000 Bengali comments with labelled 3 class sentiments.

## 1.6 Report Layout

In the layout section we will discuss the specific task each of the section does. Here the introduction section discusses about the inspiration, objective, outcome of the work, chapter 2 discusses the related research that has been done in this area, chapter 3 discusses the methodology or the process to build the proposed model and chapter 4 discusses the result of that model, chapter 5 talks about the effect of this work in society, environment and chapter 6 talks about the conclusion and future improvements about the study.

# CHAPTER 2

# Background Studies

## 2.1 Terminologies

Some basic terminologies are given below:

**Supervised learning:** In this learning, machine learns from labelled dataset that used for training. This labelled data is what make big difference to this learning comparing with unsupervised learning.

**Unsupervised learning:** The data here are unlabeled raw data and machine learns the pattern from this unlabeled data in the time of training phase.

**Machine learning algorithm:** Machine learning algorithm helps the machine to carry the task basically this algorithm takes the machine from the input to the desired result. Support Vector Machines (SVM), Decision Tress (DT), Naive Bayes (NB), Random Forest (RF) etc. are some popular used algorithms in machine learning.

**Neural Network:** Neural network is a collection of algorithms that is inspired by the human brain. Neural network is made of nodes/neurons that are interconnected with each other and it has a layered architecture. This neural network also called Artificial Neural Network. Normally a neural network has three layers. First one is an input layer, second one is a hidden layer and third one is an output layer. In the following figure 2.1.1: Neural network architecture, we can see the neurons are connected with each other and here the first two neurons are the input layer, then the five neurons are the hidden layer and the one neuron is the output layer [22]. Here it means 3 inputs will pass to 3 neurons and pass it along in the hidden layer and generate one output in the output layer. CNN, RNN, LSTM, Bi-LSTM all are the example of neural network.

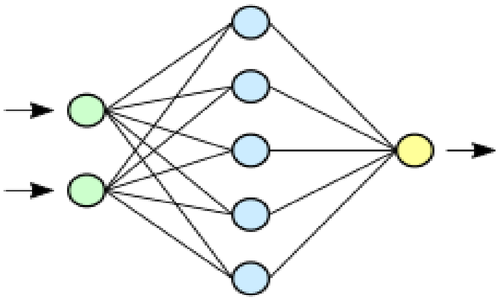


Figure 2.1: Neural network architecture

**Recurrent Neural Network:** It is an algorithm that works on sequential data or data that is in ordered series and where all the data are related to each other in that series. This network has an internal memory which remembers the previous output of the data which is very important to work on sequential data. And that is main difference of this network to other neural networks. In the following figure 2.1.2 Recurrent Neural Network, input X is passed into hidden layers A and output will be h [23]. And it the output will be cycled to the hidden layers for t time stamps. Here if fold the images then $X_0$ will be passed into hidden layers and then output $h_0$ will be generated and then this $h_0$ output will be the new input $X_1$ and will go to the hidden layer and generated another. The following figure 2.1.2: RNN architecture, input X is passed into hidden layers A and output will be h. And it the output will be cycled to the hidden layers for t time stamps. Here if fold the images then $X_0$ will be passed into hidden layers and then output $h_0$ will be generated and then this $h_0$ output will be the new input $X_1$ and will go to the hidden layer and generated another output $h_1$ and this output will again pass into hidden layer thus it will continue to t time stamps and generate final output $h_t$. So, this network remembers the past output and the put it in the present output and for this memory it is very useful to work with the sequence data.

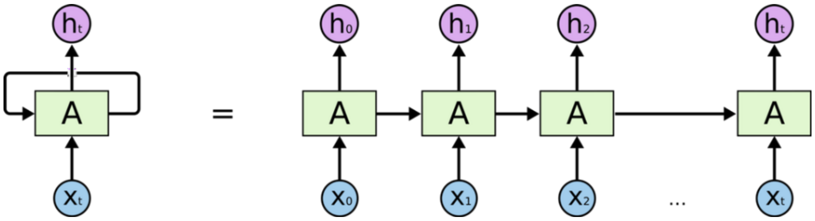


Figure 2.2:  RNN architecture

**LSTM:** LSTM is a recurrent neural network or RNN which works like a recurrent neural network but recurrent neural network has a short-term memory, LSTM has a long-term memory and that is the key difference of the RNN and LSTM network. RNN cannot work in long term dependencies which means it cannot remember past output that happened certain time ago but it can remember recent past output and to solve those issues new recurrent neural network named LSTM is introduced in the deep learning world. The following figure shows RNN has a single neural network layer but in LSTM there are 4 neural networks layers and each of them are interacting with each other. The figure 2.1.3 RNN architecture 2 shows another RNN architecture which has one neural network but in figure 3.3.4 LSTM network architecture shows 4 neural networks in its architecture [23].

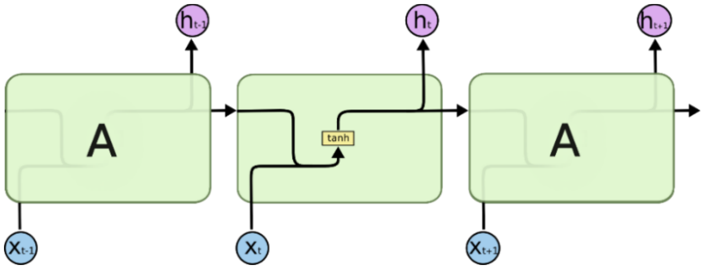


Figure 2.3: RNN architecture 2

LSTM has a state called cell state and this cell looks like a chain or a conveyor belt and it has gates and these gates control the flow of information that which information should be passed and which should be forgotten. There are 3 types of gates. Forget gate decide which information should be forgotten, Input gate decide which information should be passed or added or updated and output gate decides which information should be the output. The four neural network layer passes information to the gates and these gates passes information to the cell states. The gates play a crucial role here because by its which information is more important is added dropping the less important information. So, for these LSTM knows which information is more important and that is important for sequential data.

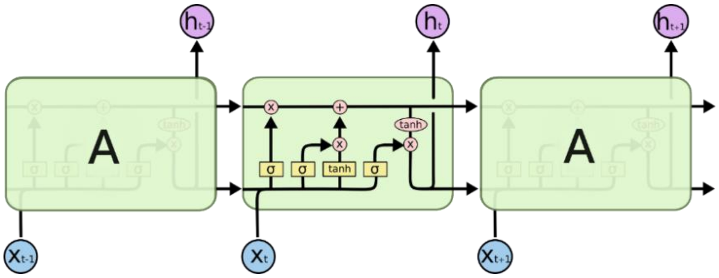


Figure 2.4: LSTM network architecture

**Bi-LSTM:** Bi-LSTM is a recurrent neural network that has two LSTM layers. Normally in LSTM network has single LSTM layer in the architecture but for Bi-LSTM it has two LSTM network layers in its architecture. From one LSTM layer the input sequence is moving in the forward way from input to output and for the other LSTM the input sequence will be moved from backward. The following figure 2.1.5 Bi-LSTM network architecture shows the architecture of Bi-LSTM network and it has two layers of LSTM where one is moving forward and another one is moving backwards [23].

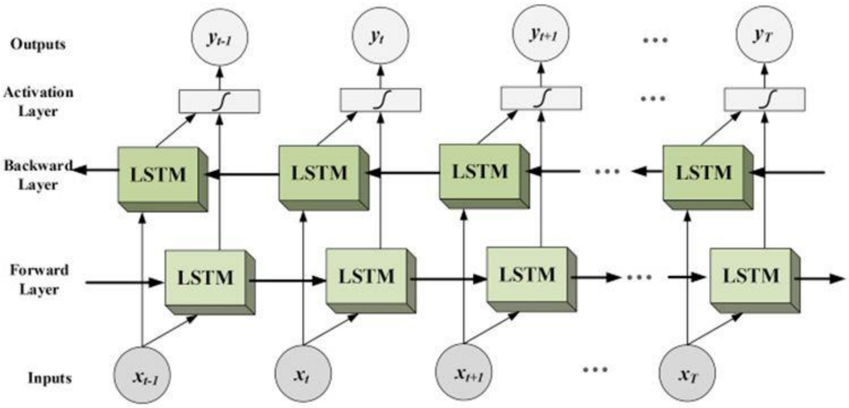


Figure 2.5: Bi-LSTM network architecture

## 2.2 Related Works

Sentiment analysis has performed on many languages, but our study focuses on Bengali language. Below some researchers done sentiment analysis work in Bengali language are studied and discussed here.

This paper shows bilingual sentiment classification using Bengali dataset and its machine translated English dataset. They have used two dataset which are from a game user comment dataset and another one is drama review from YouTube. Each dataset was manually annotated into positive, negative and neutral. balancing using SMOTE has been used on both of these datasets and it shows better performance increasing the F1 scores of the classifier such as LR, ET, RF, SVM, RR, LSTM, SVM classifier has the highest accuracy for Bengali and English dataset by 65.4% and 70.1% respectively [1].

To detect sentiment polarity of Bengali tweets this paper use 1500 Bengali tweets from SAIL contest 2015 and SentiWordNet lexical knowledge is used to augmented tweet word with sentimental tag found in the dictionary. Proposed model is built with CNN and compare it with the model with deep belief network which is DBN. The CNN outperforms DBN with 46.80% accuracy [2].

To detect emotions from Bengali text this paper proposed an automated system which detects happy, sad, tender, scared, excited, angry emotions from 7500 Bengali textual data. The proposed model is uses NB and Topical approach on both sentence and document level. Binning technique is used here if word goes missing from the dataset but available on the query sentence. Topical approach is highest with 90% on sentence level [3].

To detect depression related post from social media this paper presents Bengali sentiment analysis on depression dataset where the dataset is a combination of social media blogs with 2 emotion classes happy and sad is used. Sentiment classification run in document level where it classifies if a post is happy or sad using ML algorithms MNB, RF, DT, linear SVC. Counter vectorizer is used here and MNB get the best accuracy of 86.67% [4].

To detect real sentiments towards Bangladesh Cricket this paper shows sentiment analysis on manually created 1601 Bengali comments dataset with praise, criticism and sadness emotions and another dataset with 2979 Bengali comments with positive, negative and neutral class is chosen for the sentiment analysis. SVM, DT and MNB are performed and among them SVM outperformed them all by 64.596% accuracy in the emotion and 73.490% in the sentiment dataset. [5].

To enrich Bengali NLP fields, this paper use 7905 Bengali product reviews dataset and manually added positive, negative and neutral class sentiments to the dataset according to the user ratings of product review and oversampling was done to balance the classes. SVM, KNN, LR, RF, XGB algorithm all with 10 cross-validation has run on the model but KNN has highest with 96.25% accuracy [6].

BanglaSenti is a lexicon-based corpus or dataset of 61582 Bengali words including positive, negative and neutral labelling. By collecting data from SentiWordNet they transform the English words consisting positive, negative and neutral words into Bengali words with positive, negative and neutral words [7].

The purpose of building better feature extraction technique this paper with 2000 Bengali book review corpus detects sentiment polarity where polarities are positive and negative sentiments. Using tf-idf values with n-gram features is their main features in this whole paper. LR, MNB, RF, DT, KNN, SVM, SGD all with 10-fold cross-validation performed but MNB has highest 84% accuracy [8].

To detect the sentiments behind restaurant reviews this paper shows sentiment analysis of online restaurant review with 1000 Bengali review. For scarcity of Bengali benchmark dataset English benchmark dataset is translated into Bengali creating 1000 review dataset manually. Among the three ml algorithms DT, RF, MNB all with K fold cross-validation and MNB gave highest 80.48% accuracy with 6-fold cross validation [9].

This paper creates a benchmark dataset with 12000 Bengali reviews from YouTube. They translated the Bengali data to English data and supervised ml algorithms LR, SVM, RF, ET, unsupervised lexicon-based approach using VADER, TextBlob and SentiStrength used both on Bengali and English dataset to classify the sentiments. Used cross-domain dataset for transfer learning and SVM performs best with 93% accuracy in Bengali and 93.5% accuracy English corpus [10].

This paper shows sentiment polarity detection in Bengali tweets where the polarities are positive, negative and neutral. Dataset of 1500 Bengali tweets is used which is released in SAIL 2015 contest. SentiWordNet is used to augment the twitter dataset words with sentimental tag found in SentiWordNet. As the dataset is small for better performance 10-fold cross validation is performed with LSTM and has the best accuracy 55.27% than other models available on the contest. [11].

This paper presents the sentiment polarity detection of Facebook post and comments with 10819 Bengali textual data. Five classical approach SVM, NB, DT, AdaBoost, RF algorithm and two deep learning algorithm LSTM and CNN used here to detect the sentiment which is positive and negative. 96.95% accuracy is given by LSTM [12].

This paper proposed a network-based model with Bi-LSTM that detects sentiment polarity and categorized them into positive and negative from Bengali textual data. The data are collected from

Facebook comment and thus 30000 data collected and among them 10000 data has used here. SVM, DT, Logistic Linear Regression and also RNN with Bi-LSTM which is the proposed model has performed and RNN with Bi-LSTM outperformed them all by 85.67% accuracy [13].

This paper shows sentiment polarity detection in Bengali Facebook 7000 Bengali sentence dataset using doc2vec model with 2 sentiment class positive and negative. SGD, DT, KN classifier, LDA, GausssianNB, SM, Bi-LSTM has performed on these dataset and Bi-LSTM outperformed them all by 77.85% accuracy [14].

This paper introduces a Bengali dataset named BAN-ABSA with 9009 comments from Bengali news portal and manually annotated with aspect sport, politics, religions and others aspects are introduced with sentiment classification in positive, negative and neutral class. It uses supervised learning approaches for both aspect extraction and sentiment analysis using SVM, CNN, LSTM, Bi-LSTM algorithm and among them Bi-LSTM outperforms all the other model in aspect extraction with 78.75% accuracy [15].

This paper shows sentiment polarity detection on Bengali tweets using character n gram features and using MNB. For feature extraction they have use character n-grams. They have used Bengali tweet dataset in 2015 SAIL contest. They have also used SentiWordNet in SAIL contest 2015 their suggested model with MNB has 48.5% accuracy and outperformed all the other models [16].

This paper shows an English sentiment classification tool VADER and change it so that it can classify sentiment from Bengali language without changing the Bengali data into English data. They developed a Bengali polarity lexicon from VADER English lexicon. Also stemming, Bengali boosting words are added and to boost word valence they used bi-gram and tri-gram features. They perform a comparison of proposed VADER. The Bengali VADER performs better than this English VADER and it give faster performance than English VADER [17].

This paper presents multilingual BERT with the approach of transfer learning in Bengali sentiment analysis. Two Bengali datasets have been built with the first one is two class positive and negative sentiment and the other on is three class positive, negative, neutral sentiment. Three model one with BERT, another with word2vec and another one with fastText embedding proposed but BERT performs best with 71% and 60% accuracy for 2 and 3 class sentiments respectively [18].

This paper proposes a new approach using word2vec in this proposed model for sentiment analysis of Bengali comments with 16000 Bengali text line. They performed word embedding approach and word2vec is one of the techniques and they have created a new approach using word2vec with combining the two information where the first one is the dataset contextual information and second one is the sentiment information of the dataset comments. To create the new approach, they have built a list containing some highly positive and negative words put with the polarity score of each word is created and the accuracy is achieved to 75.5% [19].

This paper shows the study of two pre trained transformer model BERT and XLM-RoBERTa with fine tuning for sentiment analysis in Bengali language. Three Bengali dataset has been used to test the method and classified it into positive and negative. BERT and XML-RoBERTa are unsupervised language model which are pre-trained. They have fine-tuned both BERT and XLM-RoBERTa model. LSTM, CNN, GRU has use as a on top of BERT and XLM-R with three class 0.95 test accuracy and 2 class Bert with 95% accuracy and drama review from YouTube 0.97% with XLM-R book review 0.91% with Bert [20].

## 2.3 Comparative Analysis and Summary

Comparative analysis happens when one or more research works are compared in between them and also this analysis shows the similar and non-similar facts about the research works. Now in this study, all the research work that has been done in sentiment classification problems in Bengali language are discussed in the above related works section and there are twenty research papers that has been discussed there and now these research works comparative analysis and summary are discussed in this section. The following table 2.1: Comparative analysis of the mentioned research work shows the comparisons of these research works in basis of sentiment or emotion class, dataset, classification methods and maximum accuracy.

Table 2.1: Comparative analysis of the mentioned research work

| Reference Number | Number of emotions/sentiment class | Dataset | Methods of classification | Maximum accuracy |
|---|---|---|---|---|
| [1] | 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 datasets. | SVM with 65.4% for Bengali and 70.1% for English dataset. |
| [2] | 3 sentiments (Positive, Negative, Neutral) | 1500 Bengali tweet dataset which released on SAIL2015 contest | 1. CNN, DBN<br>2. Added sentiment tags from SentiWordNet | CNN with 46.80% Accuracy. |
| [3] | 6 emotions (happy, sad, tender, scared, excited, angry) | 7500 Bengali sentence datasets for sentence level analysis and 100 Bengali article from news portal for document level analysis. | 1. NB and<br>2.Topical approach<br>3. Binnig technique is also used here. | Topical approach with more than 90% accuracy. |
| [4] | 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%. |
| [5] | 3 emotions (Praise, Criticism, Sadness), 3sentiments (Positive, Negative, Neutral) | 1. Bangladesh cricket dataset of 1601 Bengali comments from Facebook and newspaper pages.<br>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.<br>2. SVM with 73.490% accuracy for BBC Bangla dataset. |

| | | | | |
|---|---|---|---|---|
| [6] | 3 sentiments (Positive, Negative, Neutral) | Product review dataset with 7905 reviews. | 1. SVM, KNN, LR, RF, XGB with all 10 cross-validation<br>2. performed oversampling for balancing the classes | KNN with 96.25% accuracy. |
| [7] | 3 sentiments (Positive, Negative, Neutral) | Collected 117660 words from SentiWordNet. | Lexicon based. | 61582 Bengali words corpus. |
| [8] | 2 sentiments (Positive, Negative) | 2000 Bengali book review corpus. | 1. LR, MNB, RF, DT, KNN, SVM, SGD with10 fold cross validation.<br>2. Used tf-idf values with n gram features. | MNB with 84% accuracy. |
| [9] | 3 sentiments (Positive, Negative, Neutral) | 1000 English translated Bengali restaurant 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. |
| [10] | 2 sentiments (Positive, Negative) | 12000 Bengali translated English reviews from YouTube | 1.LR, SVM, RF, ET for supervised learning<br>2.VADER, TextBlob, SentiStength for unsupervised lexicon-based approach.<br>3. Both approaches run on the English and Bengali dataset. | SVM with 93% accuracy. |
| [11] | 3 sentiments (Positive, Negative, Neutral) | 1500 Bengali Tweets dataset | 1.LSTM with 10 cross validation compared with SVM and MNB both with unigram and | LSTM with 55.27%. |

| | | | | |
|---|---|---|---|---|
| | | | extra addition of bigram to MNB<br>2. SentiWordNet for adding lexical knowledge on the dataset. | |
| [12] | 2 sentiments (Positive, Negative) | 10819 Bengali textual data from Facebook | SVM, NB, DT, AdaBoost, RF, LSTM and CNN | LSTM with96.95% accuracy. |
| [13] | 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. |
| [14] | 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. |
| [15] | 3 sentiments (Positive, Negative, Neutral) | 9009 Bengali comments from news portal. | SVM, CNN, LSTM, Bi-LSTM | Bi-LSTM with 78.75%. |
| [16] | 3 sentiments (Positive, Negative, Neutral) | used Bengali tweet dataset in 2015 SAIL contest | Character n gram features | MNB has 48.5% accuracy. |
| [17] | 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. |
| [18] | 2 sentiments (Positive, Negative), | 17852 comments with 3 class and 13120 comments with 2 class from online news website. | 1.BERT<br>2.word2vec<br>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. |

| | | | | |
|---|---|---|---|---|
| | 3 sentiments (Positive, Negative, Neutral) | | | |
| [19] | 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%. |
| [20] | 2 sentiments (Positive, Negative)<br>3 sentiments (Positive, Negative, Neutral) | 1.17852 Bengali data from news portal<br>2.13807 YouTube Bengali drama review<br>3.2000 Bengali book review | Transformer model BERT and XLM-RoBERTa with fine tuning is used here with 3 datasets to perform sentiment analysis in Bengali language. | XML-RoBERTa wiith 95% accuracy. |

From the above table there are many similarities and dissimilarities among the research works. Some have same sentiments class, and some have same domain in the dataset but every one work is different. The algorithms they have performed are similar, but their purpose and model are different. The dataset is very small also the data is noisy [2] also another work didn’t use social media data though their task is based on it [4] also another work training data is small [8] and another work uses Translated Bengali data which does not include the real way of expressing feelings in the original language [9]. So, these are some limitations of some research works. If we summarize our work according to that table then our dataset consists of 3 sentiments class which are positive, negative and neutral. We have used Facebook and YouTube Bengali comments as our dataset manually collecting them and labeling them with the mentioned sentiment class. For classification methods 2 models one with LSTM and another with Bi-LSTM is proposed and among them Bi-LSTM performs better, and this is the summary of it.

## 2.4 Scope of the Problem

This study will explore the Bengali dataset sentiment analysis using deep learning algorithms. The limitation that some dataset is very small and noisy [2], some study not collecting the data from social media though the task is based upon it [4], some study faces with small training dataset [8], some uses translated Bengali language which loses the originality of the language in terms of expressing opinions [9] and these are some limitations these research work faces. To solve these problems in our study we have used bigger dataset than this one [2], original Bengali language expressing opinions, better training size dataset than [8] and are collected from social media. So, these limitations we wish to fill up in our study.

## 2.5 Challenges

To fill up this limitation in our work we faced many challenges regarding our studies. Collecting Bengali data from social media to make a larger dataset with larger training data is the first challenge that we faced in this study. Bengali data is not so available such as English data and to collect it from the comments of Facebook and YouTube is very challenging. After collecting the data, we must put each data in google sheet and has labeled each comments manually with the sentiments class which we find challenging. Preprocessing the Bengali text is also a very difficult task. Removing the stop words and punctuation has been part of the data cleaning process and to clean that precisely has become very challenging for us.

# CHAPTER 3

# Research Methodology

## 3.1 Introduction

A methodology is the overall process of a research paper. A research methodology is all the steps that has been taken to solve the research problem to creating a solution is told in this section. This is the section where researchers systematically design the overall procedures to generate a better result than the previous research on the research topic. Our topic is sentiment analysis on Bengali social media dataset. Many researchers done work in sentiment analysis but very few exist in Bengali language. The methodology where data collection happens for sentiment analysis in Bengali language, many researchers faced problems because of the scarcity of Bengali data that's why some of them used English corpus and translated them into Bengali language [9] [10]. There is always a lack of benchmark dataset and that's why some created lexicon-based dictionary so that the machine can better identify the sentiment words and predict more accurately [7]. Supervised learning is the most popular method in this task [11].

Also, we have used this approach in this study. Classical algorithms and both deep learning algorithm have used in sentiment classification and some even compared both of the algorithm in their work [1] [11] [12] [13] [14] [15]. In methodology many researchers have divided the dataset into training and testing after the data collection [3] [9] and some also added a validation set to it [8]. The methodology in this study has many sections but these three sections are the most important and they are: data collection, data pre-processing, building the proposed model. To execute this section BNLP corpus and regular expression are used to clean the data and Keras library is used to tokenize the data and to build the model. Sentiment analysis of social media Bangla data is our research topic name and all the procedures that needed to build the proposed model are discussed in this chapter.

## 3.2 Research Subject and Instrumentation

Our research subject name is "Sentiment Analysis of Social Media's Bangla Data". We have named the subjects according to our work in this topic. As we wanted to work in analyzing the sentiments, we have also wished to work with Bengali language as it is our mother tongue and very few works exist here. In recent day social media is blasting with Bengali data as many people of Bengali natives are using this platform and their number is increasing each day and observing all these, we wanted to select this topic for our work and named the subject according to our task which is Sentiment analysis of Social Media’s Bangla Data. In research instrumentation research instruments are used for analyzing or measuring the data and here the LSTM and Bi-LSTM framework are used to do that. The frameworks and how it works are given below:

**LSTM**: LSTM has three gates which controls the flow of information and these gates are input gate, forget gate and output gate. The following figure shows the equation of these three gates.

$$
\begin{aligned}
i_t &= \sigma(w_i[h_{t-1}, x_t] + b_i) \\
f_t &= \sigma(w_f[h_{t-1}, x_t] + b_f) \\
o_t &= \sigma(w_o[h_{t-1}, x_t] + b_o)
\end{aligned}
$$

Figure 3.1: Equations of LSTM

In the figure 3.2.1: Equations of LSTM first equation is of the input gate which adds new information to the cell state and second equation is of forget gate which forget the unnecessary information and third one is of output gate which gives output from the whole process [24]. Also, here from the input gate $i_t$ represents input gate, from forget gate $f_t$ represents forget gate and from output gate $o_t$ represents output, $w_x$ means weights of each gate, $b_x$ means biases of each gate, $h_{t-1}$ means last LSTM block output happened at t-1 timestamp, $x_t$ means current input and t means timestamps [24]. There is another equation of cell state for LSTM. The following figure 3.2.2: Equations of LSTM 2 shows three equations and among them first one is of candidate cell state, second one is cell sate equations representing $c_t$ as cell states and third one is the final output [24].

$$
\begin{aligned}
\tilde{c_t} &= tanh(w_c[h_{t-1}, x_t] + b_c) \\
c_t &= f_t * c_{t-1} + i_t * \tilde{c_t} \\
h_t &= o_t * tanh(c^t)
\end{aligned}
$$

Figure 3.2: Equations of LSTM 2

**Bi-LSTM:** Bi-LSTM has information of sequence in every timestamp and this information is from both forward and backward [25]. The following figure 3.2.3: Structure of Bi-LSTM shows two hidden states combined where input is running from past to future and another one is from future to past acquiring it to have both past and present information [25].

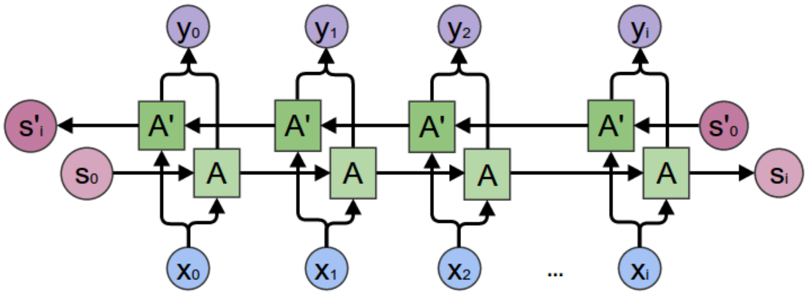


Figure 3.3: Structure of Bi-LSTM

## 3.3 Data Collection

Data collection is the first step in this overall procedure. Collecting Bengali data is difficult because it is not available such as English data. In recent, many people are expressing their emotions using Bengali language in social media [3] and so collecting Bengali data is became a little bit less difficult than the previous times but still it is not enough and the scarcity of Bengali data is still there and it is also called one of many resources poor language because of not having enough good dataset and content [8] that’s why collecting Bengali data for sentiment classification task is a bit difficult. In our manually collected dataset, we have used three sentiment class and added them with each comments manually. We have used google sheet to collect our data and add comments and sentiments column where comments are where the collected data goes and in sentiments column the sentiments according to the comments is manually put there. The following table 3.1: The Bengali dataset shows the first ten Bengali comments with its sentiments.

Table 3.1: The Bengali dataset

| **Comments** | **Sentiments** |
|---|---|
| ধন্যবাদ জানাই এই নাটকের সাথে জড়িয়ে থাকা সবাইকে। এত সুন্দর একটি নাটক আমাদের কে উপহার দেউয়ার জন্য। | Positive |
| অসাধারণ লাগে তোমার নাটক | Positive |
| এই ভাবে বিশ্বাসটা রাখতে শেখো প্রিয় | Positive |
| বাহ অসম্ভব সুন্দর লাগতেছে | Positive |
| মহিলা কি জেলে? | Neutral |
| এ ভাই আপনি তো কিছু জানেনই না,, শুধু বলেন দেখা যাক কি হয়. | Neutral |
| আলহামদুলিল্লাহ। মাএ ১৫টি ভিডিও আপলোড করে ৩৫০০০+ সাবস্ক্রাইব হলো। সবই আপনাদের ভালোবাসা। | Positive |
| তোদের মত আলতু ফালতু বিষয় নিয়ে সারাক্ষণ শুধু ভিডিও বানানোর জন্য ওরা পাগল না | Negative |
| আপনি ছাড়া আর কেহ হতাশ না উল্টা পাল্টা নিউজ না করে ভেলিড ইনফরমেশন দেন। | Negative |
| আমার তো মনে হয় খান দের থেকেও স্টার ডোম বেশি হৃতিক রোশানের | 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 different language has different rules, spelling, vowels etc. As every language is different, before applying to machine learning model data preprocessing is a crucial part to every language. For building the machine learning model successfully and fitting Bengali data into that model data preprocessing is a must [4]. Data preprocessing has two parts here. 1) Data Cleaning, 2) Tokenization Using Keras Library. In the first part the data was cleaned and the second part it was tokenized and converted into sequence of numbers.

1) **Data Cleaning:** In data preprocessing, we mainly cleaned the data. First, we removed the punctuation from the google sheets and then using this data we checked the null values in code. As our data has no null values after checking it the number of comments remain same. For cleaning the data, we have use BNLP corpus. We have defined stop words from BNLP corpus and used it to remove our Bengali dataset stop words. We have removed digits using regular expression. After that we have performed tokenization on the Bengali data using Keras library. The figure 3.4.1: Steps of data cleaning shows all the steps of data cleaning where Bengali data are cleaned well.

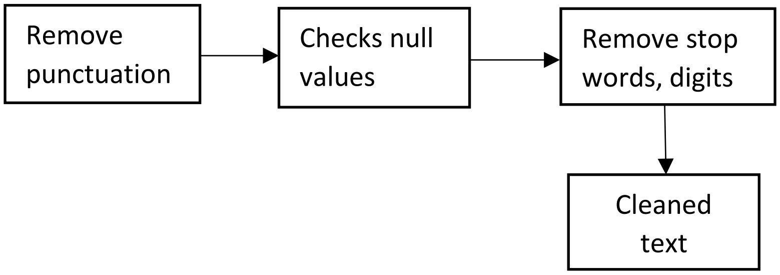


Figure 3.4: Steps of data cleaning

2) **Tokenization Using Keras Library:** Tokenization is the process where bigger text is broken into smaller text. Keras is an open-source library for implementing the neural networks. For tokenization in our Bengali dataset, Keras library is used. We have used Keras preprocessing module's tokenizer class which tokenizes our Bengali text. As we are using deep learning algorithm in our model and as we are using LSTM and Bi-LSTM algorithm in our model both needed sequence data to perform and that's why Keras tokenizer class is used here because this class does not only split every sentence into words but also convert each word into integers and whole sentence into sequence of integers values which later feeds onto that model. The two methods from the tokenizer class that turned the bunch of words into sequence of numbers are given below:

- **fit_on_texts:** It is a method which updates the vocabulary according to the text it gets from the dataset and gives vocabulary's each word a word index or an integer number based on the frequency of that word. Suppose dataset is “I am a girl. I love ice-cream. I am awesome” so according to this method it gives integer values basis on the word frequency or how many times the word occurs. So, after applying this method it will be I =1, am=2, a=3, girl=4, love=5, ice-cream=6, awesome=7. So here I occur more than am that’s why it gets more priority here. So, in this dictionary every unique word has given a number this is what the method does.

- **text_to_sequence:** It is a method where every sentence will turn into sequence of integers and these integers are the values that were used in the word index. So, if we follow the first example then this the dataset “I am a girl. I love ice-cream. I am awesome”. So, after applying fit on text all the words are given each number and then if this method performed then the result will be [[1,2,3,4], [1,5,6], [1,2,7]] so this the sequence of numbers and the numbers here are from fit on text word index number and the way these number seat together is according to the original sentence word sequence.

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

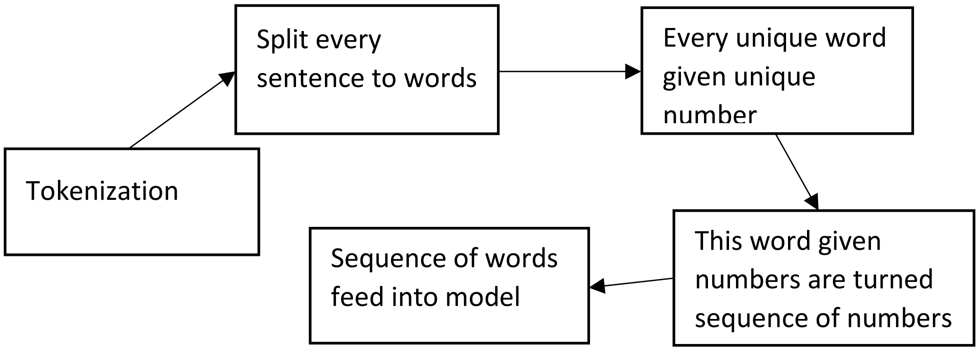

Split every sentence to words
Every unique word given unique number
Tokenization
This word given numbers are turned sequence of numbers
Sequence of words feed into model

Figure 3.5: Tokenization process

Then the sequenced numeric values are shaped into same lengths and then fed into the model. In the later stage after converting every text into a sequence of numbers then it is padded making the data into same size and shape and then categorical data which defines the sentiments were also converted into integer variable and later feed to both LSTM and Bi-LSTM model in the applied mechanism section.

## 3.5 Statistical Analysis

In this study sentiment classification task is performed and we have 3 sentiment class datasets of 4000 Bengali comments collected from social media Facebook and YouTube. Positive, negative and neutral are the 3 class sentiments. Here number of comments basis on sentiment class are given below:

| | |
|---|---|
| Positive comments: | 1244 |
| Negative comments: | 1518 |
| Neutral comments: | 1238 |

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

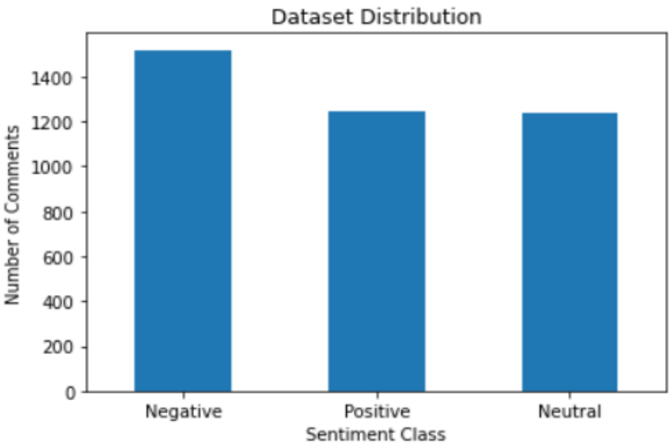


Figure 3.6: Dataset distribution

## 3.6 Applied Mechanism

Sentiment classification is a supervised learning approach in machine learning using bi-class or multi-class dataset [5]. Bi-class is where 2 sentiments class exist and multi-class is more than two class sentiments exist there. Our dataset has 3 sentiments so it is multi-class. For classifier we have selected neural networks rather than classical one because neural network performs better than classical or traditional classifier [12]. In this study, the following figure 3.6.1: Work flowchart shows the flow of procedures in the figure. There are two algorithms applied in the figure and they are one with LSTM algorithm and other with Bi-LSTM algorithm and both algorithms fall into the applied algorithm section.

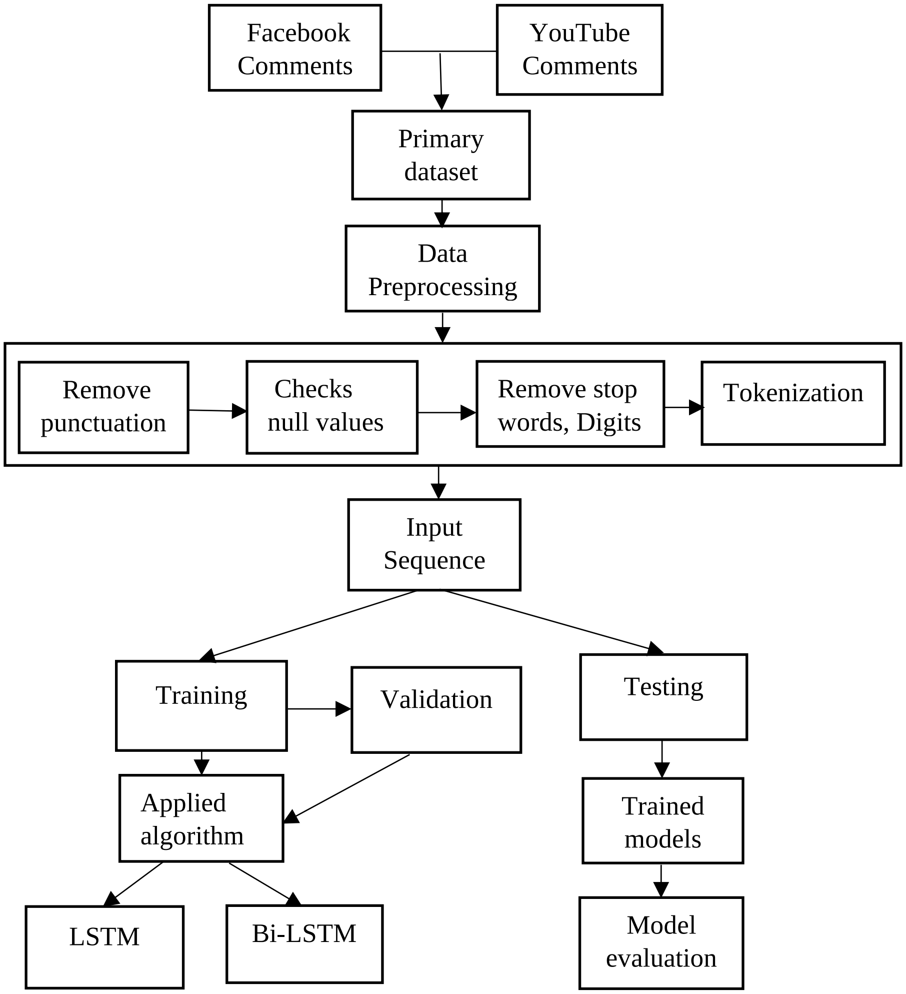


Figure 3.7: Work flowchart

Here we have used these two algorithms because they perform better than most neural network and we have followed the flowchart to build deep learning models using two algorithms such as LSTM and Bi-LSTM using Keras library. For both LSTM and Bi-LSTM model they have hidden layers where embedding layer is first of them and model has also an output layer and also one LSTM and two Bi-LSTM layers. Sequential model of Keras is called and then added layers to the model to build the LSTM and Bi-LSTM model.

**LSTM**: LSTM is an advance form of recurrent neural network. It remembers a data longer than RNN and also it using it's 3 gates it decides which data is more important and which data needs to forget and that’s why for all the sequence data related task LSTM is the best. For sentiment classification LSTM also work with sequence data and for this study LSTM uses some layers to execute the classification task and is described below. The LSTM model layers are given below:

1. **Embedding Layer:** It is the first hidden layer of neural network. Here all the integer encoded input will pass onto the next layer. Here this layer will reduce the dimension of the input vectors and pass it on the next hidden layer.
2. **LSTM Layer:** After the embedding layer input will pass into this layer. This layer has 2 neurons or nodes with a dropout of 0.2.
3. **Dense Layer/Output Layer:** It is a fully connected neural network layer where it’s 3 neurons are connected to the pervious neurons of the layers.it is an output layer with Softmax activation function and 3 neurons are used here because of 3 class sentiments and one sentiment can be the output from the 3.

The following figure 3.6.2: LSTM model shows all the layers of LSTM in the flowchart.

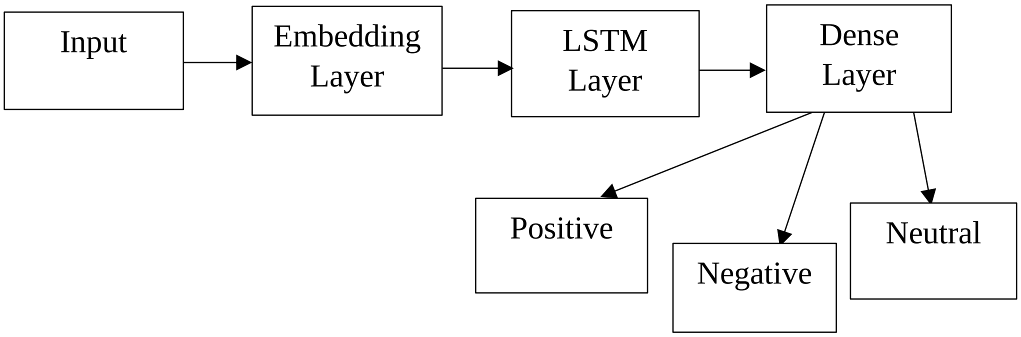


Figure 3.8: LSTM model

**Bi-LSTM:** Bi-LSTM is Bidirectional LSTM network which is an advance version of LSTM in the neural networks. The main difference is in its direction as the LSTM network we see it only uses forward direction means it only goes in one way but for Bi-LSTM it goes both ways means both forward and backward ways. But the similarities are that it uses sequential data and for sentiment classification task in this study it uses sequential data. As Bi-LSTM has some layers similar to LSTM and the layers we have used in this study are discussed below. Bi-LSTM model layers are given below:

1. **Embedding Layer:** It is first layer in this model and input enters this layer first. The input is integer coded input and this layer takes the input and reduced the dimension of it and pass it to the next layer.
2. **Bidirectional LSTM Layer:** This is a hidden layer and here 2 neurons are used here and the reduced dimension input that came from the embedding layer will pass into the Bidirectional LSTM layer and this layer has dropout rate of 0.2
3. **LSTM Layer:** This layer is another Bi-LSTM layer with 2 neurons and the input will pass from this layer to output layer which is a dense layer and also it has a dropout of 0.2
4. **Dense Layer/Output Layer:** This layer has 3 neurons in it and this layer neurons are fully connected with the previous layers. This is an output layer and as there are three sentiments the number of neurons here are 3. So, among these three sentiments class one class will be selected and will be the output. Also, in this layer for activation function Softmax is used. The following figure 3.6.3: Bi-LSTM model shows the flow of inputs in all the layers of this Bi-LSTM model.

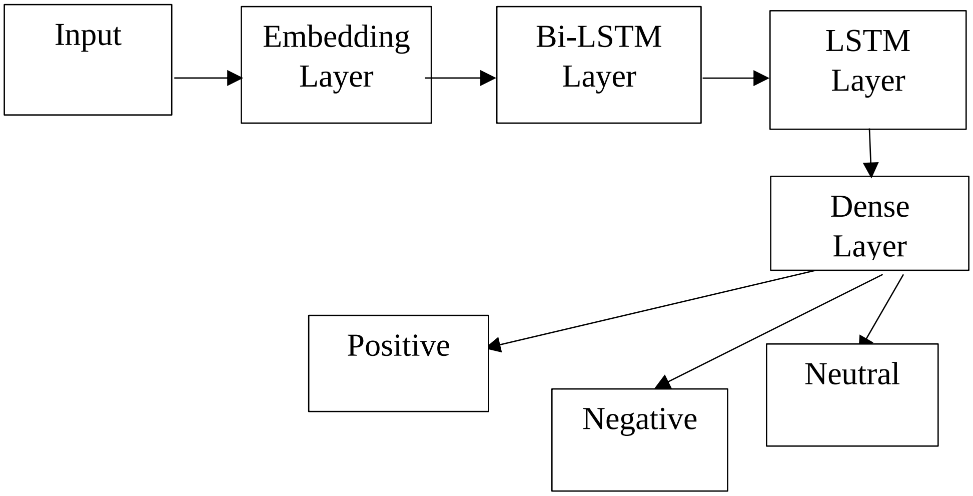


Figure 3.9: Bi-LSTM model

## 3.7 Implementation Requirements

Python language has used here to do all the programming and Google Colaboratory used as a platform to implement the language. Google Sheet was used to collect the data. Pandas library used to read the datasets and encoding categorical values to integer and BNLP corpus used for data cleaning and Keras library was used to tokenize the data and for building the models. Numpy library was used to show arrays Scikit-learn library was used for splitting the dataset. Matplotlib and seaborne library are used to visualize the data. To perform this task or sentiment analysis the components we have used are given below:

- **Google Colaboratory:** We implement python code in this platform.
- **Pandas:** Used to read and changes in the file.
- **Numpy:** Used it to show arrays of data.
- **Scikit-learn:** Used To split the dataset into training and testing.
- **Seaborn:** For visualizing the dataset.
- **Matplotlib:** For visualizing purpose.
- **Keras:** To implement the algorithms and tokenization process.
- **BNLP corpus:** For data cleaning purpose.
- **Google Sheets:** All the data are collected using Google sheet.
- **Regular Expression:** For data cleaning purpose to remove the digits.

# CHAPTER 4

## Experimental Results and Discussion

### 4.1 Introduction

Experimental results section shows the result of the research study and analyzes the result by implementing many graphical representations. Our task is sentiment analysis and many researchers have applied sentiment analysis on different dataset and they all show a result of the studies that they have studied. These results conducted as experimental results by implementing many algorithms on the training set and after that evaluating the result in the testing set. The experimental results can be represented in two ways which are statistical and graphical representations [9] where accuracy, precision, confusion matrix is example of statistical and accuracy, loss graph is example of graphical representations to analyze the results these are necessary and below is a definition of them.

**Accuracy:** It is the value of how many predicted classes are right among all the classes.

**Precision:** In the test set all the sentence from a particular class performed by the classifier and among the sentence which number of sentences are predicted right in the basis of overall this particular class prediction then this number is called precision [5]

**Recall:** In the test set all the sentence from a particular class performed by the classifier that are predicted actually and among the sentence which number of sentences are predicted right then that number is called recall [5].

**F1 score:** For a particular class it is a weighted harmonic mean for both precision and recall [5].

**Classification report:** This report is a combination of values of precision, recall, f1 score for all the classes that is used in the dataset and it is used to measure the performance of the classifier.

**Confusion matrix:** It shows the actual class and prediction class values according to the used classes in the experiment and it is used to measure the performance of the classification model [9].

## 4.2 Experimental Setup

For experimental setup first the dataset needs to be divided into training and testing set. Then the algorithm will perform on the input data and give prediction on the classes. But to increase the level of correct prediction means accuracy some parameters need to be tuned in the training time for both LSTM and Bi-LSTM model to increase the accuracy level. But first the dataset needs to split for the experiment to start. Here the data split and parameter tuning for both LSTM and Bi-LSTM are given below.

**Dataset Splitting:** The dataset is divided into two set which is training and testing set. The training set was split into 80% and testing into 20% for both LSTM and Bi-LSTM model. But among training set, a validation set was created using 10% of training data. So now training data is 70%, validation data 10%, testing data 20%. The following figure 4.2: Dataset splitting shows the split of the dataset.

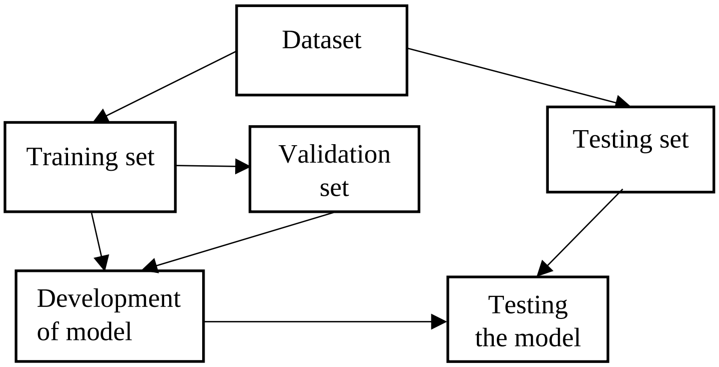


Figure 4.1: Dataset splitting

**Hyper-Parameters Tuning for LSTM:** For the layers in LSTM, dimension of the LSTM layer, epochs, batch size etc. all can play an essential role to increase the model accuracy. So, tuning this parameter the model accuracy can be increased. The final layer or output layer of LSTM has activation function Softmax with categorical crossentropy as loss function as this is a multi-class classification also Adam optimizer used here. The hyper-parameters are given below:

**Input dimension:** It is the size of the vocabulary of text and it is in embedding layer here the input dimension is 5000.

**Output dimension:** It is in embedding layer and it is the size of the output vector. Here it is 64.

**Dimension of Bi-LSTM and LSTM layer:** The dimension of LSTM layer is the number of neurons or nodes that is situated at the layer and here it is 16 neurons.

**Epoch:** Epoch is the number of times the dataset passed through forward and backward of any neural network. Here it is 40.

**Batch size:** The batch size is the how many inputs it takes at a time and here it is 12.

**Hyper-Parameters Tuning for Bi-LSTM:** For the layers in Bi-LSTM, Hyper-parameters such as epochs, batch size, input dimension etc. play an important role to increase the accuracy of the Bi-LSTM model and so tuning the parameters is very important to get a good result for the model. Here the output layer of Bi-LSTM uses categorical crossentropy as loss function and Adam as optimizer and Softmax as active function similar to LSTM because both model uses same dataset which is multiclass and that’s the main reason of using these optimizer and loss function and active function. Below the hyper-parameters are given that is used in this model.

**Input dimension:** 5000 size of input dimension is used in this model.

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

**Dimension of Bi-LSTM layer:** The dimension of Bi-LSTM layer in this model is 16 neurons and has another LSTM layer with 16 neurons.

**Epoch:** The number of epochs is 40 in this model.

**Batch size:** 12 batch size is used here

## 4.3 Experimental Results & Analysis

For the experiments the whole dataset was divided into training and testing set. From training data, a validation data is created with 10% of training data. So, it is 70% data for training, 10% for validation and 20% for testing set among 4000 text data. The training set for both models has performed and evaluated by the 20 percent test set both for LSTM and Bi-LSTM model and using metrics module form scikit library accuracy, loss for both training and testing are counted and shown in the graph.

**1. LSTM:** The accuracy for LSTM model is 96.88% and loss is 0.200 for testing set. The classification report was also created using testing values and prediction values also using the metric module from scikit library. The table and the figure are given below. The table 4.1: Classification table for LSTM shows the classification report which has precision, recall, F1 score and support where sentiment 0 means Negative, 1 means Neutral, 2 means Positive.

Table 4.1: Classification table for LSTM

| **Sentiment** | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| 0 | 0.97 | 0.97 | 0.97 | 318 |
| 1 | 0.98 | 0.95 | 0.97 | 242 |
| 2 | 0.96 | 0.97 | 0.97 | 240 |

The figure 4.3.1: LSTM model accuracy and loss in training and testing set shows the difference in accuracy and loss of training data and testing data which is used of validation split of 10% using the training data after every epoch is shown below.

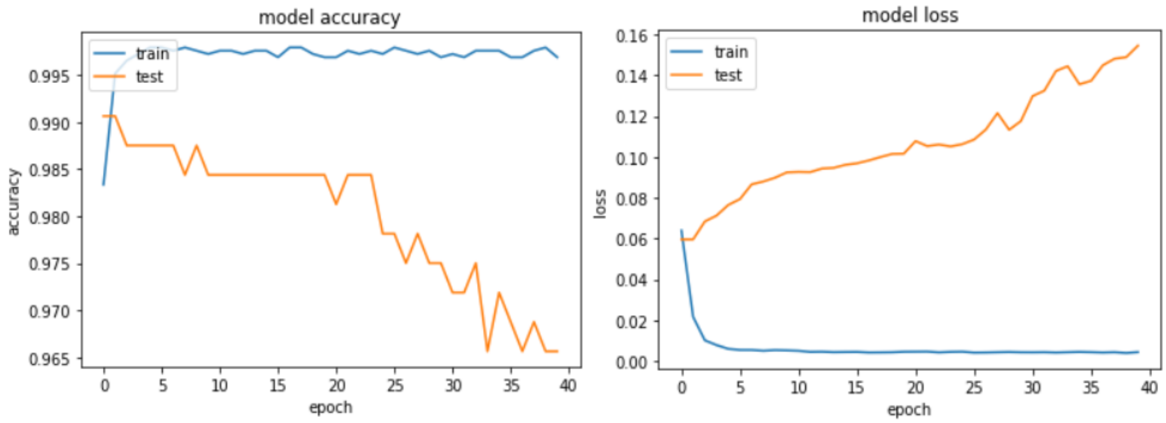


Figure 4.2: LSTM model accuracy and loss

**2. Bi-LSTM**: This model shows accuracy and loss for both training sets and for testing set 97.25% accuracy and loss 0.129 is test set accuracy and loss. The Figure shows the difference of training and testing the dataset after every epoch were testing the set using validation split 10% on the training data. In the figure 4.3.2: Bi-LSTM model accuracy and loss of training and testing set, the left one shows the accuracy for training and testing and the right one shows the loss of training and testing dataset.

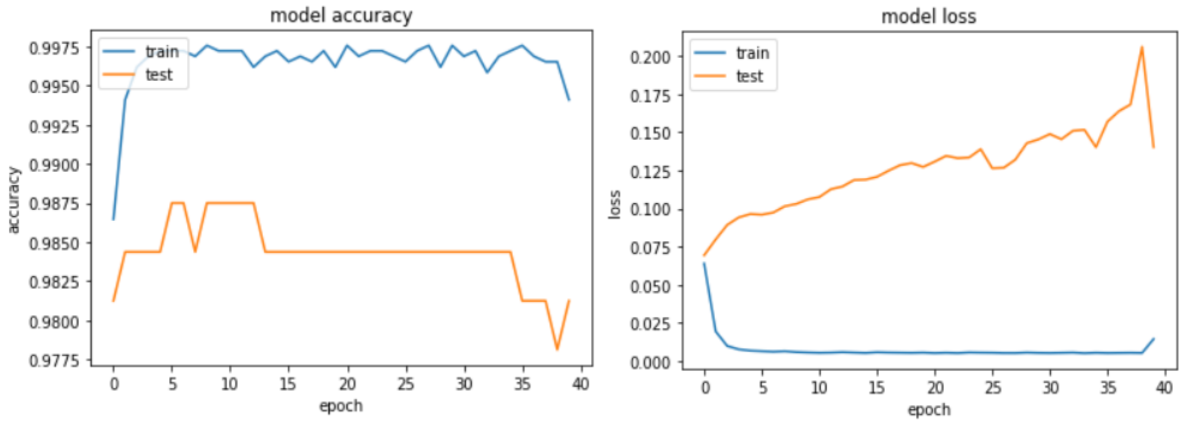


Figure 4.3: Bi-LSTM model accuracy and loss

The table 4.2: Classification table for Bi-LSTM shows the classification report of Bi-LSTM model showing precision, recall, F1 score and support for different sentiments where 0 means negative sentiment, 1 means Neutral sentiment and 2 means Positive sentiments.

Table 4.2: Classification table for Bi-LSTM

| **Sentiment** | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| 0 | 0.98 | 0.98 | 0.98 | 296 |
| 1 | 0.98 | 0.96 | 0.97 | 270 |
| 2 | 0.95 | 0.97 | 0.96 | 234 |

## 4.4 Discussion

Sentiment analysis has been performed using deep learning model LSTM and Bi-LSTM. Both of those model uses a training set of 70% data and validation data 10% and test data of 20%. The training data has been trained using LSTM and Bi-LSTM model and this model is used from Keras library. The model was trained using the data with epoch of 40, batch size of 12 for both models. After evaluating the test data, the accuracy for LSTM was 96.88% and Bi-LSTM was 97.25%. According to the classification report for LSTM it for negative class f1 score is 97, positive class score is 97 and neutral class is so 97 which means the model can correctly tell that 97% data are negative, 97% are positive and 97% are neutral. For Bi-LSTM model negative, neutral and positive f1 score are 0.98, 0.97, 0.96 respectively and so it means that Bi-LSTM model can correctly classify for both negative 98%, neutral class 97% data and positive class 96% data. So Bi-LSTM F1 score is better than LSTM. Also, as BI-LSTM is the advance version of LSTM so it will perform better than LSTM model and the classification table shows that. So, Bi-LSTM did better than LSTM in this dataset.

# CHAPTER 5

# Impact on Society, Environment and Sustainability

## 5.1 Impact on Society

Internet based life has huge impact on our society. People cannot live without internet now. Using this Internet social media is everywhere and people are using this social media every day. Online store to personal blogging social media has everything. And in these media or network often sentiment analysis is performing there to observe the sentiments of how audience is taking everything if it is positive or negative or neutral. Sentiment analysis also uses to monitor the customer behavior towards a brand or reviews towards a brand etc. And as people and business are going online many sentiments analysis has performed on the online social media business or personal blog etc. In our society everything is becoming internet-based day by day and as this internet world is growing rapidly sentiment analysis on this world is also can be seen more often and thus in society sentiment analysis has a great impact. People are getting more at ease with the technology each day. As we can see in the recent past many people did not want to buy from online shops but now more people are buying from it. As people post their reviews to the online store now all the people can see it and if sentiment analysis is performed on the reviews more accurate data that how much is the positive and how much is the negative and how much is the neutral can be seen and understood easily and owner will take steps to find out the problem if the data are more negative. So, it is seen that sentiment analysis has huge impact on our society.

## 5.2 Impact on Environment

There is a huge impact on environment in sentiment analysis. As internet is a part of it, using internet can cause damage to the environment we are living. Internet causing carbon emission which is bad for the environment. Internet has made our life easier with technology such as sentiment analysis, but it also impacts our environment as advancing of internet is making a part of the people lazy mentally and physically which hampers the way the person is living his life. If new things are created its impact on the environment should be kept in mind. As of this study shows a Bengali dataset how can we use it to give a better accuracy and make prediction so that the algorithm can predict seeing an unseen data that which class the data belong to. This will be helpful if someone try to find the public opinions about any things in social media, he or she can apply this model and can get the predicted values which help them to understand the opinions of the public more. But this can create an impact on environment as our environment is degrading day by day using the model technology is actually hampering the environment such as emitting carbon emission can cause serious issues in our climate and can cause serious issues in earth atmosphere level. So before creating something new we all should see the qualities of that product that if it will cause any issue to the environment or not.

## 5.3 Ethical Aspects

Every device to website to new product to technology everything has its ethical aspects. Sentiment analysis is mainly targeted by brands to see their buyer's opinion about their new or updated products. Such as customer review can help buyers to see the important information about the usage of products which helps the buyer and also helps the company for seeing the opinion of public in that matter. But when people misuse sentiment analysis for their personal gain that moment its losses it's ethical aspects. Ethical aspects also show how a product must be used maintaining the ethics. Ethics include privacy for a person and their consent to that matter. If sentiment analysis performs on without the person's consent and endangering their privacy, then those analysis will lose its ethical aspect. So not sentiment analysis but for any system ethical aspects and a person's privacy and consent must be kept protecting ethical aspects of that system.

## 5.3 Sustainability Plan

Sustainability is the term that measures the time of how much longer a certain thing stays and it can be a new system or a new product or a new lifestyle. If we think of sustainability, then internet is not sustainable as it is creating increasing amount of carbon emission each year. Though internet is very helpful to live in the modern world, but it is not safe for the environment as it is emitting carbon which is one of the greenhouse gases effecting our ozone layer. Only it will be sustainable if it does not damage earth nature system. Here sustainability means how much longer it remains and the plans to sustain it. For sustaining our work, we must increase accuracy so the model gives better result than any other models also as our work is useful for researcher and Bengali content creators it can be helpful for them ensuring it sustaining a longer time.

# CHAPTER 6

# Summary, Conclusion, Recommendation and Implication for Future Research

## 6.1 Summary of the Study

Here in this study using deep learning algorithm LSTM and Bi-LSTM has been implemented here for sentiment analysis of Bengali social media content's Bengali comment text data that has been collected from social media such as YouTube and Facebook. The dataset is consisting of 4000 comments and 3 sentiment classes named Positive, Negative, Neutral exist there. The dataset was manually created by collecting Bengali comments YouTube and Facebook posts and video comments. As we are working with algorithms that is popular for its sequenced data so the input must be sequenced integers and for that Keras tokenizer class with methods that will turn tokens to sequence of integer is used. After all of the text turning into numerical values the input will be passed to the LSTM and Bi-LSTM model and here the LSTM model give 96.88% accuracy and Bi-LSTM gives 97.25% accuracy. So, Bi-LSTM shows better result than LSTM.

## 6.2 Conclusions

In this era text has become treasure of information with opinions of people [1] and for these detecting sentiments from text is necessary now and for this our study has done sentiment analysis on dataset from social media. The work is focuses on Bengali language as there are few works have been done in this field. Our dataset is of 4000 Bengali comments takes from Bengali contents of Facebook and YouTube. We have performed machine learning approach on our work and we have chosen neural networks such as LSTM and Bi-LSTM for the sentiment analysis task where Bi-LSTM give better performance than LSTM with 97.25% accuracy. Using our work Bengali content creator can know public opinions about their contents without reading all the comments and researcher can use this dataset for their work.

## 6.3 Implication for Future Study

The future improvements in this study are discussed here. In this study we can see that the dataset that we have working not so large so for future improvements the size of the dataset will be increased. Also, we will add Bengali data from TikTok and Instagram and increase our horizon in terms of social media content. As the Bengali comments are collected from Bengali contents there are also some comments in Bengali language but written in English words in the Bengali contents and these issues will be solved in future implications. Also, some post such as sad post dataset are maximum negative comments which can cause class unbalancing issues to our model and we will solve these issues in our future works. Also, we will use pre-trained word embedding to use for both model in the future work which can increase the accuracy of the current models.

# REFERENCES

[1] S. Sazzed and S. Jayarathna, “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.

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

[3] R.A. Tuhin, B.K. Paul, F. Nawrine, M. Akter and A.K. Das, “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.

[4] M.R.H. Khan, U.S. Afroz, A.K.M Masum, S. Abujar and S.A. Hossain, “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.

[5] S.A. Mahtab, N. Islam and M.M. Rahaman, “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.

[6] M.T Akter, M. Begum and R. Mustafa, “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.

[7] H. Ali, M.F. Hossain, S.B. Shuvo and A. Al Marouf, “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.

[8] E. Hossain,O. Sharif, and M. Moshiul Hoque, “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.

[9] O. Sharif, M.M Hoque and E. Hossain, “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.

[10] S. Sazzed, “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.

[11] K. Sarkar, “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

[12] P. Chakraborty, F. Nawar and H.A. Chowdhury, “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.

[13] A.A. Sharfuddin, M.N. Tihami and M.S.Islam, “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.

[14] M.T. Hoque, A. Islam, E. Ahmed, K.A Mamun and M.N. Huda, “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.

[15] M. Ahmed Masum, S. Junayed Ahmed, A. Tasnim and S. Islam, “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.

[16] K. Sarkar, “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.

[17] A. Amin, I. Hossain, A. Akther, and K.M. Alam, “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.

[18] K.I. Islam, M.S. Islam and M.R. Amin, “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.

[19] M. Al-Amin, M.S. Islam and S.D. Uzzal, “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.

[20] A. Bhowmick and A. Jana, “Sentiment Analysis for Bengali Using Transformer Based Models”

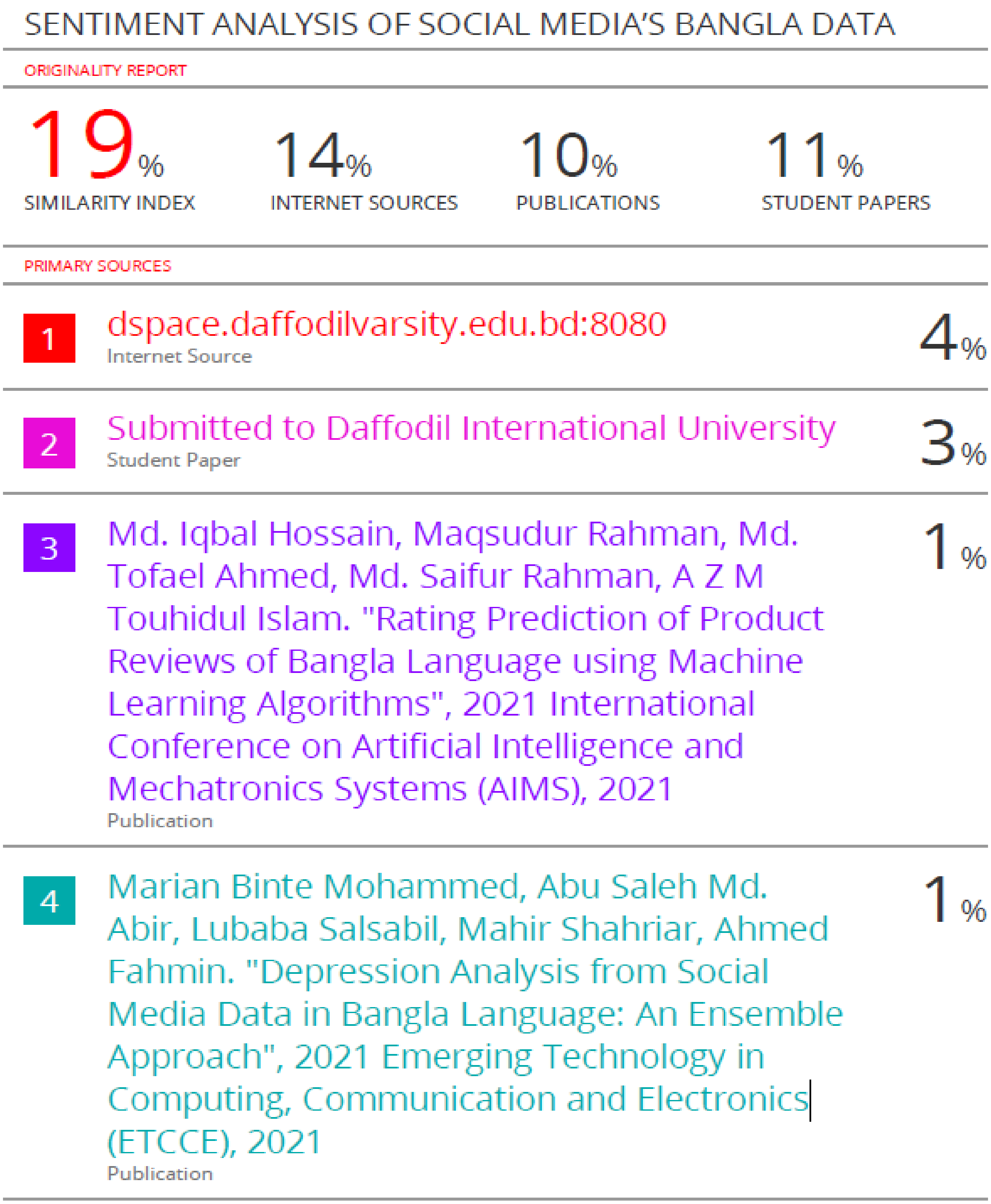
[21] Bengali language Wikipedia, available at
<<
https://en.wikipedia.org/wiki/Bengali_language#:~:text=With%20approximately%20300%20million%20native,of%20speakers%20in%20the%20world.>>, last accessed on 15-06-2022 at 8.21 pm.

[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.

[24] LSTM and its equations, available at<< https://medium.com/@divyanshu132/lstm-and-its-equations-5ee9246d04af >>, last accessed on 15-08-2022 at 12.34 pm.

[25] Bi-LSTM, available at << https://medium.com/@raghavaggarwal0089/bi-lstm-bc3d68da8bd0 >>, last accessed on 15-08-2022 at 12.40 pm



SENTIMENT ANALYSIS OF SOCIAL MEDIA'S BANGLA DATA

ORIGINALITY REPORT

| 19% | 14% | 10% | 11% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | dspace.daffodilvarsity.edu.bd:8080<br>Internet Source | 4% |
| 2 | Submitted to Daffodil International University<br>Student Paper | 3% |
| 3 | Md. Iqbal Hossain, Maqsudur Rahman, Md. Tofael Ahmed, Md. Saifur Rahman, A Z M Touhidul Islam. "Rating Prediction of Product Reviews of Bangla Language using Machine Learning Algorithms", 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), 2021<br>Publication | 1% |
| 4 | Marian Binte Mohammed, Abu Saleh Md. Abir, Lubaba Salsabil, Mahir Shahriar, Ahmed Fahmin. "Depression Analysis from Social Media Data in Bangla Language: An Ensemble Approach", 2021 Emerging Technology in Computing, Communication and Electronics (ETCCE), 2021<br>Publication | 1% |