Bachelor of Science in Computer Science & Engineering

Developing a System to Analyze Comments of Social Media and Identify Friends Category

by

Tasfia Hyder

ID: 1504053

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

May, 2021

Developing a System to Analyze Comments of Social Media and Identify Friends Category

Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Tasfia Hyder
ID: 1504053

Supervised by

Dr. Mohammad Shamsul Arefin
Professor
Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)
Chattogram-4349, Bangladesh.

The thesis titled '**Developing a System to Analyze Comments of Social Media and Identify Friends Category'** submitted by ID: 1504053, Session 2019-2020 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

______________________________ Chairman

Dr. Mohammad Shamsul Arefin
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

______________________________ Member (Ex-Officio)

Dr. Md. Mokammel Haque
Professor & Head
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

______________________________ Member (External)

Dr. Mohammed Moshiul Hoque
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and that neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Signature of the candidate

Date:

Acknowledgements

This thesis would not have been feasible without the help of a lot of individuals. First and foremost, I want to express my gratitude to Professor Dr. Mohammad Shamsul Arefin, my supervisor, for his unwavering support and assistance. His study motivation, endless persistence, empathy, and determination to let me fly on my own have all aided in my humble growth.

I would moreover like to thank my defense committee individuals, for giving me with their profitable advices and input. I am inconceivably thankful to Prof. Dr. Md. Mokammel Haque, Head, Dept. of CSE, CUET for his extraordinary back all through my undergrad instruction. I would moreover like to thank all officers/staffs at IICT, CUET for demonstrating their mind blowing Web administrations.

Finally, I would like to thank my parents,friends and family for their continuous love and support throughout my journey.

Abstract

Today, users on the social platform are expressing their emotions, ideas, proposals and views.The opinion may articulate critical opinions in various ways and may include different polarities such as positive, negative or neutral and it is often a difficult challenge for people to appreciate the feeling of each opinion and the time it takes. The analysis of the feeling in each statement will resolve this issue.This paper presents a framework to analyze social media comments e.g. Facebook and identify a category of friends. First, some public profiles are selected and comments are retrieved from different posts of them. Secondly, those comments are stored as dataset and pre-processed for sentiment analysis. After that, the pre-processed data trained and tested in a sentiment analysis model developed by us. From the sentiment of data, we then identify friends. For example, a friend with a positive sentiment of comment can be considered as a good friend. The evaluation of the performance is measured. A decent accuracy is achieved by the system.

Keywords: Social Media,Sentiment Analysis, Natural Language Processing, Facebook Comments, Long Short Term Memory(LSTM), Friends category.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Social media is an interactive computer-mediated technology that enables information, ideas, career interests and other forms of expression to be created or shared via virtual communities and networks. In today's world, social media plays an important role in our lives. It has become an important forum for expressing opinion and thoughts. Meanwhile, the popularity of internet user is rapidly increasing that use social media for expressing their opinions.

Sentiment analysis is the automatic mining of attitudes, opinions, and emotions from sources of text, speech, and database using Natural Language Processing (NLP).The object of sentiment analysis is to digitally acknowledge and express opinion. There are huge number of information in web as well as in social media. Through Sentiment Analysis we can learn about public opinion about a particular object or a public figure. It is useful for product reviews, classification of feedback, opinion mining etc.

Comment section of social networking sites like Facebook and Twitter can be represented as social media. Such social media will catch the thoughts or word of mouth of millions of people. People are now sharing facts about their lives, knowledge, experiences and opinions with the entire world through it. They express their opinions and state comments to participate in events.

As we can see opinions and comments differ from person to person. A certain thing can have positive opinion from a individual and negative from another. It is a common issue in social media as every individual's thoughts and perspective varies from another. Social media has been seen as a medium for people to make positive or negative remarks.

Our proposed framework addresses and analyzes social media comments. At first, some Facebook profiles are chosen and the data with required parameters is retrieved. Then the results are pre-processed and primed for the next level of sentiment analysis. To carry out this function, an LSTM model is built with suitable parameters. The model has been validated and trained. The comments are graded according to the score received as positive, negative and neutral. It's a part of our job. The next step will be to identify the followers we worked on. Followers are graded in three categories based on the sentiment of comments.

1.2 Framework/Design Overview

Our framework overview is given in the following diagram-

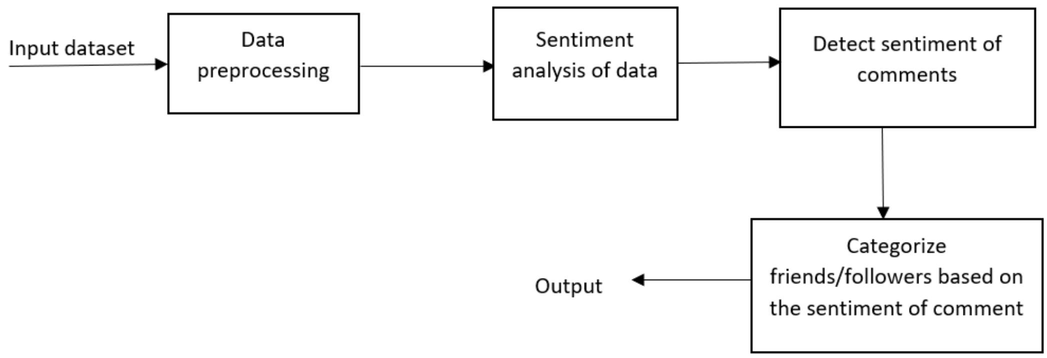

Figure 1.1: Basic Block Diagram of Proposed Framework

This framework has basically three main steps: (1) Data pre-processing

(2)Sentiment Analysis of data

(3)Categorization of friends or followers based on sentiment.

1.3 Difficulties

We encountered many challenges when designing the system:

- As it was mostly based on data mining and machine learning, collecting data was very time consuming. We had to collect it manually.
- We needed to carefully train and test our model so that we could get better result.
- We worked for two separate languages i.e English and Bengali which was very challenging.
- There were complex sentences as well as sarcastic comments therefore desired accuracy rating was not easy to get.

So it was quite challenging for us to build a successful system.

1.4 Applications

Some useful application of our thesis is given below:

- With the increasing use of social media, our system can identify the sentiment of comments thus allowing the user to identify which followers or friends are hateful or devoted towards them.
- No need to manually check all the comments to identify bullies or appreciating comments.
- Users can use this system to further block followers those who always spread negativity.

1.5 Motivation

Social media contains vast source of information. People are now sharing facts about their lives, knowledge, experiences and opinions with the entire world through it.There are thousands of comments found in social media like Facebook which are productive, optimistic ,offensive, bad in different posts.The system will help us to recognize people's opinion through their comments on different issues discussed as a post in social media and classify the people based on their comment.

Table 1.1: An Overview of Social Media Users' Posts and Comments on them

USER	POST ID	FOLLOWER ID	COMMENTS
Bill Gates	1.01578E+16	Andrias Mulambo	You are a danger to humanity.
		Arturo Jose	I love your work, Bill! Keep it up! Keep finding new ways to benefit the world and its future.
		Tania Polidoro Cortón	Bill Gates, with all the billions you have, you should give the profit to charity please! you have a lot of power in this planet, in this life.
ওবায়দুল কাদের	4.40001E+15	MB Kaniz	সুস্থ থাকুন নিরাপদ থাকুন দোয়া রইলো।
		Farhana Papia	আপনার আইডি থেকে এখন দেশ বিরোধী যত সমলোচনা হয়া সেটা নিয়ে কিছু লেখার চেস্টা করুন। আপনি একজন দলের সাধারণ সম্পাদক এইটুকু আশা করি।
		Zahidul Hasan	সাবধানে থাকুন, নিরাপদ ভাবে থাকুন,শুভ কামনা রইলো স্যার। করোনার এই ঊর্ধ্বগতিতে আপনার অসামান্য অবদান রয়েছে।
Barrister-Syed-Sayedul-Haque-Suman	1.33788E+15	মীয়াদ হোসাইন	ফুটবল খেলার নামে সারাদেশে চসে বেড়িয়েছেন আর মানুষের জটলা বাঁধিয়েছেন।
		S.A. Soukhin	আপনার কাজগুলো অবশ্যই প্রশংসার যোগ্য।।।
		Md Al-Amin	সব বিষয় আপনে লাইভে আসেন,কিন্তু জাতিয় নির্বাচন নিয়ে একটা লাইভ করেন না,কেন , কারন তোমরা শয়তানের অনুসারী
Shakib Al Hasan	3.15591E+14	Kazi Nayeem	আপনার সন্তানদের টিকা দেওয়া জন্য ধন্যবাদ, আল্লাহ আপনার ফ্যামিলিকে সুস্থ রাখুক ❤️
	3.15404E+14	Abdullah Al Arif	এইবার ভাল করতে হবে বস 👌❣️
		Mahafujar Rahman Mukut	সাকিব তোমার দ্বারা আর কিছু সম্ভব নয়,দেশে এসে কাকড়া চাষে মনোযোগ দাও 😎
Narendra Modi	1.01652E+16	Guru Datta	We are grateful to you for leading the country and doing right things in spite of criticism. 🙏😇
		Ajoy Biswas	You are one of the most failed Prime Ministers in the history of India.You have no right to play with the lives of common people.You should resign now
		Samuel Jacob	Great decision, will save lives of millions in Ventilators and gasping for breath.

From this table below we can see there are various types of comments in an individual's posts. Some are appreciating, some are hateful or bullies as well as some neutral comments. We as a human can easily detect the sentiment from reading the comments. But there are massive amount of comments found in different posts. It is too time consuming by checking those comments manually and knowing who wrote what type of comments. The main motivation is to reduce the work and help user to identify his or her adherent and adversary followers.

Analyzing the comments of social media posts can help us to recognize the public opinion. People often express their opinion aggressively whether the corresponding post is related to aggression or not. This may affect the person who is in charge of the particular social media profile. Therefore by analyzing the comments against a social media post, we can classify the comments into positive, negative and neutral category. As a result, the followers can also be classified into groups which will be a beneficial thing for the social media handler.

1.6 Objectives

The project will be carried out with an aim to achieve the following objectives:

- To select profiles from social media to analyze their followers.
- To collect comments from various posts of a particular profile and analyze the sentiment of those.
- To classify the comments into three sections – positive, negative and neutral.
- To group the followers into good, hostile and neutral category based on their comments.

1.7 Contribution of the thesis

A thesis work is conducted to attain a set of objectives whether it is to characterize a new methodology or to move forward the existing ones.In this thesis, our motive is to work on Facebook comments to identify friends category which is not done in the past.The primary contribution of this thesis is the following:

- Building a dataset of Facebook comments from various users profile.
- Developing a sentiment analysis system for classifying Facebook comments.
- Facebook friends/followers classification based on the nature of comments done by them.

1.8 Thesis Organization

The thesis report is organized as follows:

- Chapter 2 gives a brief summary of previous research works in the field of sentiment analysis of English and Bengali language data.
- Chapter 3 gives a total view of the proposed methodology.
- Chapter 4 provides the description of the working data set and analysis of the performance measure for the proposed framework.
- Chapter 5 shows the general overview of this study is included as well as a number of potential suggestions.

1.9 Conclusion

An outline of our proposed system to analyze social media comments to classify friends is given in this chapter.The summary of the framework is defined in this chapter, along with the difficulties. This section also discusses the inspiration for the job and the achievements made. The system's history and current state will be discussed in the following sections.

Chapter 2

Literature Review

2.1 Introduction

In recent years, we have seen enormous works in the field of sentiment analysis of social media data. There has been different methodology and idea for implementation for each work. In this section we will try to discuss some of the related works of past and present which have been helpful for our thesis.

2.2 Related Literature Review

In [1] Twitter data has been used to perform sentiment analysis. This study aims at developing sentiment analysis using lexicons and polarity in multiplication. Data is collected using Twitter API. The collected data is taken in form of text stored in the database and analyzed later. The findings of the study are in the form of a list of adjectives together with manual marking which is used as a lexicon and the implementation of a sentiment analysis framework. The steps includes-crawling data from Twitter API, tokenizing, slang removal, stopword removal, stemming,elevator which is useful for retrieving data and matching words with tweets. The limitation of this works is less accuracy as lexicon method is lesser accurate than machine learning techniques.

In [2] D Gurkhe et al. develop a model to train machine to extract the polarity (positive, negative, or neutral) of a social media dataset in relation to a query keyword.Using the following technique, this project proposes an approach for automatically classifying the sentiment of social media data: The training data is first fed into the Sentiment Analysis Engine, which uses a machine learning algorithm to learn. Following the completion of the learning process with competent precision, the computer begins accepting human social data in relation to the keyword.Although the findings are not sufficient, the initiative aims to mark neutral data that has not been experimented with in a substantial way in the past. With a larger volume of neutral data and a higher standard of neutral datasets, the results are bound to change.

Kaur et al. [3] did research on facebook comments to review and explore sentiments of users.This project demonstrates a new algorithm written in the Java programming language. The algorithm is applied to comments, and utility is measured using the algorithm's accuracy rate.

In [4] sentiment classification on Bangla textual content was done. The researchers used both classical and deep learning algorithms to build classifiers for many publicly accessible sentiment named datasets in this research. SVM and Random Forest are examples of classical algorithms in our research, while CNN, FastText, and transformerbased models are examples of deep learning algorithms. They used various annotated sentiment datasets consisting of Bangla information from social media for multiple domains to perform comparative experiments in this analysis. The dataset for the classification experiments was pre-processed.As compared to classical algorithms, the efficiency of deep learning algorithms is significantly better.

Sarkar et al. [5] for detecting sentiment polarity in Bengali tweets, proposed a method that employs supervised machine learning algorithms. Data cleaning and preprocessing, Attribute extraction, Model creation and classification are the phases in the proposed method.They used the classifiers Naive Bayes multinomial and SVM, which were trained with various combinations of features to find the best classifier with the best feature set. On the Bengali tweet dataset that used for tests, it was found that SVM classifiers trained with unigram and SentiWordNet features work best.

In [6] researchers perform text sentiment analysis based on LSTM technique.This paper proposes an improved RNN language model, LSTM, that successfully covers all background sequence information and outperforms traditional RNN. It is used to accomplish text emotional attribute multi-classification, and it describes text emotional attributes more reliably than a traditional RNN. The datasets were in Chinese and English languages.

Hassan et al. [7] showed sentiment analysis on Bangla and Romanized Bangla texts using deep learning approach.This paper presents a large textual dataset of both Bangla and Romanized Bangla texts that is the first of its kind, has been post-processed, has been multiple checked, and is ready for SA implementation and experiments.Furthermore, this dataset was evaluated using two types of loss functions - binary cross-entropy and categorical cross-entropy - in a Deep Recurrent model, specifically, Long Short Term Memory (LSTM), as well as some theoretical pre-training by using data from one validation to pre-train the other and vice versa. Finally, this study presents the findings as well as an interpretation.

In [8] the research deals with the sentiment analysis from a Machine Learning perspective for Facebook comments written and posted in Arabic (Modern Standard or Dialectal) language. This process starts with Facebook getting comments and writing them. Every comment is then labeled positive or negative. Features are then extracted from each comment, and pre-processed. A collection of features was made prior to classification, and finally an assessment phase allows to measure the performance of our Machine Learning process. The results of the experiments shows that the consistency of the classification model produced is based on the set of features designed with the combination of extraction and weighting schemes that they had experimented with.

In [9] a method was proposed for analyzing data from social networks to classify human behaviour. By crawling public data obtained from online social network users, a framework was developed for the collection and analysis of large data. The program will interpret user-post data in two different languages. It is a system for bilingual behavior analysis, which can categorize users based on their posts and comments in social networks. The limitation was that data was not processed correctly and data extraction process was slow.

In [10] researchers suggest methods for correctly classifying the sentiment label.They implement two methods: one approach is known as the Sentiment Classification (SCA) algorithm KNN and another is SVM. They also assess their success on the basis of actual tweets.In this article they concentrate on the division of optimistic and negative feelings of the tweets. In this work,the classifying feeling (SCA) algorithm works better than SVM.

In [11] Nabi et al.use Tf.Idf (date frequency-inverse document frequency) to obtain a better response and to achieve a more reliable outcome by extracting various features from a positive, negative or neutral word study, especially from a view of the Bangla text. In relation to the whole meaning, the author calculates the overall positive, negative or documentary value. There is sufficient instance and experiment to explain how this technique extracts feelings.The framework outlined in this paper offers a method for understanding the feeling or viewpoint and for extracting a uniqueness to improve the understanding of the feelings in Bangla.The limitation is Supervised methods cannot always be restricted, as labeled corporations are not always available. In addition, the data contain some noise.

2.3 Conclusion

This chapter contains detailed literature review related to our work. We came to know about many approaches and techniques of different researchers in the field of sentiment analysis. Each work has some limitations and future opportunity is there to make them more efficient.

Chapter 3

Methodology

3.1 Introduction

Sentiment analysis is a natural language processing technique for determining the positive, negative, or neutral nature of results. It is also known as Opinion Mining. Till this date various methodologies and techniques have been implemented for this. This section contains our proposed methodology as well as implementation.

3.2 Diagram/Overview of Framework

Figure 3.1 illustrates our system. First, we choose popular profiles at random, in this case celebrities and public figures. The feedback from some of their posts were then recovered. We collected comments in both English and Bangla from Facebook. Our job begins after we have gathered the feedback. For our upcoming sentiment analysis, we need to preprocess the information.

After that, we build a sentiment analysis model that can be used in both languages. A LSTM model was created. Since the model has been developed, a direction for testing and assessment has been established. We get a labelled dataset from the model, where we can find the sentiment score of each comment.

We then list the friends or supporters of that profile user based on the polarity score of their remarks. We receive a list of friends divided into three groups. In social media, their replies to a user's posts determine whether they are nice, supportive, or aggressive friends/followers.

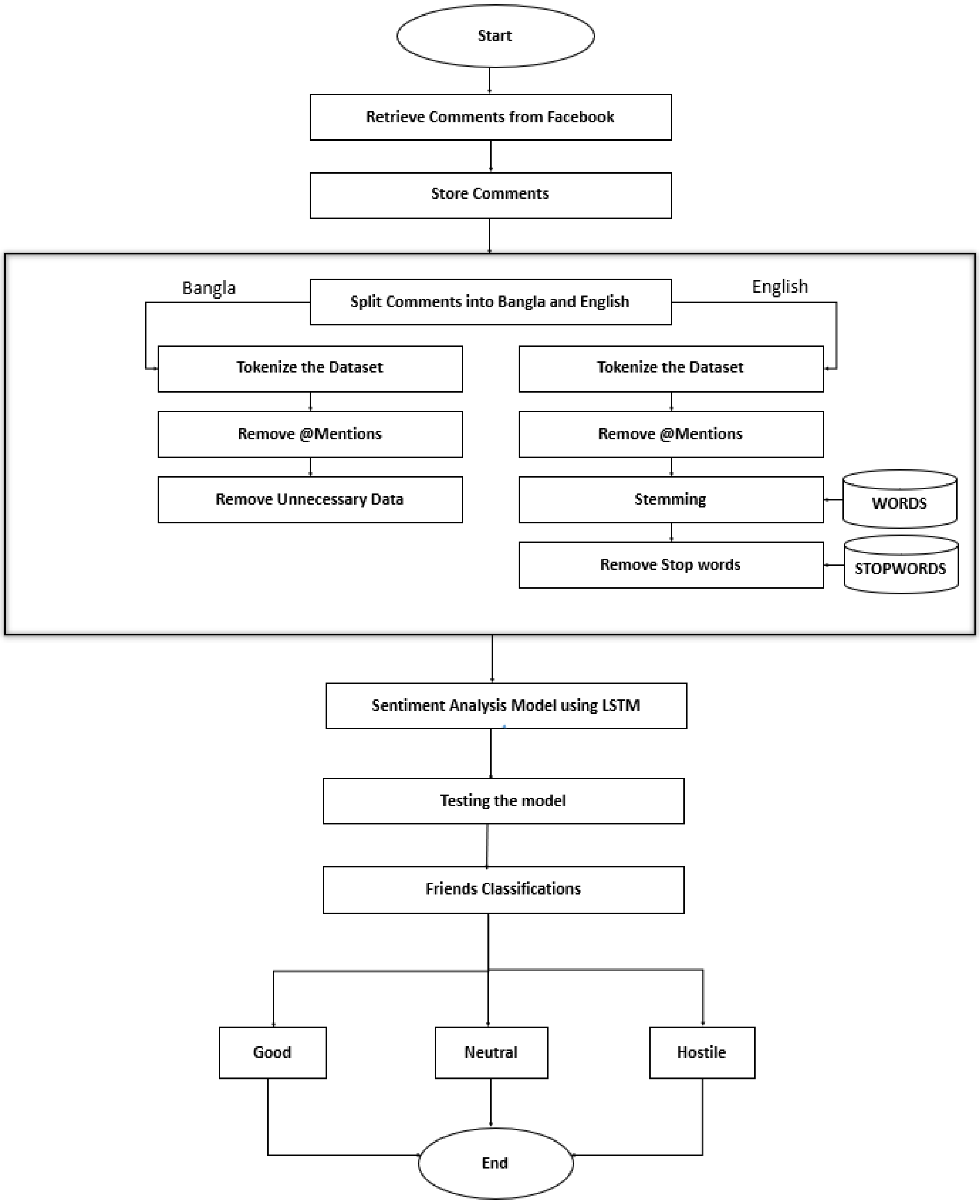

Figure 3.1: System Architecture of Proposed Framework

3.3 Detailed Explanation

3.3.1 Data Collection

Data is collected from social media along with username, id, comments, comment id, friends/follower id.Facebook is our primary data source. Bangla and English comments were collected from more than 30 user profiles and more then 3000 comments were collected manually and stored. Automated data retrieval was not possible due to Facebook's restriction for COVID-19.We keep track of our comments with labels.

3.3.2 Data Pre-processing

Data preprocessing is a data mining strategy that entails converting raw data into a format that can be understood. Real-world evidence is often unreliable, contradictory, and/or deficient in specific patterns or developments, as well as containing numerous errors. Preprocessing data is a tried and true way of addressing certain problems. At first the dataset is splited into two different datasets based on language. It was done using Unicode and ASCII code.

3.3.2.1 Tokenize the dataset

Tokenization is the process of breaking down a large piece of text into smaller tokens. Tokens may be words, characters, or subwords in this case.The most popular method of processing raw text is at the token stage, since tokens are the building blocks of Natural Language.Every word in a dataset is a token. We divide every comments into a token set by splitting it.

3.3.2.2 Remove @Mentions

We remove the words, which starts with '@' by using find method.

3.3.2.3 Remove Unnecessary Data

We remove the words which is not useful in sentiment analysis like numbers, emoji, white spaces etc.

3.3.2.4 Stemming

Stemming is a method for extracting affixes from terms in order to remove the base shape. It's the same as chopping a tree's roots down to the trunk. The stem of the words eating, eats, and eaten, for example, is eat.Porter stemming algorithm is applied to stem the words.

3.3.2.5 Remove Stopwords

The most opening words in any natural language are stopwords. These stopwords can not add much value to the context of the document when interpreting text data and constructing NLP models.Stopwords are omitted or deleted from assigned texts for purposes such as text sorting, where the text is to be classified into various groups, so that more attention can be given to the words that describe the text's context.There is a set of words which marked as Stopwords which will be removed in this stage. Like about, after, been etc. NLTK tool was used to remove stopwords.

Algorithm 1 Data Pre-Processing

Input: Comment Text Data
Output: Pre-Processed Text Data

```
begin ;
for each row in datafile do
    Drop rows which contains empty cells
end for
for each Comment in datafile do
    Apply tokenizer
end for
for each 'Token Sets' in datafile do
    Remove words, which start with "@"
end for
for each 'Token Sets' in datafile do
    Remove Non Alphabetic letters
end for
for each 'Token Sets' in datafile do
    Apply Stemming for each word
end for
for each 'Token Sets' in datafile do
    Remove Stopwords
end for
end
```

3.3.3 Sentiment Analysis Model

3.3.3.1 Sentiment Analysis Model Using LSTM

Sentiment Analysis may be a prescient demonstrating errand where the show is prepared to anticipate the extremity of literary information or opinions like Positive, Neural, and negative. It is a predictive modelling task where the model needs to be trained.Since we are as of now over-burden with parts of unstructured information it gets to be exceptionally intense to analyze the huge volume of literary information. But sentiment analysis can be exceptionally valuable.

Long-term dependency is captured by the LSTM(Long Short Time Memory), a kind of RNN network. They are also commonly used for a wide range of activities such as speech recognition, text classification, sentimental interpretation, and so on.The framework can detect long and short patterns in data and removes the disappearance gradient problem by training RNN. LSTM has been accepted in numerous applications and the language modeling course seems to be very promising as well.LSTM uses three doors, which are input gates, forget gates and exit gates, to check for the use and the updating of text history records. The memory cell and three gates enable LSTM to read, store and update historical information from a longer distance.

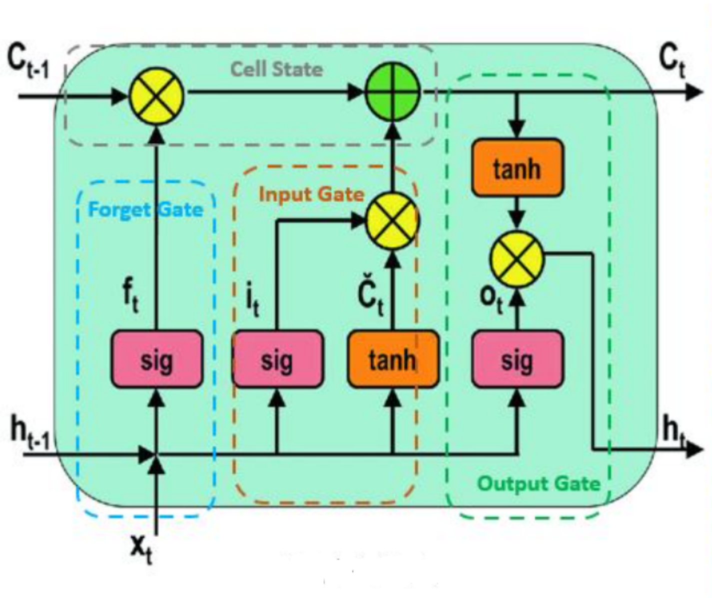

Figure 3.2: A LSTM Cell

The Tanh function is non-linear. It governs network flow values, with values from -1 to 1 maintained. A function whose second derivative can live longer is required to prevent information fading.Sigmoid is part of the non-linear function family. The door is used. In contrast to tanh, sigmoid preserves 0 to 1. It lets the network upgrade or forget details. The knowledge would be called forgotten if the multiplication is 0. The details would also remain if the meaning is 1.

The forget gate decides which facts should be considered and overlooked. The current X(t) input and hidden state h(t-1) information is transferred by the sigmoid function. Sigmoid produces values from 0 to 1. It concludes if the old output component is required (by giving the output closer to 1). This f(t) value is later used by the cell to multiply points by points.For the cell status check the input gate is performed following operations. The second sigmoid function is transferred into the present state X(t) and secret state h(t-1). The values between 0 (important) and 1 are transformed (not-important).

The next thing is to pass through the tanh mechanism the same knowledge about the secret condition and current state. The tanh operator will produce a vector (C (t)) with all possible values between -1 and 1 in order to control the network. The provided output values of the activation functions are ready to be multiplied point by point.The next move is to determine and store the new cell state details. Forget vector f(t) multiplied by the previous cell condition C(t-1) (t). If the result is 0, so cell status values are lowered.The output gate decides the next hidden state attribute. This state includes details of past inputs.

In our work for building a model for sentiment analysis we split our dataset into two different part named as Train and Test. The ratio of them is 80:20.

We used a LSTM sequential model with one embedding layer,two LSTM layer and one dense layer.

- The input layer contains the tokens that will be fed to embedding layer.
- The Embedding layer takes 128000 parameters and generate a (51,256) shape output. It enables us to convert each word into fixed length vector of

defined size. The fixed length of word vectors helps us to represent words in a better way along with reduced dimensions.

- Then the dropout layer start working with no parameters and generate a (51,256) shape output. The dropout rate is 0.3. It aims to reduce the complexity of the model with the goal to prevent overfitting.
- Now the 1st LSTM is activated for action with 525312 parameters and generate a (51,256) shape output.
- The 2nd LSTM layer activate with same size of parameters but generate a 256 size output.
- Adam optimizer is used to optimize the model and categorical_crossentropy is used as the loss function.A batch size of 32 is adopted with 20 epoch.
- Finally, in the output or dense layer we get three classes of output for the input data. This is in the form of probability. The classes are 0(positive), 1(negative) and -1(neutral).
- The target class will have the highest probability score.

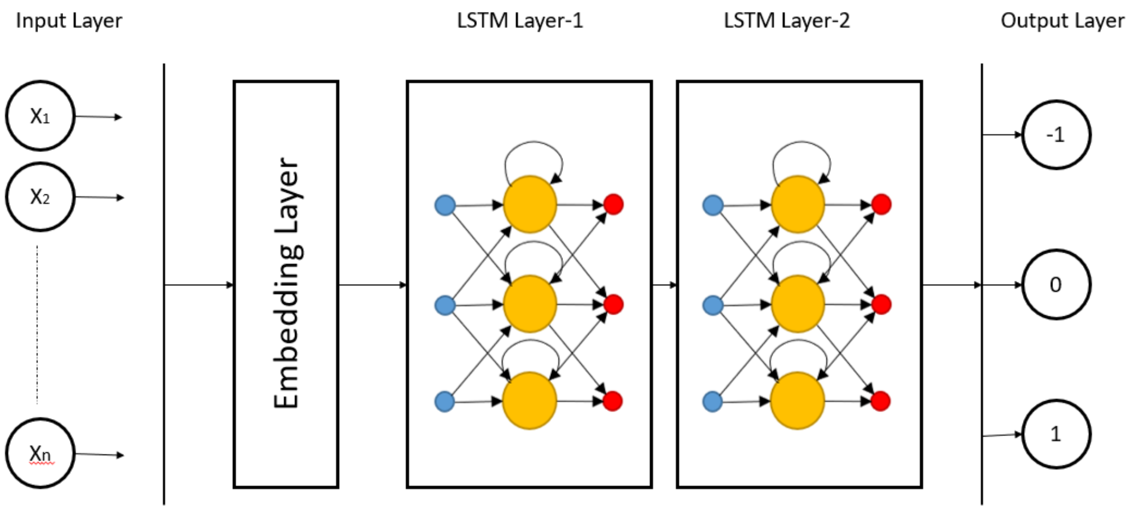

Figure 3.3: LSTM architecture for Sentiment Analysis

Algorithm 2 Sentiment Analysis Model

Input: Pre-Processed Labeled Text Data
Output: A Trained Machine Learning Model

begin ;
Define a Sequential Model ;
Add Embedding Layer-1 ;
Add Dropout Layer-1 ;
Add LSTM Layer-1 ;
Add LSTM Layer-2 ;
Add Dense Layer-1 ;
Compile model ;
Split Dataset into **Train** and **Test** ;
Fit the Model ;
Testing the Model with Test Data ;
end

3.3.4 Friends Classification

The model is constructed and we test to assess it. The following thing to do is classify the friends/followers based on their comments. Our model classify the comments into three categories- positive,negative and neutral comments.It assigns a score to each of the comments based on the sentiment. According to the polarity of comments, the persons are grouped into three class mainly Good friends, Hostile friends and Neutral friends.

Algorithm 3 Friends Classification

Input: Output Data File of Trained Model
Output: Classify Friends as Good, Hostile and Neutral

begin ;
Define Positive list, Negative list, Neutral list ;
for each row in Test Labeled DataFile **do**
    if label value = 0 **then**
        Update Positive_List **Add** Name of This Row in the List
        end
    else if label value = 1 **then**
        Update Negative_List **Add** Name of This Row in the List
        end
    else if label value = -1 **then**
        Update Neutral_List **Add** Name of This Row in the List
        end
        end
        Print Positive_List, Negative_list and Neutral_List values ;
        end

Friends with maximum number of positive comments will be considered as good friends, hostile friends will be the friends with maximum negative comments and neutral friends are there in between. In the algorithm Positive_list, Negative_list and Neutral_list indicates Good friends, Hostile friends and Neutral friends list respectively.

3.4 Conclusion

This chapters gives a detailed overview of our proposed method for sentiment analysis of social media comments and identification of friends category. We trained and tested the model with great effort. It was a challenging task to perform. In the upcoming chapter, the implementation and evaluation of this system will be discussed.

Chapter 4

Results and Discussions

4.1 Introduction

A thorough overview of the system was given in the previous chapter. The implementation and efficiency of the proposed architecture is examined in this chapter.This framework was implemented on a machine having Intel Core i3 processor and 8GB RAM. Keras with TensorFlow library is used. Python 3.6.9(version) is used for developing it.

4.2 Dataset Description

A proper dataset is needed for implementing our framework. But unfortunately there exists no standard dataset according to our requirement.On the other hand, we were unable to use Facebook API to collect data due to their restriction for COVID-19. So, we built our own dataset. We collected all the datas from Facebook manually and had to label them.

	A	B	C	D	E	F
1	UserProfile	PostID	FollowerName	FollowerID	Comment	Label
2	DonaldTrump	1.02E+16	Lori Penpek	1	you’re right again sir! I am astounded at how some of my neighbors in	0
3	DonaldTrump	1.02E+16	Walt Howard	2	Biden doesn’t have a snow-ball’s chance in hell. Continue to drain the	1
4	DonaldTrump	1.02E+16	Stephen Meggs	3	Dem mayors don’t let him stop it. He has stopped it everywhere it was	1
5	DonaldTrump	1.02E+16	Melissa N Jason Ha	4	Your presidency made me realize that I am a Republican and nothing	0
6	DonaldTrump	1.02E+16	Renato M Ochoa	5	Thank you, Mr. President! You are right, Sir! If these Democraps get int	1
7	DonaldTrump	1.02E+16	Terri Leo	6	Please make attacking a police officer a federal offense punishable by	1
8	DonaldTrump	1.02E+16	Tammy Mason Sho	7	We have to vote red people!!! We have to stand up and fight!!!!! For	-1
9	DonaldTrump	1.02E+16	John Howarth	8	Wake up America the Democrats are destroying this country and they	1
10	DonaldTrump	1.02E+16	Kenneth Collis	9	It’s not about whether Democrats or Republican leaders keep us safe..	0
11	DonaldTrump	1.02E+16	Jill North	10	I have bad anxiety about the things that are going on in the cities I hop	1
12	DonaldTrump	1.02E+16	Jean Siderio	11	I recently changed from independent to Republican because I will nev	-1
13	DonaldTrump	1.02E+16	Dominic Paulo	12	If you want America to look like Venezuela in 5 years then vote for the	0
14	DonaldTrump	1.02E+16	Susan Colangelo-g	13	It is horrible out there. And they're trying to blame you for the violenc	1
15	DonaldTrump	1.02E+16	Matt Neo Mortime	14	Too bad the suburbs are filled people who love and support the 2nd a	0
16	DonaldTrump	1.02E+16	Kurtis Barnes	15	Ya they are also starting fires everywhere on the west coast i believe th	-1
17	DonaldTrump	1.02E+16	Okech Derrick San	16	Nonsense.You've failed in your first term leadership, Biden is the only	0
18	DonaldTrump	1.02E+16	Bender Labrozzi	17	If Biden doesn't win they are still going to flood violence and destructi	1
19	DonaldTrump	1.02E+16	Aaron Michael	18	Let’s see that happen! My house has a safe zone sticker. It’s only safe f	0
20	DonaldTrump	1.02E+16	Bradley H. William	19	We will not allow it in the suburbs, we all own firearms and will defen	1
21	DonaldTrump	1.02E+16	BradnShel Smith	20	The left is committing acts of terrorism! Why can’t the US Government	1
22	DonaldTrump	1.02E+16	Jack Dos Santos	21	How can Biden say he’s all for the middle class, while he wants to dest	1
23	DonaldTrump	1.02E+16	Terry Hillig	22	Suburban voters are increasingly Democratic. Trying to scare them will	1
24	DonaldTrump	1.02E+16	Pradeep Raturi	23	Greatest president of my lifetime. Who is with me?!	0

Figure 4.1: Sample Input Dataset in English Language

	A	B	C	D	E	F
1	UserProfile	PostID	FollowerName	FollowerID	Comment	Label
2	DrShusamaReza	5.25762E+14	Arjina Jaman	1	বাস্তবসম্মত কথাগুলো তুলে ধরার জন্য ধন্যবাদ আপু।	0
3	DrShusamaReza	5.25762E+14	Jannatul Nabila	2	অসাধারণ..সময়পযোগী,যথাযথ!বারাকাল্লাহ	0
4	DrShusamaReza	5.25762E+14	Liton Chowdhury	3	সত্যিই অসাধারণ আপনার লেখনি এবং বিশ্লেষণ এর দক্ষতা।	0
5	DrShusamaReza	5.25762E+14	আরিশা ইবনে ইনায়া	4	অনেক ধন্যবাদ ম্যাম।এত সুন্দর করে গুছিয়ে লিখার জন্য!	0
6	DrShusamaReza	5.25762E+14	অনু আয়নাত	5	ধন্যবাদ ম্যাম।এত সুন্দর করে এত গুরুত্বপূর্ণ এবং সময়োপযোগী একটি বিষয় তু	0
7	DrShusamaReza	5.25762E+14	Tanzila Azizun Naher	6	ভীষণ ভালো লাগলো।শিক্ষণীয় পোস্ট।	0
8	DrShusamaReza	5.25762E+14	Muhtarima Hossain Sumiya	7	ইস! আমরা নিজেরাই নিজেদের কষ্টের কারণ হয়ে যাই।চোখের সামনে ঝুলিয়ে রা	1
9	DrShusamaReza	5.25762E+14	Mariam Patwary Chumki	8	মেম ভালো লাগলো পড়ে,আমার ও আছে এই সমস্যাটা।তবে পোস্ট টা পড়ে মন	0
10	DrShusamaReza	5.25762E+14	Rajia Sultana	9	এতো সুন্দর করে গুছিয়ে একদম মনের কথাগুলো কিভাবে বলেন আপু!!!!মহান	0
11	DrShusamaReza	5.25762E+14	Sheikh Md. Nabil	10	তুলনা আর নেগেটিভ চিন্তা ই সমাজে অপরাধঁ আর বিচ্ছেদ বাড়াচ্ছে...	1
12	DrShusamaReza	5.25762E+14	Nazmun Nahar	11	এজন্যইতো আপনাকে এত ভালোলাগে।	0
13	DrShusamaReza	5.25762E+14	Shamima Momtaz	12	কথাগুলো অসাধারন। আমরা যদি আমাদের চেয়ে যারা কম সুখে আছে তাদের স	0
14	DrShusamaReza	5.25762E+14	Shaila Ferdoush Bristy	13	আলহামদুলিল্লাহ আমি আমার পরিবার নিয়ে অনেক ভালো আছি.... মহান আল্লা	0
15	DrShusamaReza	5.25762E+14	Israt Jahan	14	তাই তুলনা হোক আগের আমি বনাম বতমান আমি-র;আমার বিবেক বনাম আমা	0
16	DrShusamaReza	5.25762E+14	Arefin Nahar Ava	15	কথা গুলো এতটাই সত্য যতটা সত্য জীবন বদলে দিতে পারে।অনেক ধন্যবাদ।	0
17	DrShusamaReza	5.25762E+14	Sajid Muhammad	16	অসাধারণ লেখেছেন আপু...,	0
18	DrShusamaReza	5.25762E+14	Md Zahirul Islam	17	অনেক ধন্যবাদ কথা গুলে মন ছুয়ে গেছে খুব ভালে লাগছে	0
19	DrShusamaReza	5.25762E+14	Sabina Yasmin	18	এতো চমৎকার একটি পোস্ট এর জন্য আপনাকে অনেক ধন্যবাদ আপু	0
20	DrShusamaReza	5.25762E+14	Munna Akanda	19	খুবই সত্যি ও বাস্তবসম্মত কথা।	0
21	DrShusamaReza	5.25762E+14	Sonali Das	20	এখন বদলে গেছে,,,,সময়ে সাতে নিজেকে বদলাতে হয়	0
22	DrShusamaReza	5.25762E+14	Ferdous Ara	21	দারুন !!! এ সময়ের জন্য অত্যন্ত প্রয়োজনীয় একটি বার্তা যার মধ্য দিয়েআমরা	0
23	DrShusamaReza	5.25762E+14	Jannatul Ferdous	22	আপু, সুন্দর লেখনির জন্য ধন্যবাদ।একদম ঠিক বলেছেন।আশা করি প্রায় ই এমন	0
24	DrShusamaReza	5.25762E+14	Eysmin Sathy	23	ধন্যবাদ এত গুছিয়ে লেখার জন্য। আসলেই আমরা অন্যদের সাথে নিজেদের তুল	0
25	DrShusamaReza	5.25762E+14	মোহাম্মদ তুষার ইমরান	24	ধন্যবাদ এমন কথাগুলো সুন্দরভাবে বোঝানোর জন্য।	0
26	DrShusamaReza	5.25762E+14	Mir Nasrin Liza	25	এক কথায় অসাধারণ। আললাহ সবাইকে এইভাবে বোঝার তৌফিক দান করুন।	0

Figure 4.2: Sample Input Dataset in Bangla Language

We collected almost 3000 comments in English and Bangla language from 30 user profiles. It is divided into 80:20 for train and test.

The datasets comprise of six columns: (1) User Profile (2) Post ID (3)Follower name (4) Follower ID (5) Comment and (6) Label of comment.

4.3 Implementation

The implementation of our Framework with the input dataset and the outputs will be discussed step by step in this segment.

- For this thesis a dataset is prepared by collecting comments on Facebook posts of different influential people and celebrities. The whole process is done manually due to limitation of Facebook API during COVID-19.
- Then the datasets are cleaned and we start our data pre-processing.

	Comment	Label
0	you're right again sir! I am astounded at how ...	0
1	Biden doesn't have a snow-ball's chance in hel...	1
2	Dem mayors don't let him stop it. He has stopp...	1
3	Your presidency made me realize that I am a Re...	0
4	Thank you, Mr. President! You are right, Sir! ...	1

Figure 4.3: Cleaned Data English

	Comment	Label
0	#mentions বাস্তবসম্মত কথাগুলো তুলে ধরার জন্য ...	0
1	অসাধারণ..সময়পযোগী,যথাযথ!বারাকাল্লাহ	0
2	সত্যিই অসাধারণ আপনার লেখনি এবং বিশ্লেষণ এর দক্...	0
3	অনেক ধন্যবাদ ম্যাম।এত সুন্দর করে গুছিয়ে লিখার ...	0
4	ধন্যবাদ ম্যাম।এত সুন্দর করে এত গুরুত্বপূর্ণ এব...	0

Figure 4.4: Cleaned data Bangla

- For both datasets removing @mentions, numbers, stopwords are done. Tokenization and Stemming is also accomplished to get a proper text which will further help in Sentiment Analysis. It is done in both languages.

	AfterStemming	text
0	[right, astound, neighbor, semi, rural, suburb...	right astound neighbor semi rural suburb even ...
1	[biden, snow, ball, chanc, hell, continu, drai...	biden snow ball chanc hell continu drain swamp...
2	[mayor, stop, stop, everywher, allow, democrat...	mayor stop stop everywher allow democrat parti...
3	[your, presid, made, realiz, republican, noth,...	your presid made realiz republican noth close ...
4	[thank, presid, right, democrap, presid, mix, ...	thank presid right democrap presid mix social ...

Figure 4.5: Data pre-processing for English

	AfterRemoveMentions	AfterTokenized
0	বাস্তবসম্মত কথাগুলো তুলে ধরার জন্য ধন্যবাদ আপু।	[বাস্তবসম্মত, কথাগুলো, তুলে, ধরার, জন্য, ধন্য...
1	অসাধারণ..সময়পযোগী,যথাযথ!বারাকাল্লাহ	[অসাধারণ..সময়পযোগী,যথাযথ!বারাকাল্লাহ]
2	সত্যিই অসাধারণ আপনার লেখনি এবং বিশ্লেষণ এর দক্...	[সত্যিই, অসাধারণ, আপনার, লেখনি, এবং, বিশ্লেষণ,...
3	অনেক ধন্যবাদ ম্যাম।এত সুন্দর করে গুছিয়ে লিখার ...	[অনেক, ধন্যবাদ, ম্যাম।এত, সুন্দর, করে, গুছিয়ে,...
4	ধন্যবাদ ম্যাম।এত সুন্দর করে এত গুরুত্বপূর্ণ এব...	[ধন্যবাদ, ম্যাম।এত, সুন্দর, করে, এত, গুরুত্বপূ...

Figure 4.6: Data pre-processing for Bangla

- A LSTM based Sentiment Analysis Model is built. A sequential model is initialized. One embedding layer, two LSTM layers, one dropout layer and a dense layer with three output are added with softmax activation function.We use categorical_crossentropy loss and Adam optimizer to train the model. Also, we set accuracy as the metric for measuring model's performance. For the training of LSTM neural network we use batch_size=32 and train the network for 20 epochs.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 1907, 256)         1280000
_________________________________________________________________
dropout (Dropout)            (None, 1907, 256)         0
_________________________________________________________________
lstm (LSTM)                  (None, 1907, 256)         525312
_________________________________________________________________
lstm_1 (LSTM)                (None, 256)               525312
_________________________________________________________________
dense (Dense)                (None, 3)                 771
=================================================================
Total params: 2,331,395
Trainable params: 2,331,395
Non-trainable params: 0
_________________________________________________________________
```

Figure 4.7: LSTM architecture(English)

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_1 (Embedding)      (None, 51, 256)           128000
_________________________________________________________________
dropout_1 (Dropout)          (None, 51, 256)           0
_________________________________________________________________
lstm_2 (LSTM)                (None, 51, 256)           525312
_________________________________________________________________
lstm_3 (LSTM)                (None, 256)               525312
_________________________________________________________________
dense_1 (Dense)              (None, 3)                 771
=================================================================
Total params: 1,179,395
Trainable params: 1,179,395
Non-trainable params: 0
```

Figure 4.8: LSTM architecture(Bangla)

- The last step is to categorize the followers based on their comments in posts of those user profiles. We classify them into three groups- (1) Good followers (2) Hostile followers and (3) Neutral followers.

  For Donald Trump and Zunaid Ahmed Palak we get this following list of classified friends:

```
Positive Followers List:
['Lori Penpek', 'Melissa N Jason Hathaway', 'Kenneth Collis', 'Dominic Paulo', 'Matt Neo Mortimer', 'Okech Derrick Sandeh', 'Aaron Michael', 'Pradeep F
Negative Followers List:
['Walt Howard', 'Stephen Meggs', 'Renato M Ochoa', 'Terri Leo', 'John Howarth', 'Jill North', 'Susan Colangelo-galione', 'Bender Labrozzi', 'Bradley H.
Neutral Followers List:
['Tammy Mason Short', 'Jean Siderio', 'Kurtis Barnes', 'Katie Whittemore', 'Kelly Turner', 'Scott Wattenbarger', 'Justin Bayliss', 'Eric Franco', 'Sama
```

```
Good Followers List:
['MD Shek Rasel', 'AJ Ariyan Johan', 'Md Alauddin', 'Md Ariful', 'Rouf Khan', 'Abir Ahammed', 'Md. Ashraful Islam', 'শাহেদ আহমেদ', 'Atm Shamsul', 'Md
Hostile Followers List:
['Lotif Ali', 'Badrul Alam', 'স্বপ্নই তো মানুষের জীবন', 'MD Habibul Bashar', 'Sikder Portugal', 'Rezaul Molla Nur', 'Abu Bakkar Siddiqe', 'Mami Ninetyseve
Neutral Followers List:
['মোহাম্মদ মহিউদ্দিন', 'আমরা নতুন প্রজন্মের মুক্তিযোদ্ধা', 'Qushalla Roy', 'Safiq Islam', 'Rks Rasel', 'AJ Ariyan Johan', 'Rk Rashed', 'Monirul Islam Monir']
```

Figure 4.9: Classified Friends List

4.4 Impact Analysis

Our system has a significant social influence. We have seen a huge amount of people using social media in recent years. The app was used for many purposes according to Facebook, by 2.80 billion users. Some people use it as a corporate forum, some to communicate etc.

4.4.1 Social and Ethical Impact

Celebrities or popular social media users have hundreds and thousands of followers. In every post of them, there are thousands of comments. It is very time consuming as well as tiring to read all the comments and to know people's opinion about them. Most of the comments are done thinking of spreading hate and negativity. So, our user can identify the people who do this and he/she can take further steps to prevent it. Social media platform can be safe from any cynical person.

4.5 Evaluation of Framework

After establishing a system, an adequate assessment is required. It helps to find out how the framework works correctly.To evaluate the efficacy of our framework we considered few metrics.Measuring performance metrics can be specified accordingly –

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.1}$$

$$Recall = \frac{\text{TP}}{\text{TP + FN}} \tag{4.2}$$

The harmonic mean of precision recall is F1-score.

$$F1\_Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.3}$$

$$Accuracy = \frac{\text{No of correct predictions}}{\text{Total no of predictions}} * 100\% \tag{4.4}$$

We achieved the following accuracy and loss curve :

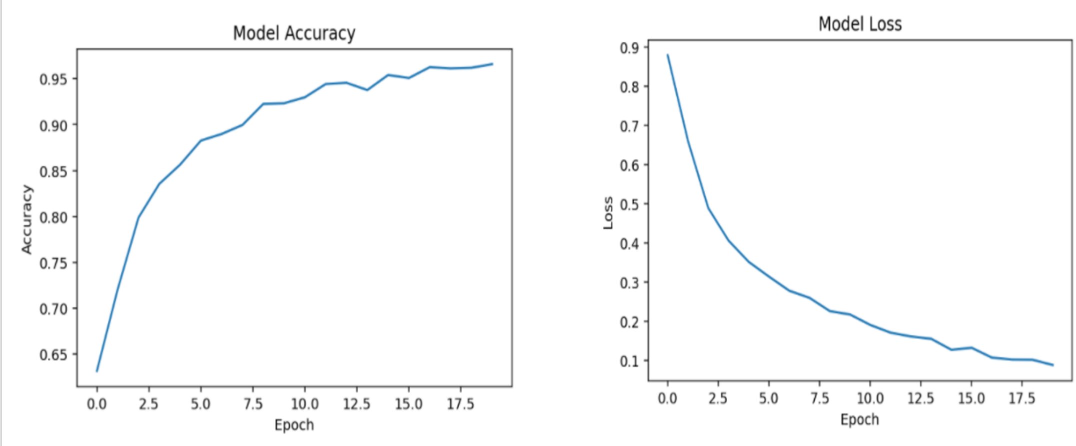

Figure 4.10: Accuracy and Loss curve(English)

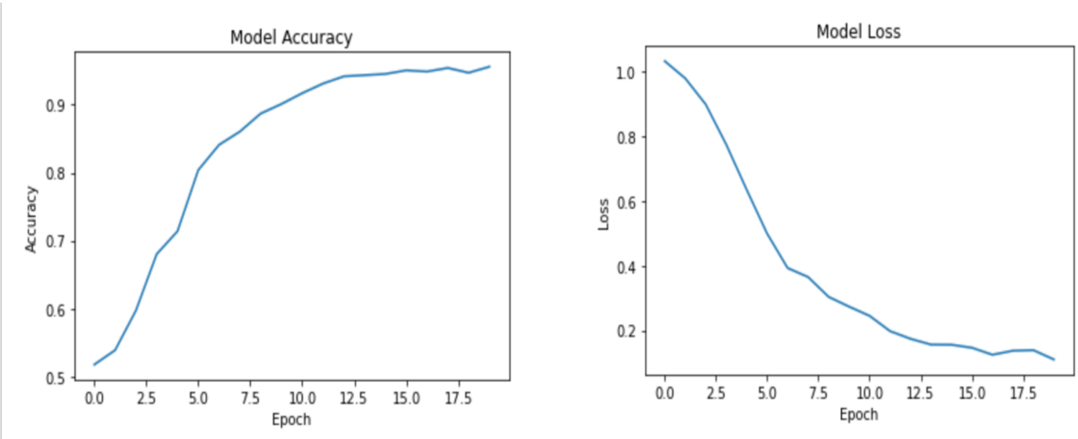

Figure 4.11: Accuracy and Loss curve(Bangla)

Table 4.1: Performance measure of Proposed Framework for English

English	Precision	Recall	F1 score
Neutral	0.89	0.99	0.94
Positive	1.00	0.98	0.99
Negative	0.98	0.98	0.98
Accuracy	0.98		

Table 4.2: Performance measure of Proposed Framework for Bangla

Bangla	Precision	Recall	F1 score
Neutral	0.97	0.95	0.96
Positive	0.97	1.00	0.98
Negative	0.99	0.95	0.97
Accuracy	0.98		

4.6 Comparison with other existing framework

The purpose of the proposed method is to perform sentiment analysis of Facebook comments and classify the users based on the score of their comment.It needs to be added that not much research works have been done on Facebook comments.On the other hand, friends classification depending on sentiment of comments is totally new in this field. There have been many sentiment analysis done on Twitter data(tweets),Movie review, Product review,Facebook posts i.e statuses. In [6] sentiment analysis was done using a LSTM and RNN network and text was classified into positive,negative and neutral. The accuracy is 89%. In [7] a Bangla and romanized Bangla dataset was developed and sentiment analysis is performed. The accuracy of that system is 78%. In that case, our system performs better having a 98% accuracy.

4.7 Conclusion

We have addressed the implementation and performance assessment of our proposed framework in this chapter.This chapter also discusses a comparison with other current systems in our method.In the upcoming chapter we discuss the future scopes of this research work.

Chapter 5

Conclusion

5.1 Conclusion

In this thesis work, our objective is to develop a system that can analyze comments of social media i.e Facebook of a paricular user and classify the follwers of him. The main purpose is to help a user to identify his/her good and hostile followers automatically. checking through thousands of comments everyday is not very effective. So, this system will come in handy. We retrieved the comments from different public figure profiles. Then a sentiment analysis model was created with an algorithm by us. After pre-processing the datas, they are fed to model and the model generated required score and result.Another algorithm was introduced to classify the friends on the basis of the score of their comments. Finally, we could classify the friends into our decided category. At the end, an adequate accuracy has been reached for this proposed method.

5.2 Future Work

While doing our thesis, we experienced many obstacles to achieve our goal. A proper dataset was missing. We had to build our own dataset. Moreover it was not labelled either. Therefore it was really challenging for us.There are some scope in future where we can make it a better system by changing certain things. The scope is discussed below:

- Automatic crawling of data from Facebook will help to achieve a large dataset. It will save time.
- If dataset is increased,the system will be more suitable for practical use and accuracy will be better.

- We have worked in English and Bangla language. But in social media there are transliteration comments also. In future, we can work on it to make the system more diverse.

References

[1] M. Mashuri *et al.*, 'Sentiment analysis in twitter using lexicon based and polarity multiplication,' in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, IEEE, 2019, pp. 365–368 (cit. on p. 7).

[2] D. Gurkhe, N. Pal and R. Bhatia, 'Effective sentiment analysis of social media datasets using naive bayesian classification,' *International Journal of Computer Applications*, vol. 975, no. 8887, p. 99, 2014 (cit. on p. 7).

[3] R. Kaur, H. Singh and G. Gupta, 'Sentimental analysis on facebook comments using data mining technique,' *International Journal of Computer Science and Mobile Computing*, vol. 8, no. 8, pp. 17–21, 2019 (cit. on p. 8).

[4] M. Hasan, J. Tajrin, S. A. Chowdhury, F. Alam *et al.*, 'Sentiment classification in bangla textual content: A comparative study,' *arXiv preprint arXiv:2011.10106*, 2020 (cit. on p. 8).

[5] K. Sarkar and M. Bhowmick, 'Sentiment polarity detection in bengali tweets using multinomial naıve bayes and support vector machines,' in *2017 IEEE Calcutta Conference (CALCON)*, IEEE, 2017, pp. 31–36 (cit. on p. 8).

[6] D. Li and J. Qian, 'Text sentiment analysis based on long short-term memory,' in *2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)*, IEEE, 2016, pp. 471–475 (cit. on pp. 8, 26).

[7] A. Hassan, M. R. Amin, A. K. Al Azad and N. Mohammed, 'Sentiment analysis on bangla and romanized bangla text using deep recurrent models,' in *2016 International Workshop on Computational Intelligence (IWCI)*, IEEE, 2016, pp. 51–56 (cit. on pp. 9, 26).

[8] A. Elouardighi, M. Maghfour, H. Hammia and F.-z. Aazi, 'A machine learning approach for sentiment analysis in the standard or dialectal arabic facebook comments,' in *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*, IEEE, 2017, pp. 1–8 (cit. on p. 9).

[9] K. T. Alam, S. M. M. Hossain and M. S. Arefin, 'Developing a framework for analyzing social networks to identify human behaviours,' in *2016 2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, IEEE, 2016, pp. 1–4 (cit. on p. 9).

[10] M. R. Huq, A. Ali and A. Rahman, 'Sentiment analysis on twitter data using knn and svm,' *Int J Adv Comput Sci Appl*, vol. 8, no. 6, pp. 19–25, 2017 (cit. on p. 10).

[11] M. M. Nabi, M. T. Altaf and S. Ismail, 'Detecting sentiment from bangla text using machine learning technique and feature analysis,' *International Journal of Computer Applications*, vol. 153, no. 11, pp. 28–34, 2016 (cit. on p. 10).