

Cyberbullying Detection using Machine  
Learning from Social Media Comments  
in Bangla Language

by

Saikat Halder Tuhin

18301063

MD Touhidul Islam

18301106

MD. Tauhidul Islam

19101276

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science

Department of Computer Science and Engineering  
Brac University  
May 2022

© 2022. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

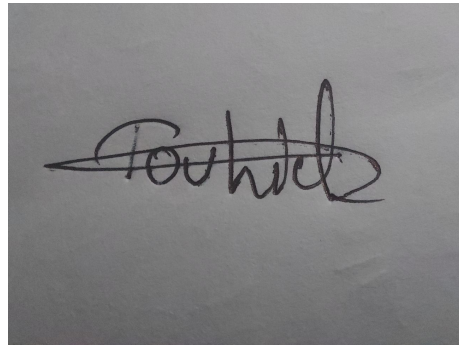
1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

**Student's Full Name & Signature:**



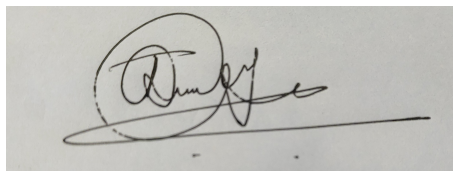
---

Saikat Halder Tuhin  
18301063



---

MD Touhidul Islam  
18301106



---

MD. Tauhidul Islam  
19101276

# Approval

The thesis/project titled “Cyberbullying Detection using Machine Learning from Social Media Comments in Bangla Language ” submitted by

1. Saikat Halder Tuhin(18301063)
2. MD Touhidul Islam(18301106)
3. MD. Tauhidul Islam(19101276)

Of Summer, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 26, 2022.

## Examining Committee:

Supervisor:  
(Member)



---

Mr. Tanvir Rahman  
Lecturer

Department of Computer Science and Engineering  
Brac University

Co-Supervisor:  
(Member)



---

Mr.Faisal Bin Ahsraf  
Lecturer

Department of Computer Science and Engineering  
Brac University

Head of Department:  
(Chair)

---

Sadia Hamid Kazi, PhD  
Chairperson and Associate Professor  
Department of Computer Science and Engineering  
Brac University

# Abstract

Cyberbullying which is defined as bullying perpetrated through the use of information and communication technology is a serious problem nowadays. As a result of the invention of social networks friendships through different social media, relationships, and social communications have all gone to a new level with new definitions. In fact, people become friends with someone whom he/she cannot even know face to face. With such a huge amount of users on the internet, cyberbullying has become a widespread global phenomenon. It not only makes a person mentally low but also has become one of the most important reasons for committing suicide. Being the seventh most speaking language in the world and increasing usage of the online platform, Bangla speaking people badly need an effective cyberbullying detection to handle this issue. In this thesis paper, we explore the spread of cyberbullying influence through the pairwise interactions between users. For cyberbullying through language, we will collect users' unique comments from social media and check them with the help of psychological references. After that, those comments will be categorized using Word embedding, an evaluation tool to categorize text, so that the dataset will be shortened and ready for classification. Lastly, the dataset will be to a machine learning classifier named Random Forest in detecting the cyberbullying comments. The performance and accuracy of numerous frequently used machine learning approaches on Bangla text are investigated in this study. In addition, the influence of user-specific information, such as location, age, gender, number of likes, number of comments, and so on, is examined for the identification of Bangla cyberbullying. Random Forest is the top effective algorithm for Bangla cyberbullying identification when just posts or comments are used to identify, according to experimental data, with 95.78% accuracy. Therefore, Random Forest is used for applying the approach on social media since it works better.

**Keywords:** Cyberbullying; Social Media; Suicide; Bangla Language; Word Embedding; Machine Learning; Random Forest;

## **Dedication**

Our work is dedicated to our parents, without whom we could never come this far in our lives. And a special Thanks to our supervisor who provided us their utmost support.

## **Acknowledgement**

Firstly, all praise to the Almighty Allah for whom our thesis have been completed without any major interruption.

Secondly, to our honourable advisor Mr. Tanvir Rahman and co-advisor Mr. Faisal Bin Ashraf sir for his kind support and advice in our work. He helped us whenever we needed help.

And finally to our beloved parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

# Table of Contents

<b>Declaration</b>	<b>i</b>
<b>Approval</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>Dedication</b>	<b>iv</b>
<b>Acknowledgment</b>	<b>v</b>
<b>Table of Contents</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>Nomenclature</b>	<b>x</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Research Problem . . . . .	2
1.3 Research Objectives . . . . .	2
<b>2 Related Work</b>	<b>4</b>
2.1 Cyberbullying . . . . .	4
2.2 ML . . . . .	7
2.3 Detection Models for Cyberbullying . . . . .	7
<b>3 Literature Review</b>	<b>8</b>
<b>4 Work Plan</b>	<b>12</b>
4.1 System Architecture . . . . .	12
4.2 Dataset Description . . . . .	16
4.3 Metrics and Evaluation . . . . .	18
<b>5 Dataset preparation and methodology</b>	<b>20</b>
5.1 Dataset Preprocessing . . . . .	20
5.1.1 Tokenization . . . . .	20
5.1.2 Text cleaning . . . . .	22
5.1.3 Word correction . . . . .	24

5.1.4	Zero Padding . . . . .	24
5.2	Feature Extraction and Feature Selection . . . . .	25
5.2.1	Word Embedding . . . . .	25
5.2.2	Word2Vec . . . . .	25
5.3	Ensemble Learning . . . . .	27
<b>6</b>	<b>Implementation</b>	<b>28</b>
6.1	Implementation of baseline model . . . . .	28
6.2	Individual Model Description . . . . .	29
6.2.1	SVM . . . . .	29
6.2.2	J48 . . . . .	30
6.2.3	KNN . . . . .	31
6.2.4	Random Forest . . . . .	31
<b>7</b>	<b>Result</b>	<b>35</b>
<b>8</b>	<b>Discussion and future scope</b>	<b>39</b>
<b>9</b>	<b>Conclusion</b>	<b>41</b>
	<b>Bibliography</b>	<b>44</b>



# List of Figures

2.1	Various Ways of Cyberbullying . . . . .	4
2.2	Cyberbullying Occurrence . . . . .	5
2.3	Cyberbullying issues . . . . .	5
2.4	Death Rate . . . . .	6
2.5	Death Rate . . . . .	6
4.1	Model Flow . . . . .	13
4.2	System Architecture . . . . .	13
4.3	Model Validation . . . . .	14
4.4	Flow Diagram . . . . .	15
4.5	Social Media Overview . . . . .	16
4.6	Victim's Profession . . . . .	17
4.7	Victim's Gender . . . . .	18
4.8	Percentage of Comments in Each Category . . . . .	18
5.1	Count vectorizer . . . . .	21
5.2	Bangla Word Cleaning Example . . . . .	24
5.3	Word embedding visualization . . . . .	25
5.4	CBOW and Skip-Gram . . . . .	26
5.5	CBOW and Skip-Gram Comparison . . . . .	26
6.1	Implementation Model . . . . .	28
6.2	SVM Vector System . . . . .	29
6.3	SVM Working Method (1) . . . . .	30
6.4	SVM Working Method (2) . . . . .	30
6.5	J48 Diagram . . . . .	30
6.6	Random Forest Process . . . . .	32
6.7	Random Forest Classifier . . . . .	34
7.1	Comparison of Different Outcomes . . . . .	35
7.2	Naïve Bayes . . . . .	36
7.3	J48 . . . . .	36
7.4	SVM . . . . .	36
7.5	KNN(1-nearest) . . . . .	37
7.6	KNN(3-nearest) . . . . .	37
7.7	Random Forest . . . . .	37
7.8	Comparison of Various Model Performance . . . . .	38
8.1	Non Bullying Data and Result . . . . .	39
8.2	Bullying Data and Result . . . . .	40

# List of Tables

4.1	Dataset's Variables . . . . .	16
7.1	Comparison of Different Methods . . . . .	35

# Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

*BLSTM* Bidirectional Long Short Time Memory

*CBOW* Continuous Bag of Words

*CNN* convolutional Neural Networks

*FPR* False Positive Rate

*GRU* Gated Recurrent Units

*HTML* HyperText Markup Language

*IDF* Inverse Document Frequency

*KNN* K-nearest Neighbor

*LSTM* Long Short-Term Memory

*ML* Machine Learning

*NCPC* National Crime Prevention Council

*NLP* Natural Language Processing

*OOV* Out of Vocabulary

*RBF* Radial Basis Function

*ROC* Receiver Operating Characteristics

*SOTA* State of The ART

*SVM* Support Vector Machine

*TF* Term Frequency

*TPR* True Positive Rate

*UNK* Unknown Word

*WEKA* Waikato Environment for Knowledge Analysis

*XML* Extensible Markup Language

# Chapter 1

## Introduction

### 1.1 Motivation

Nowadays communication through the internet is the most popular way to communicate with each other around the globe. In this modern era, we can chat and share our thoughts over networking sites. But we regret that, with the rising number of users, bullying is also being increased. Bullying through the internet or with the use of electronic communications is called cyberbullying which is now a great concern for us. In the past years, we did not look into this serious matter. A recent study in 2019 shows that in Dhaka, around The probability of women who are subjected to internet harassment are between the ages of 15 and 25 is 0.7, which is dangerous since it is causing a worse impact on teenagers. Cyberbullying for 18% of the harassment complaints and cases brought in front of the country's only cyber-crime tribunal [15]. According to a study by the National Crime Prevention Council (NCPC), Cyberbullying is when someone uses their phone, video game application, or other medium to transmit or post text, photos, or videos to purposefully injure or shame another person. Moreover, cyberbullying is also a hazard to teenagers' psychosocial well-being, according to research, as it causes symptoms of stress, despair, and anxiety that are more extreme than those seen in traditional bullying [22], [4]. Thus prevention of cyberbullying is so important for us and we need to handle it as soon as possible.

The issues in detecting cyberbullying include some steps such as when it comes to internet stages, finding obscene and indecent terms and sentences is a challenge and finding the aggressiveness and emotion behind those words. Several platforms and organizations are trying to stop cyberbullying by using various methods. Many researchers are trying their best to prevent it by using various algorithms to detect cyberbullying. Since cyberbullying should be treated and understood for various different purposes, many researchers have discovered that machine learning algorithms may correctly classify data and produce an accurate result. As a result, many researchers have chosen to utilize them in conjunction with natural language processing (NLP) approaches to detect cyberbullying.

Valuable information can be extracted from a dataset which is called predictive analytics, which can be used to classify the words. It also creates the provision for quantifying the relationship between the samples. By that, the relation between all the entities and attributes of the data can be described by predictable analysis which is essential in decision making. (Seigel, 2013). After using some steps to process the

data, we need to use some specific ML findings to detect cyberbullying.

## 1.2 Research Problem

Since communication through social media is increasing day by day, cyberbullying is significantly increasing too. Troll, harassment, racism, body shaming, spreading fake news with fake names, etc. are general forms of cyberbullying which victimize general people. Furthermore, this severe crime hampers the victim's life and even worse, provokes them to commit suicidal attempts.

Various recent reports on online portals and newspapers have already shown the worst side of cyberbullying. A piece of recent news from a renowned newspaper [29] showed that one of the most talented Bangladeshi actor Chanchol Chowdhury was severely bullied on his social media account for one of his personal pictures which he uploaded there. The comment section was flooded with bad comments and harassing texts. Again, a similar incident happened to another talented actress of Bangladesh, Ashna Habib Vabna where she posted a photo and get bullied over social media. Moreover, opening fake ID using others' photo and bullying others with the help of that is a generous problem nowadays which urgently need to be solved.

As young people use social media to convey various parts of their lives, their on-line safety is becoming a major worry. To raise cyber safety awareness among Bangladeshi social media users, it is necessary to collect and analyze statistical data on their cyber activities.

Though social media platforms use some sort of detection algorithm, it is very poor to detect. Moreover, this works if someone reports a particular comment. Thus, this research gives a strong study on automatic cyberbullying detection so that the comment detection and further process can be implied. Many types of research have already been done on this purpose and some of them had shown great results. Here, by using combination of algorithms to make a hybrid model for detecting cyberbullying more precisely.

Additionally, there are two methods to detect cyberbullying since bullying is not only done by text but also by photos. Here, this paper is focusing on the text based bullying by collecting it from the comment section of social media. Different researches use different data pre-processing methods such as tokenization, lowering text etc. After that, different deep learning/ ML algorithms are used to classify the dataset using J48, SVM, Binary Classification, N-gram etc. In this paper, we are going to use Word Embedding and to classify Random Forest classifier will be used.

## 1.3 Research Objectives

Cyberbullying is a dangerous and harmful behavior that can lead to suicide attempts or have a long-term detrimental impact on the victims. The researchers should develop a design for recognizing and preventing abusive behavior on digital platforms, based on the above-mentioned issues. The desire to develop a reliable multi-model detection system stem from users' objectionable conduct as it manifests itself across many social media sites. The majority of previous researchers in this

field have relied on supervised machine learning cluster algorithms, with researchers focusing on detecting cyberbullying on the basis of textual data. This study have gathered a review of groundbreaking research in the field of text-based cyberbullying detection. By evaluating findings from prior studies, the current effort intends to collect and debate research on the application of machine learning for cyberbullying detection. Focusing on reviews allows us to showcase the current debate and theoretical perspectives on the topic while also indicating the following points we need to answer:

- What are the highest used features to mechanized detection of cyberbullying?
- What are machine learning procedure (i.e. algorithms) and their evaluation method?
- What are the preventative implications of ML?
- What are drawbacks in predicting cyberbullying with machine learning breakthrough?

For detecting purpose, we will construct a cyberbullying detection system by using Word Embedding, an evaluation tool to categorize text. Furthermore, we will use a ML classifier named Random Forest in detecting the cyberbullying comments from the dataset. The objects of the research are:

- To deeply understand how cyberbullying occurs.
- To find the causes and effects of cyberbullying on victims.
- To deeply understand the automatic detection techniques.
- To have proper knowledge of various machine learning algorithms.
- To develop a model for detection.
- To gain knowledge and understand the difference and usage of the ML classifiers.
- To evaluate our model.
- To predict the accuracy rate of our model with respect to other research on this topic.
- To make a usable and reliable model to prevent cyberbullying.
- To offer further recommendations on improving our model.

# Chapter 2

## Related Work

### 2.1 Cyberbullying

According to Kowalski, Limber and Agatston (2012) bullying is a human behavioral pattern that is aggressive and is the reason for discomfort or harm to another person. The Internet is one of the most important inventions for this modern era, but like coins, it has some dark sides too, where cyberbullying is one of those. Cyberbullying is done by various social media through the internet by bad commenting, giving threats, body shaming, harassing sexually and religiously, etc. According to some studies cyberbullying is in fact more dangerous than traditional bullying. Taunting jokes and humiliation cause a severe impact on victims' minds. Moreover, giving threats, spreading false news, and harassing sexually by creating fake IDs using other's pictures is now a great matter of concern. Figure 2.1 shows various ways of cyberbullying.

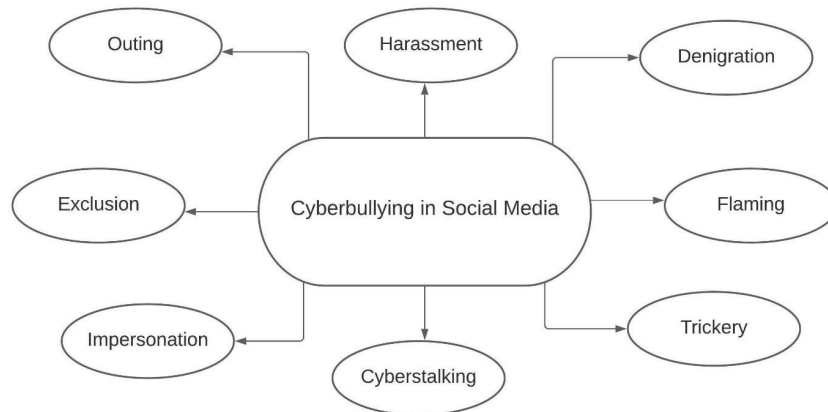


Figure 2.1: Various Ways of Cyberbullying

A recent (2021) study [30] in figure 2.2 shows the percentage of cyberbullying incident from various social media where Instagram and Facebook are the most severe platforms to bully others.

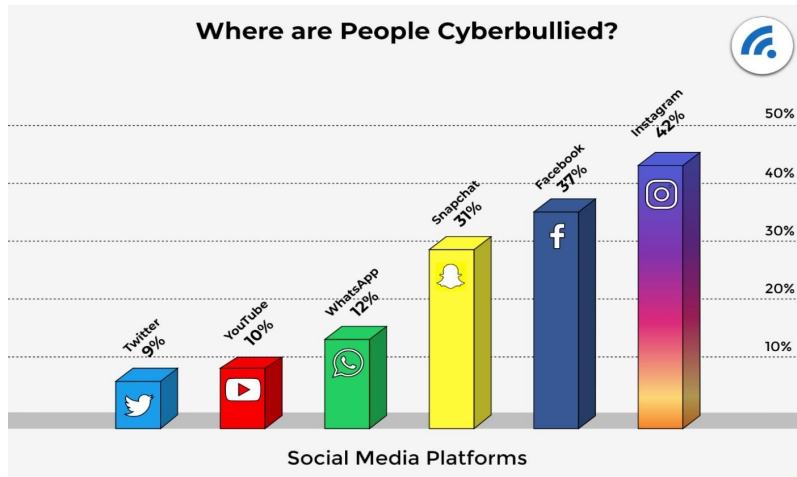


Figure 2.2: Cyberbullying Occurrence

Cyberbullying effects the normal life of a victim, in fact it increases the suicide rate. The below statistics in figure 2.3 [30] shows how it effects our kids:

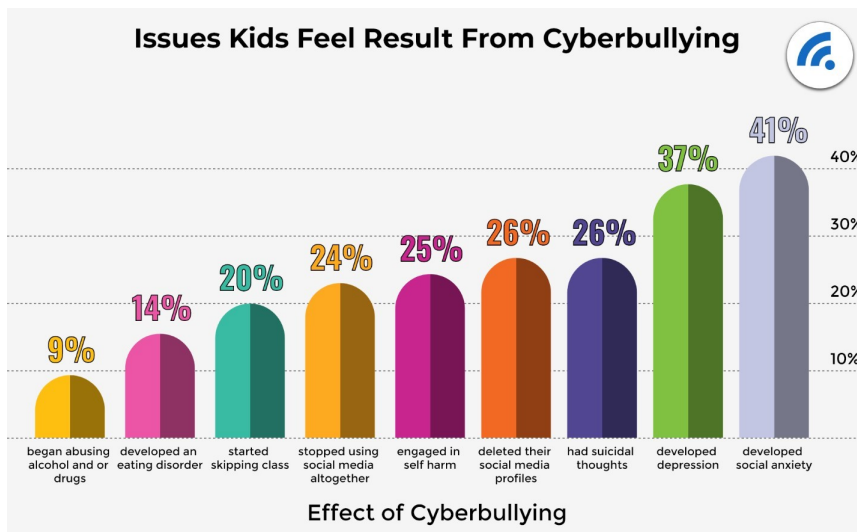


Figure 2.3: Cyberbullying issues



Another study in figure 2.4 [26] in US shows the suicide rate due to cyberbullying which is so horrifying and shower why it should be solved urgently:

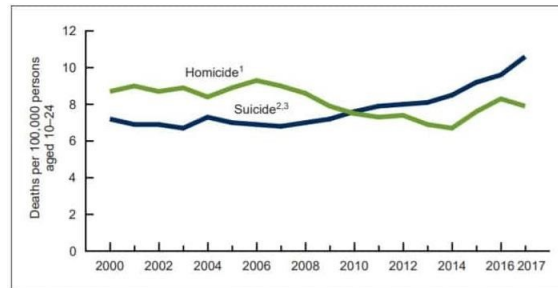


Figure 2.4: Death Rate

Furthermore, in our country, Bangladesh cyberbullying is now-a-days a great concern for the safety of teenagers and kids. The report from figure 2.5 [17] below shows the rate of cyberbullying happened to the school going kids in Bangladesh.

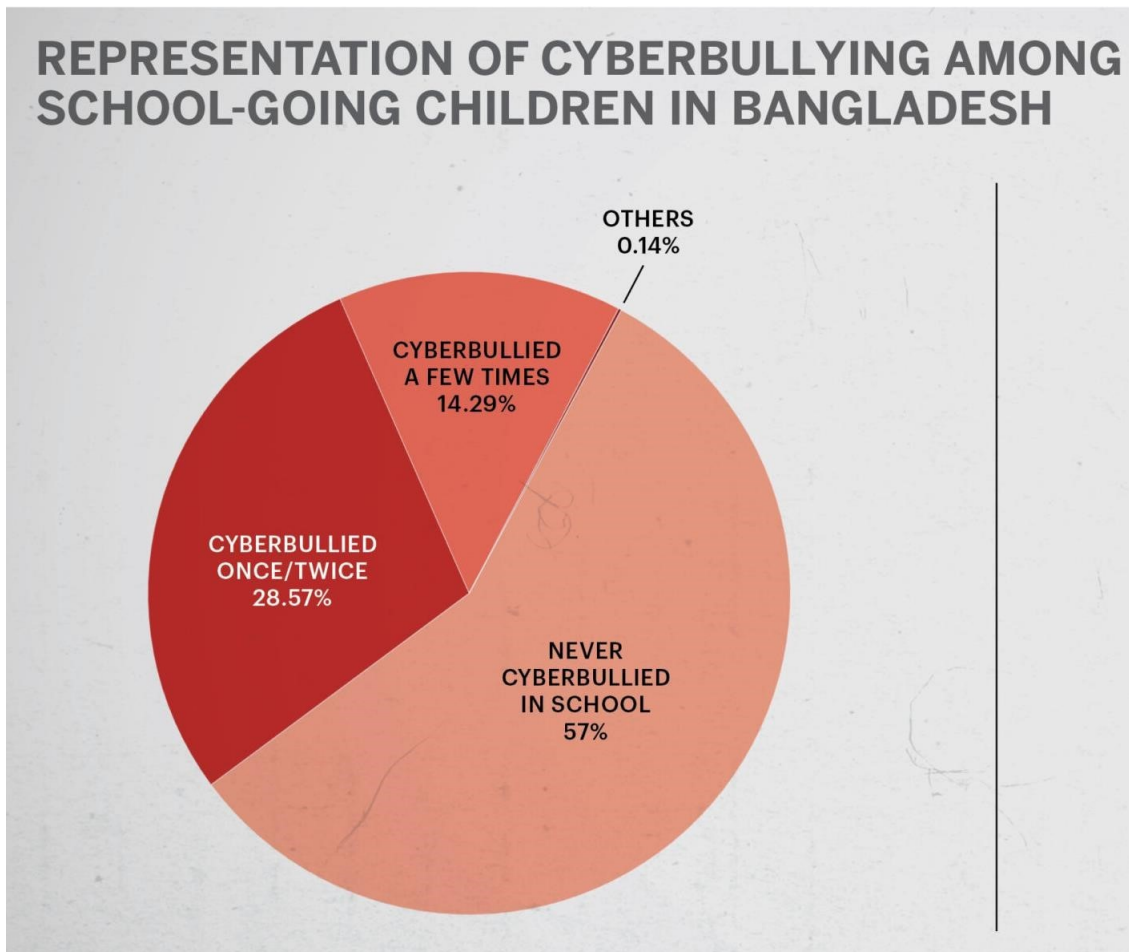


Figure 2.5: Death Rate

## 2.2 ML

ML is an AI which act as an application. It can give the feature of automatic learning. It is also able to help to find patterns. Machine learning makes use of preexisting techniques and datasets to create software applications that can solve problems. Its process begins with data observations, followed by the recognition of patterns in the data, and the making of advance judgments to be used in the future. This based on the previously recognized patterns. Machine learning helps computer to learn things by itself. Again, this application helps to adjust outcomes accordingly. By analyzing massive amounts of data, machine learning offers more accurate findings in a shorter amount of time. In today's industry, machine learning is being employed enthusiastically and actively.

## 2.3 Detection Models for Cyberbullying

There are so many tasks and researchers have already been done by ML algorithm to detect or predict cyberbullying actions. The classifiers of machine learning classify the content of text into 2 types; bullying and non-bullying. Deep learning algorithms with Neural Language Processing are also a growing algorithm where CNN is mostly used to classify. Since we are going to use machine learning, here a short description of machine learning is shown. ML have a lot of techniques which helps this type of application to learn from dataset. There are three techniques of machine learning are clustered by researchers: Supervised: The family of supervised machine learning algorithms is predicated on the availability of labeled continuous or categorical results [11]. In categorical outcome models, for example, if there is a pattern, the model learns it, by its own. After having some input data it will categorize data. Naive Bayes, KNN, Decision Trees, and random forests are famous as classification method.

Unsupervised: These machine learning techniques are based on the assumption that uncategorized data will be used to train. When outcomes aren't known ahead of time, these techniques are frequently employed to explain the data structure. K is well-known as a clustering algorithms, is an unsupervised ML algorithm.

Deep learning: The structure and functionality of biological brain networks are replicated by these algorithms [14]. Its aim to replicate cognitive learning, and perceptual mechanisms. Its like a human brain in artificial system which can do variety of tasks. The underlying structure of CNNs, for example, resembles the operation of the visual cortex of the person in order to accomplish tasks involving face identification [2] , [10] .

# Chapter 3

## Literature Review

The review of literature covers the summarized discussion on the prevalent issue of cyberbullying on social platforms. It continues by reviewing related works on the detection of cyberbullying. Many researchers researched on the topic of cyberbullying detection using machine learning or deep learning which approached across multiple social network platform like Facebook, Twitter, Instagram, YouTube etc. Though most of studies is based on text, few video or image based cyberbullying is also a matter of their concern. Among these studies, many has come up with better accuracy with the use of supervised learning algorithm, text-mining techniques for over a decade.

A model by collecting corpus by extracting online feeds to detect cyberbullying with a supervised learning approach was developed in 2009 by Yin. After collecting data, with the use of data dimension classification they trained the data on a support vector machine classifier based on different features. Yin et al. (2009) trained the classifier to find harassment which is based on the contents of feed, but unfortunately failed in analyzing the characteristic of the user behind posting these feeds. They used techniques of comprising the frequency of foul words, implications of Ngrams and lastly TF-IDF weighting. In NLP (Natural Language Processing), N-grams is defined as a contiguous sequence of ‘n’ number of syntactical characteristics which can be found in a segment of text. Furthermore, TF-IDF stands for Term Frequency-Inverse Document Frequency, is a numerical statistic that reflects how necessary a word is to a context in a collection of data. However, improvement has been shown for the results established on these baselines.

In another study, John et al. (2019) proposed approaches containing processing, feature extraction, and classification step in terms of detecting cyberbullying through machine learning [12]. They have used various steps in the processing part which include Tokenization, lowering texts, stopping word and encoding cleaning, and finally word correction. By extracting the polarity using the text Blob library, they also feature the data using the TF-IDF method. Moreover, for classification purposes, they used SVM (Support Vector Machine) and Neural Network which contains three layers called input, hidden, and output. They used different n-gram language models and run it several times to do experiments. To sum up, between these experiments, they found the highest accuracy rate (92.8%) using 3-gram on the NN.

In 2020 Saloni and Vidya presented an automated system where they included the notion of CNN implementation. Multiple layers are used in CNN to provide an accurate and efficient analysis. It is inspired by the examinations of the central

nervous system of mammals [23]. A class of neural networks consisting of a significant number of layers of neurons, which are capable of learning by themselves is termed as deep learning. Deep learning in general consists of 3 layers. Firstly, they used data pre-processing to clean the data using the vector model which accepts the vector form of data by making the data into the lowercase format and converting the data into an o vector. Secondly, for processing, they proposed the CNN model which includes sequential layers. Finally, the precise model prediction uses some parameters named epochs, batch size, and validation data. The overall study showed good theoretical results on the purpose of detecting cyberbullying.

Vimala et al (2019) took Twitter users and tried to find improved detection techniques in spite of detection cyberbullying [18]. They carefully studied users' personalities, sentiments, emotions, and various twitter features. After collecting a dataset from Twitter, they label users into 3 categories as Bully, Aggressor, and Spammer. Multiple experiments were conducted to ensure the effectiveness of the detection model based on the various features, namely, personality, sentiment, and emotion. The aspects were evaluated both individually and jointly with other aspects, and executed using Random Forest, J48, and Naive Bayes. Since Naive Bayes performed poorly during preliminary experimental analysis it was eliminated while Random Forest and J48 continued to perform well. Despite having some limitations, the study showed good results and wished to extend and examine for further models.

Dadvar and De Jong in 2012 proposed an approach that can incorporate users' information, characteristics, behavior behind the post, and contents of the conversation and can improve the accuracy of cyberbullying detection [6]. The envision of an algorithm that would go through the text and classify it whether it is bullying or non-bullying is shown there. They used a supervised learning approach to train a classifier for detecting harassment which is the SVM model in WEKA. To develop the model and train the classifier, they employ the TF-IDF value of profane words in each post. Furthermore, they split their dataset into male and female-authored posts or based on the age groups, into adult and teenager authored posts. Thus, the classifiers were trained separately for each group, and finally, they track their reactions using cross-system modeling. They then calculated the final result, based on the proportion of each group in the whole corpus (34% female, and 66% male). To evaluate the classification accuracy, they used 10-fold cross-validation and calculated corresponding precision, recall, and F-measure and found satisfying results in their study.

In another research paper, Ingle et al. in 2020 did research on detecting cyberbullying by using machine learning where they used the Twitter dataset. According to them, CNN yields a more precise output than SVM [28] and they were able to show the output of their study. The main purpose of their study is to find an optimal algorithm to detect cyberbullying using SVM which stands for Support Vector Machine and Convolutional Neural Network simply CNN. The end result shows that the SVM accuracy is around 0.86 which varies due to N-grams and in comparison, to it CNN gives 0.9 accuracy. The study reviewed the existing literature for various machine learning algorithms and they identified that SVM is the most efficient. The SVM model was compared with CNN where CNN performs with a better accuracy rate. To conclude, they had acknowledged that CNN helps to reduce the difficulty of explicitly selecting features and preserves word semantics which helps in providing better performance on complex data content as well.

In a recent study on cyberbullying detection from social media comments in the Bangla language (2020), Faisal et al. used Deep Neural Network to detect the corpus. They used a binary classification model which gave 87.91% accuracy [25]. After collecting comments from Facebook, the authors divide them into 5 classes according to Non-bully, Sexual, Troll, Religious, and threat. By removing bad characters, punctuation, etc. from the raw data, they pre-processed the information in order to fit it in their proposed model. They used a property of tensor flow, which is called 'Tokenizer' in order to put a value on the most frequently used words. By clustering the words together in a multidimensional vector space, word embedding was done successfully. Lastly, they used Binary classification and Multiclass classification to compare the end result and found that Binary classification gives better accuracy. However, they used multiple algorithms to find the accuracy rate such as SVM, RNN, Random Forest, and Naïve Bayes whereas SVM showed the most accurate rate.

In another recent study from 2020, Jacopo and Giulia discussed about how technology can prevent cyberbullying [20]. To find a solution they at first select ten articles and do research on those. After that they present some strategy and implications of ML for the automatic detection of cyberbullying. According to them as concerns algorithms such as Support Vector Machine (SVM), Naïve Bayes and Convolutional Neural Networks were shown to be the most performing algorithms, with SVM being the most efficient one. If the process is optimized perfectly then it can be a good assist to prevent the bullying.

Now-a-days, internet helping the world becoming a Global village. Especially, social media playing on of the important vital role, as 4.48 billion people using [27], and it is growing day by day. The authors [21] argue that neural network does not give the satisfactory result. And they tried to detect insulting words by using deep learning architecture. They took help from a dataset. The system will perform a preprocessing before categorization. They tested some deep learning models. However, suggest BLSTM (Bidirectional long short-term memory). They state that their proposed algorithm has more accuracy than others. Though, they accept that their proposed model has more computational complexity and also costly in terms of performance. The authors [19] showed various stages and multiple methodology systems, the first of which employs crowdsourcing for post and hashtag annotation and the second of which uses machine learning algorithms to discover further posts for annotation. This is legitimate study because we may compare our current results to those obtained from this paper. They came to the conclusion that if the dataset is trained using their method, the models will perform well. Meanwhile, their findings from various variate logistic regression techniques reveal that physical violence is connected with peers who smoke or with another aspect of carrying weapons prior to the cyberbullying enactment, as per [24], [16] physical violence is connected with peers who smoke or with another trait of carrying weapons before the cyberbullying enactment, according to their findings from various variate logistic regression approaches. So, if we want to increase cyberbullying prevention initiatives, they emphasize the need of ensuring positive and reliable monitoring by parents, teachers, or peers.

In [5] authors proposed to use the language-based method of detecting cyberbullying. They held a survey to understand the thinking of general people, and they recognize this as a subjective task. They focused on a number of bad words and the

density of those. They proposed that J48 is a machine learning algorithm. Vilas S. Chavan and Shylaja S. S. [7], proposed to classify the comments on social media as bullying and non-bullying with machine learning. They used N-grams to detect words. They also proposed a Chi-squared test for their feature matrix. Their proposed algorithm can only classify comments like a binary system. It cannot check the severity of bullying.

# Chapter 4

## Work Plan

### 4.1 System Architecture

This section gives the idea of the methods which we use to define cyberbullying and non-cyber bullying comments. Firstly, we collect the dataset while maintaining proper standards, we try to define the psychometric properties of the comments in some steps. Data collection is the process of acquiring information for the model's training and testing. This information might come from existing datasets or messages and comments from a specific web-based medium. For our research, we collected various comments from Facebook and preprocessed the dataset later. Since the data obtained may contain noise, the dataset have to be ready before the representation can be properly prepared. Data pre-processing removes hyperlinks, hashtags, usernames, as, what, who, and other special characters that are not necessary for the model's training. For identifying cyberbullying text, feature extraction is so important. Furthermore, in the feature selection stage, important features must be selected to increase the accuracy and reduce overfitting. This can be done using Filter, Wrapper, Embedded, and Hybrid methods. To send the data for the Random Forest classifier, we have used the word embedding process as feature selection. Lastly, the classification must be done using JRip, SVM, J48, naive Bayes, neural networks, etc. whereas we used the Random Forest classifier since it ensures the highest probability in case of identifying cyberbullying. In figure 4.1 and 4.2 the demo flow has been shown:



Figure 4.1: Model Flow

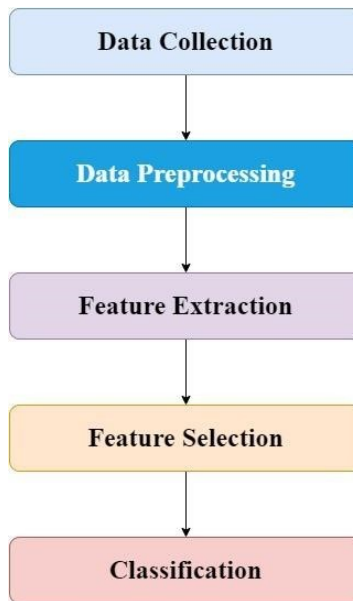


Figure 4.2: System Architecture



The purpose of the proposed cyberbullying detection model has been described earlier. In order to do that, we need a work plan according to which we have to move forward. The initial steps at figure 4.3 [9] are shown below:

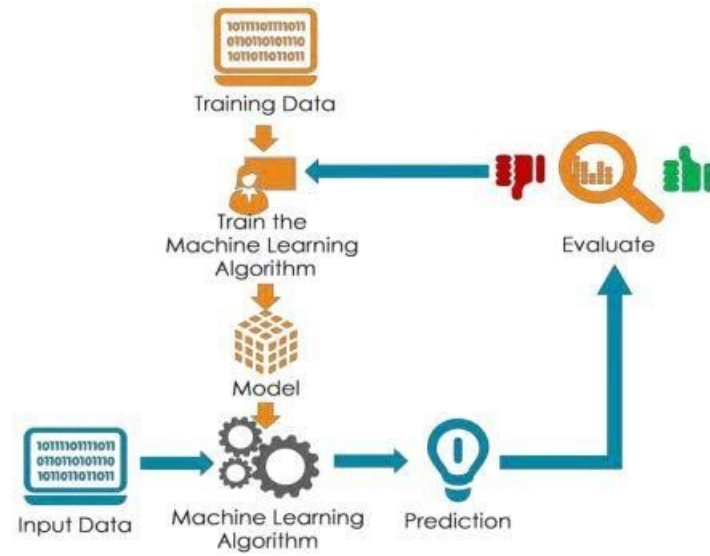


Figure 4.3: Model Validation

The work plan of this research has shortly described in the System flow diagram below in figure 4.4:

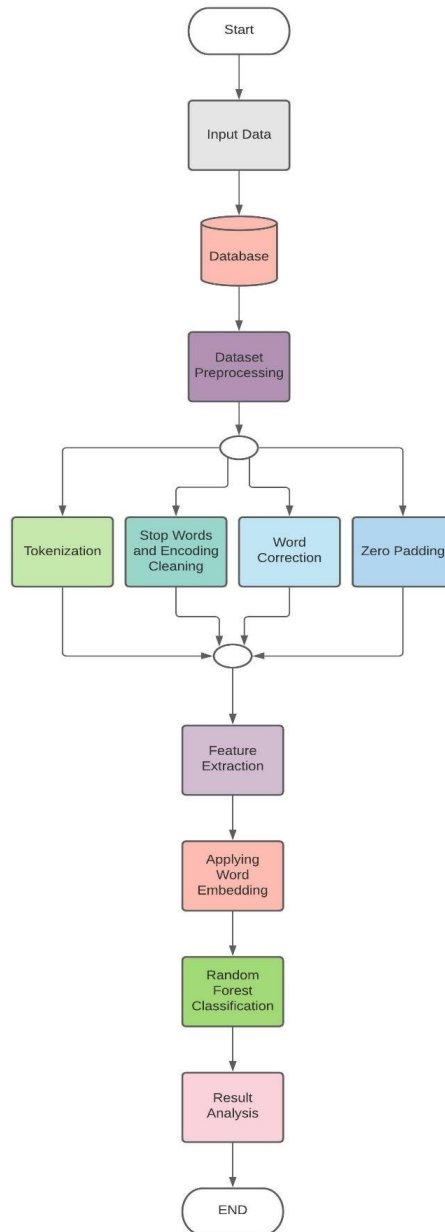


Figure 4.4: Flow Diagram

## 4.2 Dataset Description

Data collection is the way of acquiring and evaluating information on specific variables in a application. Its also mean to allowing one to answer question and guess outcomes. Sample data collection is a process where we collect data to use techniques on them. Data collection is the primary part of the research. It is a must for a researcher to use a trustworthy dataset in a research to have a good result. Building representative models for cyberbullying requires a proper dataset. Fig 4.5 shows how social media users in Bangladesh is rapidly increasing.

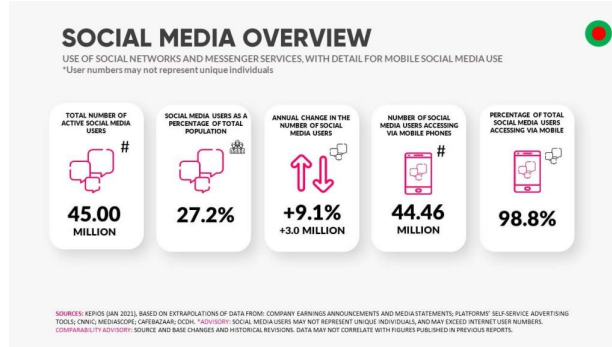


Figure 4.5: Social Media Overview

In this research we used a previously built dataset [25] to detect cyberbullying daily life internet users platform. This dataset was gathered from the comment section of actor, politicians, singers. And also famous sports person. 44001 comments were collected. The comments are in Bangla and from this dataset initially, we can see that women are getting more bullied than men. According to the data [29], 31.9% male, while 68.1% female. In addition, we can also see so many famous people were also victims of bullying. The variables that were used in the dataset are listed in Table 4.1. The dataset's labels are detailed further down: Non-Bully- Normally, these types of comments are not intended to attack anyone personally. For example, “আলহামদুলিল্লাহ। এ সজীবতা লিয়ে সুস্থ ও নিরাপদ থাকুন এ দায়ো করি ” Sexual- “তুই তাকে শালা \*\* ” Threat- “জুতা পেটা করা দরকার শালীবে” Troll- “বাংলার সানি লিওন” Religious- “নাস্তিক, ওরে সব জয়গাতেই বয়কট করা হাকে”

Variable	Description	Type
Comment	Comments which are collected from the users	Categorical
Category	Category	Categorical
Categorical	Victim's gender	Categorical
Number of react	Reaction numbers of that comment	Integer
Label	Harassment Type	Categorical

Table 4.1: Dataset's Variables

Fig 4.6 and fig 4.7 show the victim's profession and gender respectively.

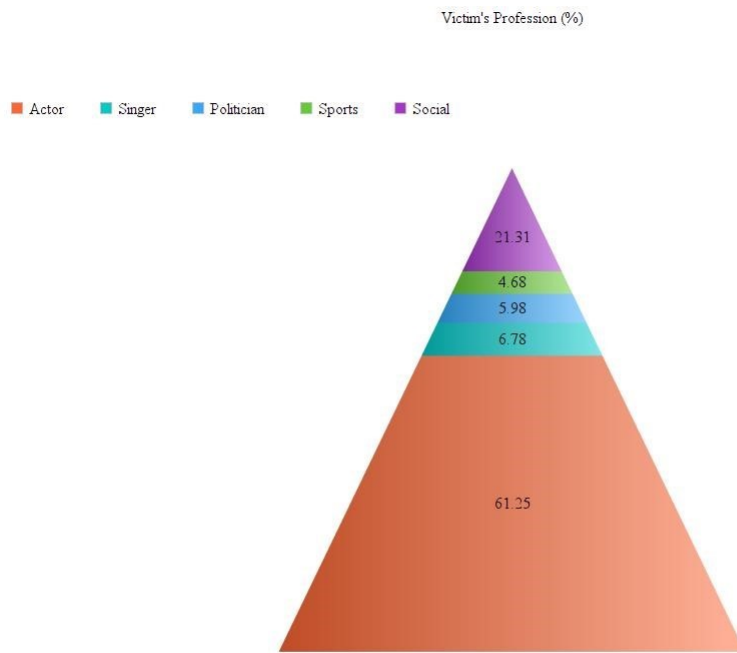


Figure 4.6: Victim's Profession

Fig 4.8 shows the percentage of comments in each individuals.

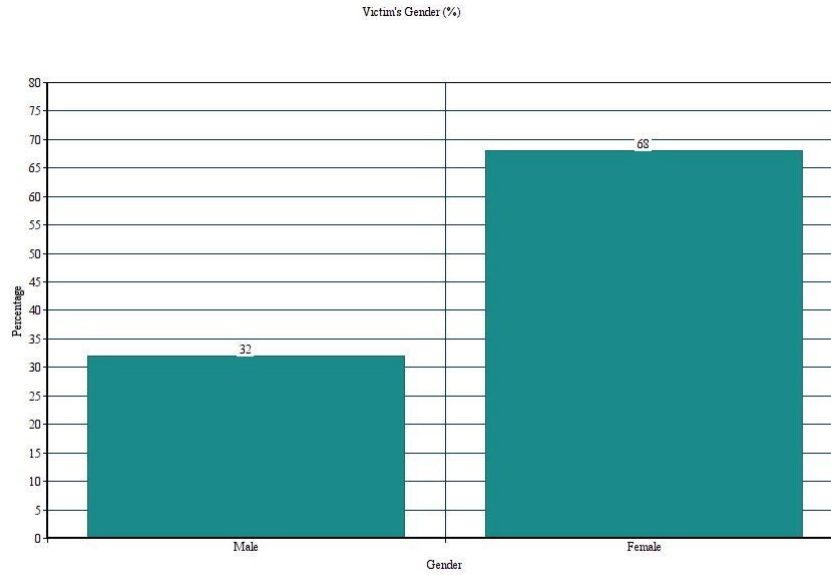


Figure 4.7: Victim's Gender

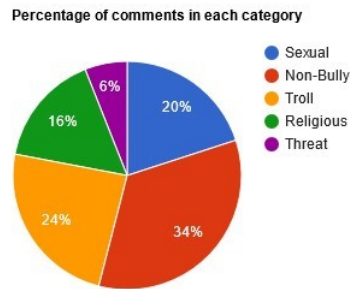


Figure 4.8: Percentage of Comments in Each Category

### 4.3 Metrics and Evaluation

In the process of classification models evaluating, we used training datasets. Training datasets were needed to train the model. Again, cross-validation was needed for k-fold method.  $t \in \{1, 2, k\}$

- Accuracy: It is the most straightforward performance statistic. It is just the proportion of correctly predicted events to all observations. Only when symmetric samples are obtainer with almost same values due to false positives and false negatives, accuracy is an important metric. The model's accuracy is calculated using the classification data acquired throughout each test phase as follows:

$$\text{Accuracy (\%)} = (c/n) * 100\%$$

- Precision: Out of the total number of instances provided under each tag: yes class and no class, the proportion of examples the classifier classified as right. In our example, the samples which belong to the positive class have been identified as bullied. Whereas, those which have not been recognized as bullied belong to the negative class. If precision is P and true positive and false positive are represented respectively as TP and FP, then:

$$P = TP / (TP + FP)$$

- Recall: The term "recall" refers to the proportion of accurately anticipated positive occurrences to all the observations in the class. If recall is R and true positive and false negative are represented respectively as TP and FN, then:  
 $R = TP / (TP + FN)$

- F1 Score: It is calculated by averaging Precision and Recall. It is a metric for how accurate a categorization system is.

$$F1 \text{ Score} = 2 * (R * P) / (R + P)$$

For binary and multiclass classification, F1 scores are utilized. The F1 score has a range of 0 to 1, with a higher number indicating greater prediction. The ROC curvature is a illustrative depiction of the classifier output's TPR and FPR. The F-1 Score obtained from the classifier's ROC curve, was used to choose the classification model which is best performing.

# Chapter 5

## Dataset preparation and methodology

### 5.1 Dataset Preprocessing

On Facebook millions of user from all over the world produce a large number of comments that are posted. Those comments are collected and stored in the dataset. Secondly, we cleaned up the unprocessed data after gathering all of the responses by eliminating incorrect symbols, grammar, and other mistakes before feeding it to our NN. The pre-processing phases for our dataset were split up into several segments: tokenization, stop words and encoding cleaning, word correction, and zero padding.

#### 5.1.1 Tokenization

Tokenization is an essential part of NLP. It is a critical level in not only classic NLP approaches like Count Vectorizer but also Advanced Deep Learning which is on the basis of designs like transformers. Tokens are the foundational elements of NL. It is a method of dividing a passage of text into narrower components known as tokens. It is in this context may be words, characters, or subwords. As a result, tokenization may be roughly categorized into three types: word, character, and subword (n-gram characters) tokenization. Consider the phrase "পটিকা মাছ". The most frequent method of creating tokens is based on space. Using space as a delimiter, the tokenization of the statement yields two tokens – পটিকা-মাছ. It is an illustration of Word tokenization since every token is a single word. Tokens, like characters, may be either characters or subwords. Consider the word

"অসাধারণ" for a moment:

- Character tokens: অ-সা-ধা-র-ণ
- Subword tokens: অ-সাধারণ

RNN accepts tokens and analyzes them at a distinct time step. As a result, the first step in modeling text data is to tokenize it. Tokenization is used to create tokens from the corpus. Following that, the upcoming tokens are utilized to construct a vocabulary. The term "vocabulary" refers to the collection of distinct tokens included inside the corpus. Bear in mind that vocabulary may be built by examining every one of the corpus's distinct tokens or by examining the highest K Most Commonly

Used Words. The vocabulary’s utilization in both traditional and advanced deep learning-based natural language processing approaches. After tokenizing, we got a large number of tokens, which consists of a lot of vocabulary.

- Vocabulary is used as a feature in traditional NLP algorithms such as the Count Vectorizer and the TF-IDF. Each word in the vocabulary is considered to have a distinct feature in figure 5.1:

	I	play	cricket	football	tennis
Doc 1	1	1	1	1	1
Doc 2	1	1	0	1	0
Doc 3	0	1	1	0	0
Doc 4	1	1	0	0	1

Figure 5.1: Count vectorizer

- Vocabulary is used to generate tokenized input phrases in Advanced Deep Learning-based NLP frameworks. Finally, the model is fed the tokens from these phrases.

## Word Tokenization

Word tokenization is used most frequently as a text token algorithm. It separates a string of words into separate words. The existence of a defined delimiter is required for these text chunks. Depending on the delimiters, new word based tokens are formed. Previously trained word embeddings, such as Word2Vec and GloVe, are used in word tokenization. One of the most perplexing aspects of word tokens is the handling of OOV terms. OOV words relate to new terms found during testing. These new terms do not exist in the English language. As a result, these approaches are incapable of processing OOV terms. Word tokenizers may be rescued from OOV terms using a simple method. The approach is to construct the vocabulary using the Top K Frequent Terms and to substitute unknown tokens for unusual words in the training data (UNK). This assists the model in learning how to represent OOV words in terms of UNK tokens. As a result, throughout the test period, every term not found in the lexicon will be mapped to a UNK token. This is one way to address the OOV issue with word tokenizers. The disadvantage of this technique is that we lose the complete meaning of the word when mapping OOV to UNK tokens. The word’s structure may aid inappropriately portraying the term. Additionally, each OOV word receives the similar delegation. Another difficulty with word tokens is related to vocabulary size. By and large, pre-trained models are trained on a vast corpus of text. Therefore, consider developing a vocabulary from all the unique terms included in such a vast corpus. This is why we need Character Tokenization.

## Character Tokenization

Conversion of a single line of text to a collection of characters is called Character tokenization. It eliminates the disadvantages of Word Tokenization discussed above. By maintaining the word’s information, Character Tokenizers handle OOV words



logically. It breaks the OOV word down into symbols and then conveys it using those characters. While character tokens address the OOV issue, the measurement of the input and output phrases quickly rises since each sentence is represented by a series of characters. As a consequence, it becomes difficult to understand the characters' relationships in order to make meaningful phrases. So, we need another process to fix all these problems we need Sub Word Tokenization. Subword Tokenization denotes the division of a passage of text into subwords. For instance, the word lower may be divided as low-er and the phrase smartest as smart-est. Transformed-based models — referred to as SOTA in natural language processing — prepare vocabulary using Sub Word Tokenization methods. The Byte Pair Encoding method is one of the most often used Sub-Word Tokenization algorithms.

## 5.1.2 Text cleaning

Text is a kind of dataset that has been existing throughout human history for millennia. Text is nothing more than a collection of words, or more accurately, a collection of characters [8]. However, when dealing with language modeling or NLP, we are often more interested in the words as a whole, rather than the character-level depth of our text data. One explanation for this is, that individual characters in language models lack a great deal of "context." While individual characters such as 'd', 'r', 'a', and 'e' lack context, when combined in the shape of a word, they may produce the word "read," which may describe some activity we're presumably performing right now. Vectorization is a technique for converting words to lengthy lists of numbers that may have some kind of sophisticated structure and are intended to be interpreted by a computer through a machine learning or data mining program. However, prior to that, we must run a series of operations on the text in order to "clean" it. The method of "cleaning" data varies according to the source of the data. The following are the primary phases in text data purification, along with their interpretations:

### Removing Unwanted Characters

This is a preliminary stage of the text cleaning method. If we scrape texts from HTML/XML sources, we must remove tags, HTML entities, punctuation, non-alphabets, and other non-standard characters. The most common techniques for this kind of cleaning include regular expressions, which may be used to colander away the majority of undesirable material. Certain systems maintain significant English characters such as full stops, question marks, and exclamation symbols. Consider the following scenario: we wish to do sentiment analysis on human-generated tweets and categorize them as extremely furious, angry, neutral, pleased, or very happy. Simple sentiment analysis may struggle to distinguish between a happy and a very joyful mood since there are certain instances that words cannot adequately describe. For example, two statements having the similar semantic sense can be compared: "This picture is awesome", and "This. Picture. Is. Awesome!!!!!!!!". The excess of punctuation, which suggests some form of "extra" experience, is the only indicator that the identical sentences transmit radically different emotions. Emoticons,

which are composed of non-alphabet characters, also contribute to sentiment analysis. ”:),:(, - -:D, xD”, all of them may contribute to a more accurate sentiment analysis when properly processed. Even if our goal is to create a system that can categorize whether a statement is sarcastic or not, such little subtleties might be beneficial. Apostrophes are a critical punctuation character that must be handled with care, since they may appear in a large amount of text. Terms such as 'আমার', 'তুমি', 'তুই', 'আমাদের' and 'তোমার' are entering internet papers like a sickness, and fortunately, we also have a treatment. There is a handy vocabulary of all these word contractions, which we may always refer to when converting apostrophe-containing nouns to their formal English equivalents separated by a space. For our Bangla language, we removed “|”, “,”, “-“ etc., and value them as per those individual meanings. For example, “ধন্যবাদ আপনাদের!!!!” has four exclamatory signs which sort of means extra emotion and expectation over some issue.

## Removing Stopwords

This cleaning stage is also dependent on what we want to do with the data once it has been preprocessed. Stopwords are words that are used so frequently that they have lost their semantic significance. Stopwords include words like ”of, are, the, it, is,” and ”of, are, the, it, is.” Eliminating stopwords can be a good idea in applications. Such as document search engines and document categorization. This is the thing where keywords are more essential than generic phrases, but stop words might be significant in applications like music lyrics search or search particular quotations. Consider phrases like ”To be or not to be,” ”Look what we made me do,” and so on. Stopwords serve a crucial function in these statements and should not be eliminated. There are two typical methods for deleting stopwords, both of which are extremely simple. The first method is to count all of the word occurrences, apply a termination value to the count, and eliminate all terms/words that occur more than the threshold number. Another option is to create a list of stopwords that may be eliminated from the tokens/tokenized sentences list. While, some human phrases might be useful, but we need a more formal style of application, these expressions may be eliminated. Words can be put together with no space between them in text data. Most social media hashtags are used, such as ”#GG, #DataScientist, #BlackFriday,” and so on. In Bangla some of them are '#ছাগল\_ইফষ্ট', '#তেলসিটি', '#মানবতার\_ফেরীওয়ালা'. Such words must be handled as well, and a simple method to do so is to separate them off based on capital letters, which is achievable if we kept the capitalization. If we do not want to keep the capitalization, this step should be done just before turning everything lowercase, during the tokenization process.

Bangla stopwords such that “ও”, “দিয়ে”, “আর” are handled carefully such that they can be removed as well their emotion and meaning can be declared properly. In Bangla literature, we have similar kinds of stop words as well like এবং, অথচ, অতএব অথবা, তাছাড়া etc. have found a list of this type of stop words and removed those from our dataset. By doing this, we had made the dataset smaller, more efficient, and accurate as it is spam less and more perfect to define cyberbullying words.

## Stemming and Lemmatization

Stemming and lemmatization have the purpose of reducing a word's inflectional and occasionally derivationally linked variants to an unified base form. As a result, stemming/lemmatizing aids in the reduction of the total number of words to a few "root" concepts. Organizer, organization, and organized can all be boiled down to a single word, perhaps "organization." Stemming is a simple approach of reducing sentences to their simplest form by simply defining criteria for chopping off particular letters at the end of the word, which provides sufficient results in most circumstances. Some examples are shown in figure 5.2.

Rule	Example
<u>ইহই</u> → ই	<u>যাইহইহই</u> → যাই
<u>উউউ</u> → উ	<u>আপউউউ</u> → আপ
<u>আআআ</u> →	<u>নাআআআ</u> → না
<u>গুনো গুনো</u> → <u>গুনো</u>	<u>ছাগনগুনো গুনো</u> → <u>ছাগন গুনো</u>

Figure 5.2: Bangla Word Cleaning Example

Lemmatization is a highly systematic version of stemming, but it also necessitates significant lexicon and anatomical examination. Because affixes of words hold extra details that can be exploited, it should only be employed when absolutely necessary. The words "ক্রততর" and "ক্রততম" For example, have the same origin but have different semantic meanings. So, If this program is just concerned with the term, since several browsers and content clustering methods are, it may be a viable option, but stemming and lemmatization may be eliminated for applications that need some semantic analysis.

### 5.1.3 Word correction

Word correction In this segment, we used a format of spell checking and corrects the word by providing a similar meaning [13]. Through this process, we can minimize the error and be able to find the cyberbullying words more correctly.

### 5.1.4 Zero Padding

By using NN, we require data to be homogeneous in size when feeding it for testing. For making the phrases into a homogeneous text sequence, we employed zero padding. Following the creation of sequences by the tokenizer, these sequences are handed to pad sequences to be padded in a similar pattern. As per consequence, a matrix was formed from a list of sentence, with every row having a maximum length of 120 characters. We had to place an appropriate amount of zeros following the text for short sentences.

## 5.2 Feature Extraction and Feature Selection

### 5.2.1 Word Embedding

Word embedding, a kind of ML method which helps to represent the word with meaning which can be compared. It helps to solve the Natural language processing problems. In word embedding, there are pre-determined vector spaces. These vector spaces have real-valued vectors. It helps to find individual words' real-valued vectors. In word embedding each word is considered as a single vector. The proposed value of that vector (word) resembles the neural network. It gives similar words similar types of representation as in vector space. Thus it captures the meaning of similar words more easily and frequently. An individual word has several kinds of expressions. They cannot be easily defined unless they are found expressively. The neural network has three-layer. To construct word embedding, a neural network is needed. Word embedding are widely used, because it provides a better solution. Let us make a list and discuss each of these uses. We used the Word2Vec [3] embedding model. In our model, we used 19469 vocabularies with an embedding dimension of 16. As a result, we may see our words primarily clustered on two opposing edges in figure 5.3.

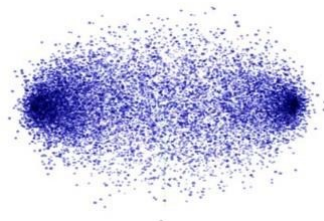


Figure 5.3: Word embedding visualization

Compute related words: Embedding indicates predicting the word. According to the prediction it suggests similar words. It also gives the suggestion of most used words which have similar kinds of meanings. Make a list of related words: It makes a semantic group. One group is for similar kinds of functionalities of words and the other group is for dissimilar functionalities. Text classification feature: Text is converted into vector arrays and input into the model for training and prediction. The string can not be used to train text-based classifier models; hence the text converts. There are many techniques available for word embedding such as embedding layer, word2vec, GloVe etc. In this paper, we will discuss Word2vec.

### 5.2.2 Word2Vec

Word2vec is a natural language processing method. It produces word embedding for better representation of the word. With the help of word embedding, we can solve many NLP problems. It helps to interpret the human language into a machine language and shows the works. You may think of them as a vectored text representation. Word2Vec produces word embedding. It can be used in many ways. Such as finding text similarity, recommendation systems, sentiment analysis, and more. Word2vec is a two-layer network that is better and more efficient. Word2vec shows spoken communication in vector area likeness. Neural networks don't learn documents or else they appreciate solely numbers. Word Implanting determines a

habit to change a text to a mathematical heading.

Word2vec rebuilds the linguistics process of the dispute. Concisely, the most objective of a sentence is a framework. Discussion or sentence encompassing verbalized or human communication (announcement) helps indecisive the intention of the framework. Word2vec learns the heading likeness of words through the circumstances. The example is shown in Figures 5.4 and 5.5.

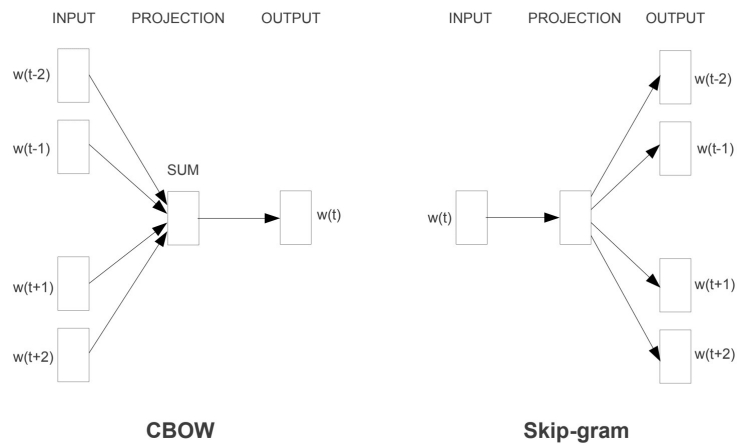
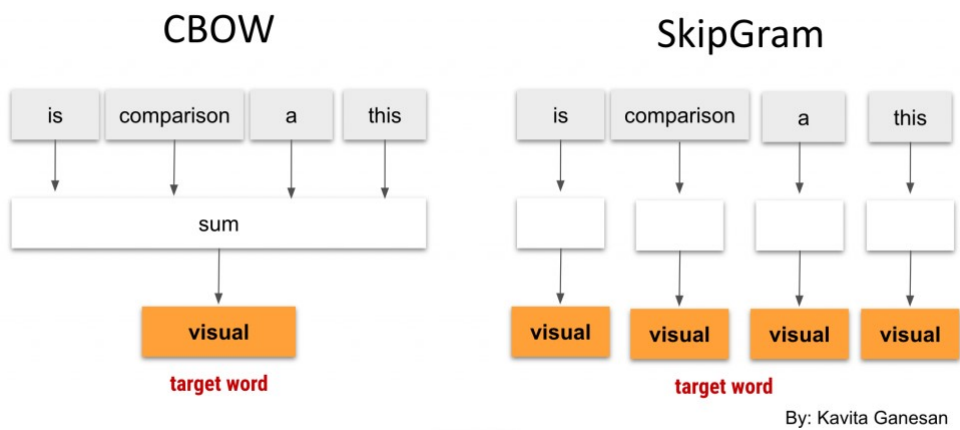


Figure 5.4: CBOW and Skip-Gram



This is a visual comparison

Figure 5.5: CBOW and Skip-Gram Comparison

A corpus can be represented as an N-dimensional vector, with each element in N representing a word in the corpus. We have a pair of target and context words during the training phase, and the input array will have 0 in all elements except the target word. The target word will have a value of one. The hidden layer will learn each word's embedding representation. And as a result, the representation will be in a d-dimensional embedding space. A dense layer with a SoftMax activation function serves as the output layer. The output layer will essentially produce a vector of a similar size as the input, with every element consisting of a probability. This probability represents the degree of similarity between the chosen word and the linked word in the sentence.

### 5.3 Ensemble Learning

Convolutional Neural Networks and related neural network models all suffer from variance, which is a problem with fitting. This is due to the fact that neural networks develop in a stochastic manner and are susceptible to the training data. As a result, even after a large number of epochs, the ultimate model could not be generalizable. As a result, the model will struggle to categorize inputs that are unfamiliar to it. Surprisingly, running fewer epochs scarcely helps because it increases the bias, which causes the model to perform miserably even in the training dataset. Running numerous models and integrating their results for a consistent and accurate forecast helps lessen the trade-off between variance and bias. Ensemble learning is another name for this method.

Several neural network architectures with various hyperparameters are performed in parallel in ensemble learning. An ensemble often consists of three, five, or seven models. Most notably, these models are packed in such a way that their forecasts are both extremely accurate and diversified. The ultimate output from these different models is determined using various approaches such as averaging, max voting, and so on. This ultimate result will always be superior to any of the individual models. For our proposed work, we will use a Random Forest classifier for maximum accuracy. We used a Max voting technique, which is quite popular for classification issues like ours, after creating numerous trees in order to reach the ultimate outcome from such an arrangement. This strategy examines each model's predictions and then focuses the final outcome on the label anticipated by the majority of the models.

# Chapter 6

## Implementation

### 6.1 Implementation of baseline model

This section explains how we used Python to implement our suggested model. After training, this model is preserved for use in our ensemble computing approach. We categorize the statement into two types whether it is bully or not bully and set this as goal. Figure 6.1 shows our binary classification neural network in action. We used a 1D convolution phase after incorporating the raw inputs. We utilized the activation function 'relu' and declare the parameters to 32, the size of the 1D convolution frame to 3, and the number of outcome filters to 32. Words will now be divided into three groups of three, then convolutions will be trained to map categorization to the appropriate output. The figure 6.1 shows the entire process.

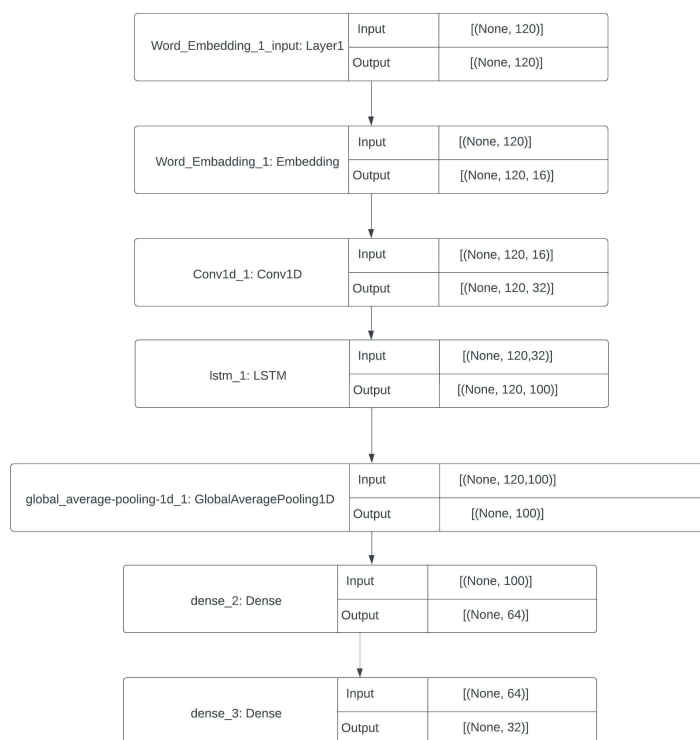


Figure 6.1: Implementation Model

Then, both for quicker training and better performance, we added an LSTM layer.

The number of outcomes in this layer was set to 100. We utilized a 0.2 dropout and recurring dropout rate to minimize overfitting. After this, we utilized the next layer as global mean pooling 1D to flatten the vector by averaging over it. The outputs were then input into a deep NN. Furthermore, some activation function are also used in this stage. Finally, because we are just categorizing two classes, we built the neural network with binary cross entropy. Lastly, we implemented our ensemble technique by importing random forest algorithm.

## 6.2 Individual Model Description

### 6.2.1 SVM

A supervised ML approach, the SVM, is used. This method is utilized in tasks like as classification and regression. It does, however, aid in categorisation. Every data item is treated as a point in an n-dimensional space, with n equaling the variety of features with a value. These values are the SVM algorithm's value for a certain position. Then we locate the hyperplane that best differentiates the two classes. Thus we complete classification (figure 6.2). Let's assume that there are two classes

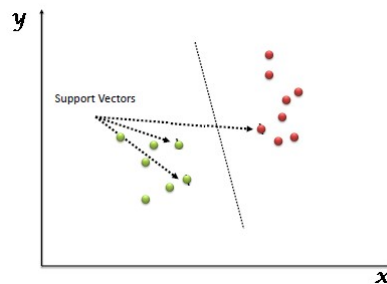


Figure 6.2: SVM Vector System

such as A: circle and B: Triangle. SVM considers all of the data points and produces a line called a 'Hyperplane,' which divides both groups. This line is known as the 'decision border.' Anything in circle class will belong to class A, and vice versa. There may be many hyperplanes visible, but the best hyperplane that separates the two classes is the hyperplane with a considerable distance from the hyperplane from both classes. SVM's main goal is to locate such optimum hyperplanes. Different dimensions are possible depending on the features we have. When there are more than three aspects, it is difficult to visualize. The methods are shown in figure 6.3 and 6.4.

The advantages of SVM are:

- performs in a clear margin of distinction
- When the amount of dimensions surpasses the quantity of data, SVM performs exceptionally well.



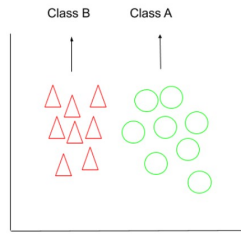


Figure 6.3: SVM Working Method (1)

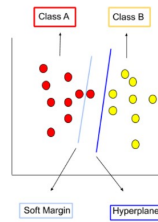


Figure 6.4: SVM Working Method (2)

- It needs a little amount of memory to perform.

The disadvantages of SVM are:

- For big datasets, the SVM approach is ineffective.
- If the amount of characteristics for each data point surpasses the amount of training data items, it will not function effectively.

## 6.2.2 J48

J48 is an algorithm that produces a decision tree. J48 has many features. Such as accounting for missing data, decision tree pruning, continuous attribute value ranges, rule generation, and other features. J48 supports categorization. With the help of decision trees or rules derived from them, J48 categorizes. This method constructs decision trees with the help of training datasets. It uses the idea of information entropy (figure 6.5).

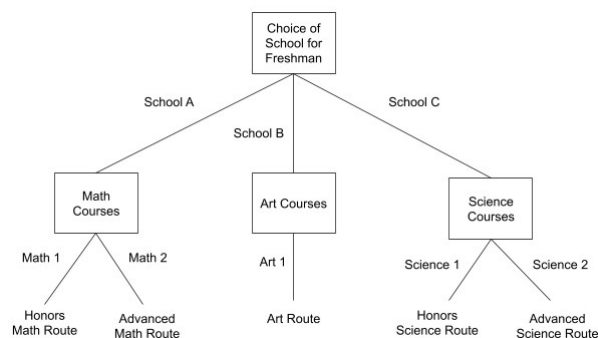


Figure 6.5: J48 Diagram

Some Advantages of Using J48 Algorithm

- This technique produces a list of all samples. These samples are in a given class.
- It assesses characteristics to find and gain any kind of information; if no gain is feasible, this technique adds a node further up in the tree using the class's predicted value.
- Although this algorithm finds an unknown class, it constructs a node higher up in the tree with the anticipated value.

#### Some Disadvantages of Using J48 Algorithm

- **Branches:** There are several rules for creating the J48 algorithm. One of them constructs trees with meaningful values. It will not help to create any classes. it makes the are wider and more convoluted.
- **Insignificant Branches:** A decision tree can be built with the same number of potential divisions as the number of distinct attributes chosen. These insignificant branches not only make decision trees less usable but also increase the risk of overfitting.

### 6.2.3 KNN

In this classifier, we have to decide on the number K of neighbors. Then we have to declare the Euclidean distance between K neighbors. Then we have to obtain a certain Euclidean distance. After that, with the help of Euclidean distance, we have to find the K's closest neighbors. Then we have to calculate the data points number in every section among these k neighbors. After that, we have to apply the new data points to the section with the greatest number of neighbors. Lastly, the model is now fulfilled.

The advantages of KNN:

- simple to use.
- larger training data this algorithm provides more effective results.

The disadvantages of KNN:

- Loss of storage.
- Slow prediction rat
- All of the training data is saved.
- The process becomes slower if the independent variable increases.

### 6.2.4 Random Forest

A good statistical learning model is Random Forest. It's a decision tree-based supervised machine learning method. A random forest is a combination of tree-based architectures that have been trained on arbitrary subsets of the training data, as the name indicates [1]. It's used in a lot of classification and regression software. To

create decision trees from the collected data, it employs supermajority for categorization and the mean for regression. It's a flexible, user-friendly machine learning approach that consistently generates outstanding results even when hyper parameters are not changed. It is also one of the most extensively used algorithms due to its simplicity and versatility. The Random Forest method has the privilege of being able to accommodate large datasets with both continuous and categorical variables, which is advantageous for regression and classification. It is based on ensemble learning, which is a method of integrating multiple classifiers to tackle a complex problem and increase the model's effectiveness. The random forest strategy, on the other hand, may be able to readily overcome the disadvantages of the decision tree algorithm. It enhances accuracy while reducing overfitting in datasets. It produces predictions without the need for a long number of package parameters. Random Forest process has been shown in figure 6.6.

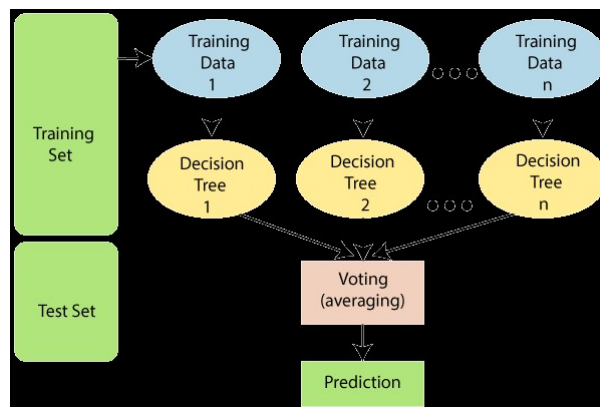


Figure 6.6: Random Forest Process

### Features of Random Forest algorithm

The Random Forest method outperforms the decision tree algorithm in terms of efficacy. It provides a realistic answer for data that is lacking. Without hyper-parameter tinkering, it can deliver a reasonable expectation. The flexibility of random forest is one of its most remarkable features. It may be used to solve regression and classification issues, and the relative importance of the input characteristics is obvious. Random forest is an additional effective approach since the default backpropagation it uses offer accurate predictions. Because there are fewer hyper parameters, they are easier to comprehend. It improves the model's predictability and speed. By deleting properties from each tree, it also saves space. The random forest relies on parallelization to allow each tree to be built individually based on its unique characteristics. Furthermore, we do not need to utilize a random forest to partition the sample for training and testing since this decision tree will never see more than 30% of the data. Because the overall average is used, it is more stable. It also improves the accurateness and reduces the problem of overfitting.

## Workflow of Random Forest

We can better understand random forest algorithms if we first understand judgment trees. The decision nodes, leaf nodes, and root nodes are the three sections of a decision tree. A training sample is split into branches using a classification tree approach, which are then further segmented. Until a leaf node has been found, the sequence will continue. The leaf node, on the other hand, is no longer detachable. The behavior of decision trees is extensively discussed in information theory. Knowledge acquisition and entropy are the two most crucial components of decision trees. As a result, a fundamental understanding of these characteristics will assist us in better understanding how decision trees are created. Uncertainty is calculated using the entropy measure. In this system, information gain refers to the process of learning about a target variable (class) through the usage of independent variables (features). The information gain is determined using the target attribute's (Y) entropy and the dependent entropy of Y. In this scenario, the conditional volatility is subtracted from the volatility of Y. Information collection is used in the training of judgment trees. It also helps to lower the level of ambiguity in these trees. A huge quantity of ambiguity has been decreased as a result of a large information gain. Splitting branches, a key event in decision tree growth, necessitates entropy and gain ratio.

An ensemble is a grouping of models. As a result, rather than using a single structure to create estimations, a collection of simulations is employed. Ensemble employs two sorts of methods: Baggage (which generates a different training subset with a substitute from a sample training dataset, and the final outcome is ascertained by majority voting) and the other one is Boosting (it transforms weak students into strong students by creating high-precision sequential models.). Random Forest makes use of the bagging concept.

## Bootstrap Aggregation

Bagging, or Bootstrap Aggregation, is a random forest clustering approach. Bagging is a method of storing an arbitrary dataset from the entire collection. As a consequence, in each model, row sampling is employed to replace the observations (Bootstrap Samples) provided by the Source Dataset. Bootstrapping is a term that refers to row sampling with replenishment. As a result, each model is now separately trained and produces outcomes. After all of the models' findings have been merged, the final conclusion is selected by majority vote. Aggregation is the procedure of amalgamate all of the facts and arriving at a conclusion based on a majority vote. Figure 6.7 implies Bootstrap Aggregation method.

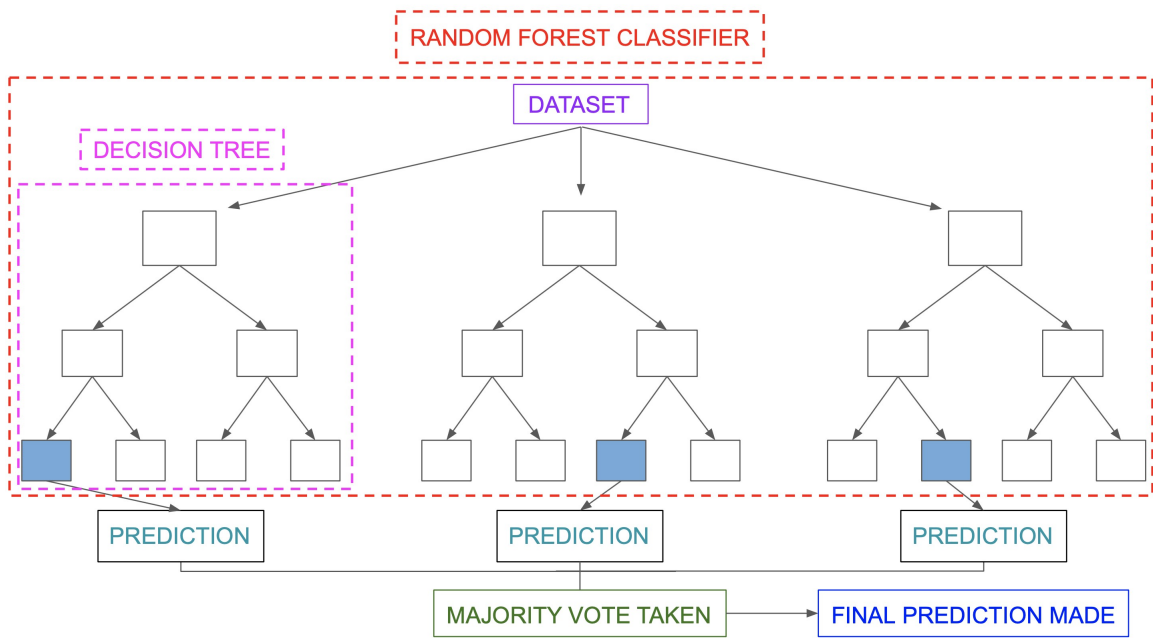


Figure 6.7: Random Forest Classifier

# Chapter 7

## Result

We have compared our result with other classifiers which has been shown in table 7.1. In figure 7.1 we have compared the classifiers outcomes with actual outcome. Figure 7.2, 7.3, 7.4, 7.5, 7.6, and 7.7 describes how precision, F1-Score, ROC, and accuracy differ respectively on Naïve Bayes, J48, SVM, KNN(1-Nearest), KNN(3-Nearest), and Random Forest.

Algorithm	Precision	F1-Score	ROC	Accuracy
Naïve Bayes	84%	67%	68%	59.36%
J48	79%	84%	51%	89.78%
SVM	91%	90%	71%	92.10%
KNN (1-Nearest)	88%	87%	61%	90.50%
KNN (3-nearest)	89%	87%	74%	90.74%
Random Forest	93%	92%	76%	95.78%

Table 7.1: Comparison of Different Methods

Example Sentence	Actual	J48	KNN	NB	SVM	RF
বোকা <span style="background-color: black; color: black;">████████</span> একটা	Y	Y	Y	Y	Y	Y
<span style="background-color: black; color: black;">████████</span> তামাশা শুরু করছস	Y	Y	Y	N	Y	Y
<span style="background-color: black; color: black;">████████</span> র বাচ্চা সাফা কবির	Y	Y	N	Y	N	Y
কি হবে এই পাসপোর্ট দিয়ে!	N	Y	Y	N	N	N
জয় কিছু ফালত কথা বলে আর <span style="background-color: black; color: black;">████████</span> উপস্থাপনা করে যা ক্যামেরার সামনে উচিত নয়!!	N	Y	Y	N	N	N

Figure 7.1: Comparison of Different Outcomes

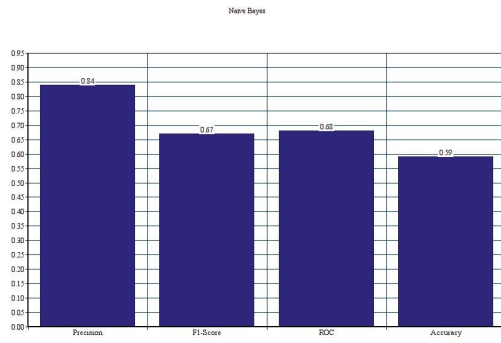


Figure 7.2: Naïve Bayes

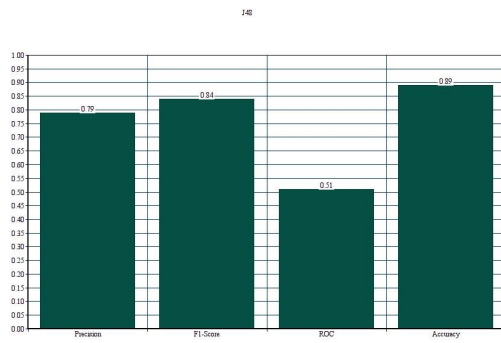


Figure 7.3: J48

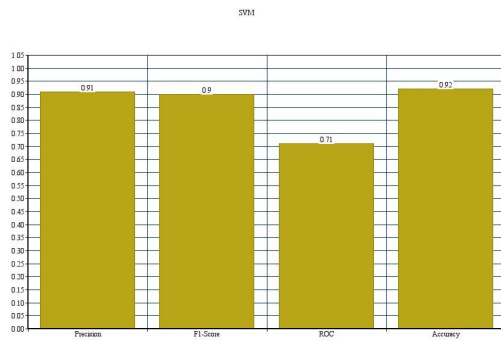


Figure 7.4: SVM

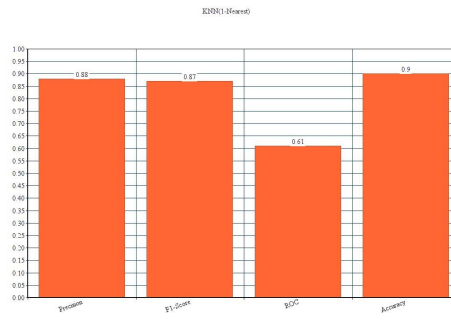


Figure 7.5: KNN(1-nearest)

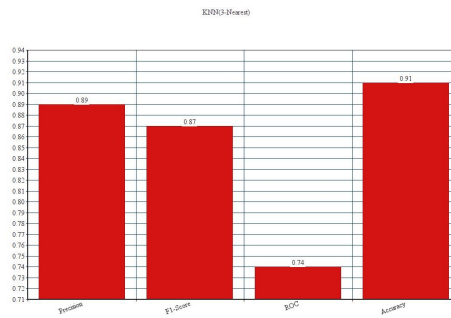


Figure 7.6: KNN(3-nearest)

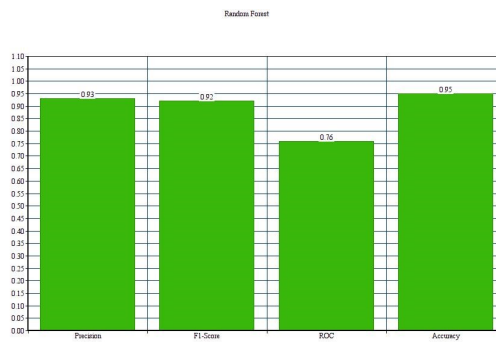


Figure 7.7: Random Forest



According to study which was previously done on our dataset, multiclass classification showed 79.29% accuracy [19]. 15 epochs were used and one and all took almost 261.4 seconds. We got 0.6210 as highest validation deprivation. This classifier model's F1-score of 76%, precision 81%, and recall 74%. This approach correctly predicts 85% of non-bullying comments. As well as 80% sexual, 48% threat. There were also 73% troll remarks and 82% religious.

For our dataset, after word embedding through Word2Vec, we pass the pre trained small dataset from the large repository to Random Forest classifier. It showed almost 96% accuracy. In addition to that, it has some limitations in terms of length, it shows better results than the other classifiers. The line chart in figure 7.8 shows the comparison with other classifier techniques.

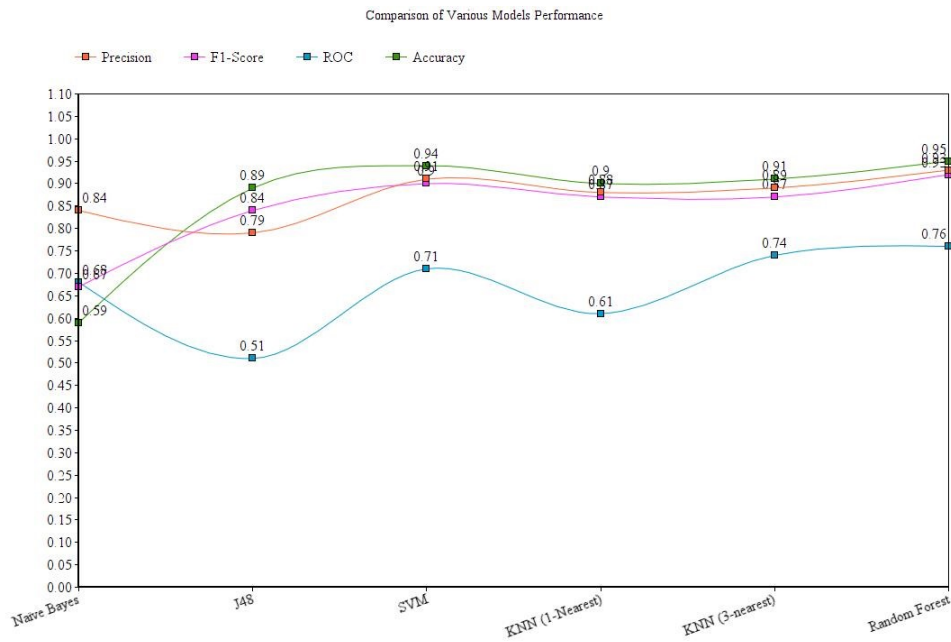


Figure 7.8: Comparison of Various Model Performance

# Chapter 8

## Discussion and future scope

After starting the system, we used social media dataset for training that has already been loaded. It generated two files:

- Non Bullying (which has been shown in figure 8.1 where N means non bully)
- Bullying (which has been shown in figure 8.2 where Y means bullying)

Data	Result
সাফা কি এখন পরকাল বিশ্বাস করে ?	N
আল্লাহ তোমাকে হেদায়েত দান করুক	N
মাশা আল্লাহ খব ভালো লিখেছেন ভাই	N
বসে বসে বকততায় কাজ হবে না।	N
আল্লাহ সুন্দর এবং উদার মনের এই দুটি প্রাণকে আপনি মানব কল্যাণের জন্য করল এবং নেক হায়াৎ দান করুন।	N
আপনার কাজগুলো অতলনীয় প্রিয়। আপনার গানগুলো একটু বেশিই ভালোবাসি। ভালোবাসা অভিরাম আপনার জন্য এবং আপনার কাজের জন্য।	N
চল গুলো একটু সরাইলে ক্রিকেটা সুন্দর হয়তো	N
জয় বঙ্গবন্ধু।	N
আপনাকে ধন্যবাদ আপনি আপনার ভুল বুঝতে পেরেছেন। আল্লাহ যেনো আপনাকে মাফ ও হেদায়েত দান করেন।	N
আল্লাহতালো আপনাকে অনেক উচ্চ মর্যাদা দিয়েছেন আরো বেশি মর্যাদা দেখ আপনার বয়স অনুযায়ী আপনি অনেক দানশীল আপনাকে খব ভালো লাগে সবসময় আপনার পোস্ট গুলো দেখি আপনি আসলে একজন রিয়েল হিরো স্যালট আপনাকে	N
ঢাকায় জ্যাম ছিলো?	N
আলম বাংলার গরিবের হিরো যার নাম হিরো আলম♥♥♥♥ জাহেদ খান তুই হয়ে যাবি তোমার বাড়ি হিন্দুয়া না পাকিস্তান	N
এগিয়ে যাচ্ছে নারী এগিয়ে যাচ্ছে দেশ	N
আরে গুরুতে <span style="background-color: black; color: black;">          </span> দিছে	N

Figure 8.1: Non Bullying Data and Result

Data that results positive	Result
বারো [REDACTED]	Y
আগে শাক দিয়ে মাছ ঢাকত এখন চল দিয়ে [REDACTED]	Y
তুই [REDACTED] নাস্তিকের জাত	Y
অাপিনারে [REDACTED] তো? তাইলেই হবে	Y
[REDACTED] তো কথা বলতে পারে না।	Y
সাবা [REDACTED]	Y
এরা আসলে [REDACTED] বাচ্ছা	Y
[REDACTED] সাফা কাবির [REDACTED] তুই	Y
বড়িরে [REDACTED] লাগে	Y
তুই প্রামাণ করলি তোর জন্মের ঠিক নাই সে টা। [REDACTED]	Y
নাস্তার ছারা [REDACTED]	Y
চল গুলা সাবা [REDACTED] একটু দেইখা লই	Y
এই [REDACTED] দের কাজ আর কি হবে	Y
জতা পিটা করা দরকার জায়েদ [REDACTED]	Y

Figure 8.2: Bullying Data and Result

This research made significant contributions to the automated identification of cyberbullying on Bangla Text. Because of the machine learning-based cyber bullying detection technique, this method is effective and competent for Bangla Text categorization. This program has made the procedure much easier and more efficient than any other available solution for Bangla Text. The tasks were determining the optimal machine learning algorithm for Bangla Text categorization and developing a social media deployment plan for Bangla text. During this phase, there was also a challenge in gathering the training data source. It was difficult to do the task adequately due to the lack of a preset instrument for data extraction.

In the future, we expect to cope up the limitations and can find a better performance. More accurate extraction technology can be made for Bangla language to extract the sentences properly. Spelling mistakes in comment section makes the cleaning process challenging which can be solved through more privileged auto correct system. In addition to that, further appropriate machine learning algorithms can be studied in the future to enhance accuracy. The dataset we used here is real time. It has a large enough quantity of data to offer high-accuracy results. Therefore, it can be updated further by taking a large number of social media comments.

# Chapter 9

## Conclusion

The research examined the efficiency of utilizing a multi-dimensional training dataset on machine learning classifiers to predict cyberbullying comments. The multi-dimensional dataset was built on the basis of linguistic pragmatics. On the other hand, Word Embedding demanded that the text file be cleaned according to its specifications. As a result, the program was able to produce an output that was as accurate as possible. Furthermore, Word embedding helps to cluster words. In a multi-dimensional vector space, it also aids in the association of words. We anticipate achieving an accuracy rate of more than 95% by employing the Random Forest method. This study will help to fill up the gap of previously used algorithms in terms of detecting cyberbullying. We also compared our work to another relevant study by using the same dataset and discovered that our approach and machine learning algorithm surpassed their classifiers in terms of accuracy. Our findings will undoubtedly improve cyberbullying detection and help people use social media safely as a result of this precision. The amount of the training data, however, limits the detection of cyberbullying patterns. As a result, more cyberbullying data is required to increase effectiveness. Furthermore, we have the ability to simplify our system in the future. As per consequence, machine learning approaches will be suitable for the huge dataset. Because, as per result, it has the best accuracy over all the techniques.

# Bibliography

- [1] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001. DOI: 10.1023/A:1010950718922.
- [2] A. Smola and S. Vishwanathan, “Introduction to machine learning,” *Cambridge: Cambridge University Press*, 2008.
- [3] K. S. Cho, J. Y. Yoon, I. J. Kim, J. Y. Lim, S. K. Kim, and U.-M. Kim, “Mining information of anonymous user on a social network service,” in *2011 International Conference on Advances in Social Networks Analysis and Mining*, 2011, pp. 450–453. DOI: 10.1109/ASONAM.2011.19.
- [4] R. J. I. J. Wang and J. Luk, “Peer victimization and academic adjustment among early adolescents: Moderation by gender and mediation by perceived classmate support,” *Journal of School Health* 81, 2011,, pp. 386–392, 2011. DOI: 10.1111/j.1746-1561.2011.00606.x. [Online]. Available: <https://doi.org/10.1111/j.1746-1561.2011.00606.x>.
- [5] K. Reynolds, A. Edwards, and L. Edwards, “Using machine learning to detect cyberbullying,” *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, vol. 2, Dec. 2011. DOI: 10.1109/ICMLA.2011.152.
- [6] M. Dadvar and F. de Jong, “Cyberbullying detection; a step toward a safer internet yard,” Apr. 2012. DOI: 10.1145/2187980.2187995.
- [7] V. Chavan and S. S S, “Machine learning approach for detection of cyber-aggressive comments by peers on social media network,” Aug. 2015, pp. 2354–2358. DOI: 10.1109/ICACCI.2015.7275970.
- [8] S. Vijayarani and M. Janani, “Text mining: Open source tokenization tools – an analysis,” *Advanced Computational Intelligence: An International Journal (ACII)*, vol. 3, Jan. 2016. DOI: 10.5121/acii.2016.3104.
- [9] Abdhullah-Al-Mamun and S. Akhter, “Social media bullying detection using machine learning on bangla text,” Dec. 2018, pp. 385–388. DOI: 10.1109/ICECE.2018.8636797.
- [10] E. Bartoloni, C. Baldini, F. Ferro, A. Alunno, F. Carubbi, G. Cafaro, S. Bombardieri, R. Gerli, and E. Grossi, “Application of artificial neural network analysis in the evaluation of cardiovascular risk in primary sjögren’s syndrome: A novel pathogenetic scenario?” *Clinical and experimental rheumatology*, vol. 37 Suppl 118, no. 3, pp. 133–139, 2019, ISSN: 0392-856X. [Online]. Available: <http://europepmc.org/abstract/MED/31464678>.

- [11] A. Borghesi, A. Bartolini, M. Lombardi, M. Milano, and L. Benini, “Anomaly detection using autoencoders in high performance computing systems,” vol. 33, 2019. DOI: 10.1609/aaai.v33i01.33019428. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/4993%5C%7D>.
- [12] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, “Social media cyberbullying detection using machine learning,” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, 2019. DOI: 10.14569/IJACSA.2019.0100587. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2019.0100587>.
- [13] M. I. Islam, F. Kasem, R. Meem, A. Rakshit, and M. Habib, “Bangla spell checking and correction using edit distance,” Apr. 2019. DOI: 10.1109/ICASERT.2019.8934536.
- [14] G. Orrù, M. Monaro, C. Conversano, A. Gemignani, and G. Sartori, “Machine learning in psychometrics and psychological research,” *Frontiers in Psychology*, vol. 10, Dec. 2019. DOI: 10.3389/fpsyg.2019.02970.
- [15] D. reporter, *70% of women facing cyber harassment are 15-25 years in age*, DhakaTribune, Bangladesh, Sep. 2019. [Online]. Available: <https://www.dhakatribune.com/bangladesh/dhaka/2019/09/24/70-of-women-facing-cyberharassment-are-15-25-years-in-age>.
- [16] P. Siriaraya, Y. Zhang, Y. Wang, Y. Kawai, M. Mittal, P. Jeszenszky, and A. Jatowt, “Witnessing crime through tweets: A crime investigation tool based on social media,” Nov. 2019, pp. 568–571, ISBN: 978-1-4503-6909-1. DOI: 10.1145/3347146.3359082.
- [17] Z. news 2020, *Bangladesh teen wins international peace prize for anti-bullying app*, The Tennessee Tribune, Nov. 2020. [Online]. Available: <https://tntribune.com/bangladesh-teen-winsinternational-peace-prize-for-anti-bullying-app/>.
- [18] V. Balakrishnan, S. Khan, and H. Arabnia, “Improving cyberbullying detection using twitter users’ psychological features and machine learning,” *Computers Security*, vol. 90, p. 101710, Mar. 2020. DOI: 10.1016/j.cose.2019.101710.
- [19] D. Bruwaene, Q. Huang, and D. Inkpen, “A multi-platform dataset for detecting cyberbullying in social media,” *Language Resources and Evaluation*, vol. 54, Dec. 2020. DOI: 10.1007/s10579-020-09488-3.
- [20] J. De Angelis and G. Perasso, “Cyberbullying detection through machine learning: Can technology help to prevent internet bullying?” *International Journal of Management and Humanities*, vol. 4, p. 57, Jul. 2020. DOI: 10.35940/ijmh.K1056.0741120.
- [21] C. Iwendi, G. Srivastava, S. Khan, and P. Reddy, “Cyberbullying detection solutions based on deep learning architectures,” *Multimedia Systems*, Oct. 2020. DOI: 10.1007/s00530-020-00701-5.
- [22] L. L.-R. K. Hellfeldt and H. Andershed, “Cyberbullying and psychological wellbeing in young adolescence: The potential protective mediation effects of social support from family, friends, and teachers.,” *International Journal of Environmental Research and Public Health* 17,, p. 45, 2020. DOI: 10.3390/ijerph17010045. [Online]. Available: <https://doi.org/10.3390/ijerph17010045>.

- [23] S. M. Kargutkar and V. Chitre, “A study of cyberbullying detection using machine learning techniques,” in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 734–739. DOI: 10.1109/ICCMC48092.2020.ICCMC-000137.
- [24] G. Paez, “Assessing predictors of cyberbullying perpetration among adolescents: The influence of individual factors, attachments, and prior victimization,” *International Journal of Bullying Prevention*, vol. 2, pp. 1–11, Jun. 2020. DOI: 10.1007/s42380-019-00025-7.
- [25] M. F. Ahmed, Z. Mahmud, Z. Biash, A. Ryen, A. Hossain, and F. Ashraf, *Cyberbullying detection using deep neural network from social media comments in bangla language*, Jun. 2021.
- [26] S. Cook, *Cyberbullying facts and statistics for 2018-2021*, Comparitechh, 2021. [Online]. Available: <https://www.comparitech.com/internet-providers/cyberbullying-statistics/>.
- [27] B. Dean, *Social network usage growth statistics: How many people use social media in 2021?* 2021. [Online]. Available: <https://backlinko.com/social-media-users>.
- [28] M. Islam, M. A. Uddin, L. Islam, A. Akhter, S. Sharmin, and U. Acharjee, “Cyberbullying detection on social networks using machine learning approaches,” Apr. 2021. DOI: 10.1109/CSDE50874.2020.9411601.
- [29] D. reporter, *Chanchol amader vai*, Prothom Alo, Bangladesh, May 2021. [Online]. Available: <https://www.prothomalo.com/entertainment/dhallywood/%E0%A6%9A%E0%A6%9E%E0%A7%8D%E0%A6%9A%E0%A6%B2%E0%A6%9A%E0%A7%8C%E0%A6%A7%E0%A7%81%E0%A6%B0%E0%A7%80%E0%A6%86%E0%A6%AE%E0%A6%BE%E0%A6%A6%E0%A7%87%E0%A6%B0%E0%A6%AD%E0%A6%BE%E0%A6%87>.
- [30] N. reporter, *Survey report 2021, all the latest cyberbullying statistics and what they mean in 2021*. BroadBand Search, 2021. [Online]. Available: <https://www.broadbandsearch.net/blog/cyber-bullying-statistics>.