DETECTING SOCIAL MEDIA CYBERBULLYING ON BANGLA LANGUAGE USING MACHINE LEARNING

BY

MAHDI HASAN NILOY
ID: 211-25-017

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

Supervised By

Dr. S. M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Project titled **"DETECTING SOCIAL MEDIA CYBERBULLYING ON BANGLA LANGUAGE USING MACHINE LEARNING"**, submitted by **MAHDI HASAN NILOY**, ID No: **211-25-017** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 19-01-2022

BOARD OF EXAMINERS

Chairman

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

Internal Examiner

Dr. Fizar Ahmed
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Naznin Sultana
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that, this thesis has been done by me under the supervision of **Dr. S. M. Aminul Haque, Associate Professor, Department of CSE** Daffodil International University. I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. S. M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Mahdi Hasan Niloy
ID: 211-25-017
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing that make me possible to complete the final year thesis successfully.

I am really grateful and wish my profound and indebtedness to **Supervisor Dr. S. M. Aminul Haque**, **Associate Professor**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "*Machine Learning*" instructed to carry out this Thesis. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this thesis.

I would like to express my heartiest gratitude to, **Professor Dr. Touhid Bhuiyan**, and Head**,** Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

ABSTRACT

On the internet, the number of cyberbullying importunity in the Bangla language is adding in a noteworthy way. All kinds of people like men, women and youths are being the victims of cyberbullying substantially through social medias. There's hardly the system of discovery on the cyberbullying in Bangla language. My ideal is to descry cyberbullying and to argue out of the bullying using machine learning. To complete this ideal there's the need of Bangla dataset, but unfortunately this dataset is veritably rare to find. So I collected the data from Youtube, Facebook etc. using some scrapper tools. The dataset is labelled as cyberbullying " YES" or " NO". Machine learning is the stylish way of approach for my work. I've used many a type of algorithms like Natural Language Processing (NLP), Logistic Regression (LR), Multinomial Naïve Bayes (MNB), Support Vector Classifier (SVC), Random Forest Classifier (RFC), Decision Tree Classifier, KNeighbors Classifier, AdaBoost Classifier, Bagging Classifier, ExtraTreeClassifier, GradeintBoosting Classifier, XGB Classifier. After applying all these algorithms, the exactitude is plant in Logistic Regression (LR) 89.81%, Multinomial Naïve Bayes (MNB) 89.38%, Support Vector Classifier (SVC)90.0%, Random Forest Classifier (RFC) 89.91%, Decision Tree Classifier 86.39%, GradeintBoosting Classifier 89.81%. And the maximum exactitude in Support Vector Classifier (SVC), Which is 90.0%.

Keywords-Cyber Bullying, Machine Learning, Natural Language Processing, Logistic Regression, Multinomial Naïve Bayes, Support Vector Classifier, Decision Tree Classifier.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

About 300 million people speak in the Bangla language [1], which is recognized as the 5th native language in the world and its rate is 4% [2]. Bangladesh is the only land that sacrificed their blood for the sake of their mother tongue. Only because of this movement held in Bangladesh in 1952, The International Mother Language Day, established by UNESCO in 1999, is honored by individuals around the globe [3]. Bangla language is the main native language and culture of our country. We respect our mother tongue from the core of our hearts. That's why we should not let this language used for any kind of misuse. But unfortunately, it is sad to know that, this language is being misused continuously on the Internet.

The internet penetration rate on the globe is 4.88 billion [4]. Among these, 128.78 million in September 2021 internet users are from Bangladesh [5]. Which is 28.8% of the whole. 7.7 million (19%+) internet users increased between 2020-2021 [6]. Social Media is used by 4.48 billion people on the earth [7]. Among them in Bangladesh 45 million people users social media [8]. The rate of cyberbullying victims in Bangladesh is 60%, in which the women and children are the highest victims by BIID survey (June 2021) [9]. The rate of women victims is 80% aged between 14-22 among the cyberbullying rate [10]. In this 64% is city girls and 33% are rural areas that intake sexually explicit videos, messages, and photos. 49% of Bangladeshi school pupils face cyberbullying [9]. Young people are the main culprit for cyberbullying. Most of cyberbullying is mainly affected through social media such as Youtube, Facebook, Instagram, twitters, etc [11]. Most of the time harassment is done through the Bangla language in these social media in Bangladesh. If a word incorporates racist or bigoted obscenities, assaults or condemns any ethnic or spiritual position, or inspires criminal activity, it is considered threatening or offensive. Inflammatory language has a severe negative impact on teenagers, and it can even lead to

violence. So, we should take initiatives to prevent this cyberbullying and need to make a strong system to detect cyberbullying words and lines.

Machine learning is the best way to detect cyberbullying words or lines. This type of works in the Bangla language is very rare. The Bangla dataset is also very rare to find. So, a dataset has been made to detect the words. This data is collected through Youtube, Facebook comments using Facepager, instant data scraper tools. Several Machine Learning-based algorithms and Natural Language approaches are used in this research. A new dataset is built to identify cyberbullying words or lines in the Bangla language. Dataset is labeled in two categories- Cyberbullying "YES" or "NO". Machine Learning-based algorithms and Natural Language Processing techniques have been applied in this dataset to detect cyberbullying in the Bangla language.

1.2 Related Work

Since last few years, many researches have been done in automatic cyberbullying detection. It has been done in several language such as English, Arabic, Spanish language. The field of knowledge done in Bangla is really limited.

Abdullah Al Mamun, Shahin Akter researched cyberbullying detection in Bangla text and collected data from social media like Facebook, Twitter. They partitioned dataset into two attributes such as "Not Bullied" & "Bullied". Used tf-idf in the feature extract. Stemmined & tokenized the data. Using Machine learning algorithms such as Naïve Bayes, J48, SVM and KNN. All the algorithms were used in weka platform. They found their best accuracy in support vector machine(SVM). They compared the accuracy of cyberbullying detection in between Bangla and English language [12].

Puja Chakraborty, Md. Hanif Seddiqui collected 5644 data from Facebook, Prothom Alo, BBC Bangla. Tokenized the data through NLP & then they used various types of classifiers. Used Natural language Processing (NLP) for automatic detecting abusive language in Bengali. Considered Unicode emotions & characters in Bangla language to their system. They Used three different types of algorithms like MNB, SVM, CNN-LSTM. They got the best accuracy in SVM algorithms which is 77.5% [13].

Estiak Ahmed Emon and his team mates collected data from Prothom Alo, Youtube. The size of data 4700. Stemming their dataset through NLP. They used different types of the classifier of deep learning, machine learning techniques. In the feature extract used count vectorizer & if-idf vectorizer. Used separate types of deep learning algorithms like ANN, RNN, LSTM. They found their best accuracy in RNN which rat is 82.20% better than classifier of machine learning [14].

Gabriel A. Leon-Paredes and his team mates they detected cyberbullying through Twitter using machine learning in Spanish language. In their paper, they used Naïve Bayes, SVM, Logistic Regression. Also used Natural language processing(NLP). They got 93% accuracy in SVM through studying case three times [15].

Ricardo Martins applied emotional words to detect cyberbullying. They got best accuracy in SVM which rat is 80.56% [16].

Shovon Ahammed collected data from Facebook in Bangla language. They labelled dataset in two categories. They also had neutral data in their dataset. They analyzed data in different type of diagram & histogram. In the feature extract used count vectorizer. Using only two algorithms like Naïve Bayes, SVM. They found their best accuracy 72% in Naïve Bayes [17].

Batul Haider detected cyberbullying in Arabic language. In this paper they applied deep learning algorithms, NLP. They used these algorithms in 32890 Arabic Characters. They compared both validation and test accuracy. Build FFNN model with 4 hidden layers and found validation accuracy 94.56% and test accuracy 92.53% [18].

1.3 Methodology

In this section I describe all the work process in a diagram. Everything has been described here very clearly starting from data collection to result.

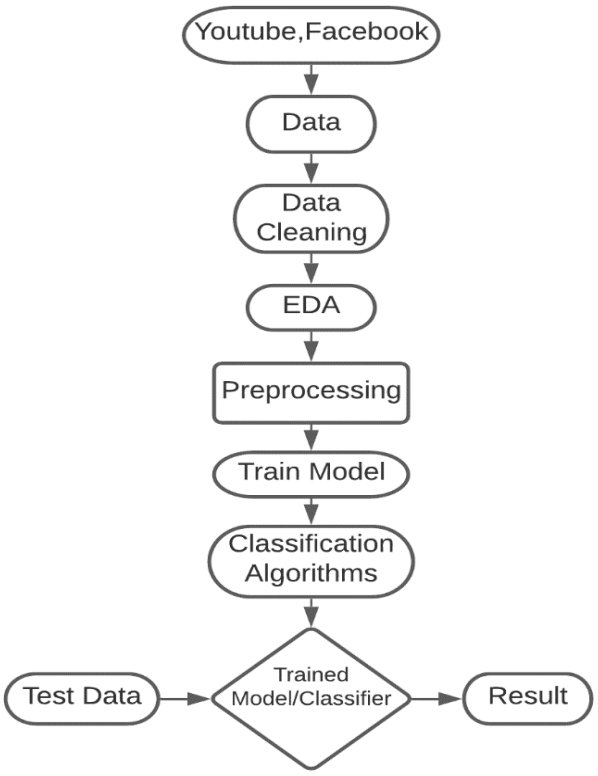


Figure 1.3.1: Methodology flow chart

Fig.1.3.1 describes the full process of work. The work is completed in separate steps. At the first, I have collected the data from Youtube, Facebook with the help of some tools. After wise, I have created a dataset. Then I have done the data cleaning. I have done Exploratory Data Analysis(EDA) of the dataset through different figures, diagram and it is shown through different charts. In the EDA section, I have described the data analysis, Natural Language Processing(NLP), data description, data correlation. In NLP data has been tokenized. Data has been described by mean & standard equations shown in some tables. After that, I have preprocessed the dataset. In this section, I have cleaned some unnecessary space, emoji & all languages excluding Bengali. Then using Count Vectorizer,

tf-idf Transformer. I have used these Count Vectorizer, if-idf Transformer through the pipeline, fit and trained the pipeline model. 70% of data have been trained & 30% of data have been tested. In my model, I have been used some popular evaluation metrics to measure performance. Different classifiers/Algorithms are used in this model. After applying all the classifications in my model, I could predict the data and find the best accuracy.

CHAPTER 2
DATA

2.1 Data set

Social media like Youtube, Facebook is the best medium of collecting dataset. Firstly, I have selected the roasting videos, news channels, celebrating pages, roasting pages from Youtube & Facebook and collected the Bangla comments and some information. Some scraper tools like Facepager apps and Instant Data Scraper Extension file in browser is used for collecting data. Total 10910 data have been collected. Most of the data are collected from Youtube. Labelled the data in 2 categories, one of them is Cyber Bullying "YES" & "NO".

Example:

1. বাংলাদেশের পচা মাল ভারতে রপ্তানি করে বাংলাদেশ গর্বিত

   English: Bangladesh is proud to expert its rotten goods to India

   This is a normal sentence so labelled as Cyber Bullying, "NO"

2. তোর মত বেশ্যা কে ত্যাগ করে ভাল আছে তাহসান।

   English: Tahsan, it is better to leave a prostitute like you

   This sentence is sentimentally hurts so labelled as Cyber Bullying, "YES"

3. রপ্তানিকৃত বেশ্যা মিথিলা।

   English: Exported prostitute Mithila.

   This sentence is bad sentence so labelled as Cyber Bullying, "YES".

4. লজ্জা বলে একটা শব্দ আছে

   English: There is a word for shame

   This is a normal sentence so labelled as Cyber Bullying, "NO"

5. বাবা, মেয়েকে খুব মানায়ছে

   English: Dad, the girl is very agreeable

   This is a normal sentence so labelled as Cyber Bullying, "YES"

That was the hard part of this paper for labelling data. This dataset file is saved as "CSV UTF-8" otherwise it would be not supported Bangla text.

2.2 Data Cleaning

Data cleaning means identify & correcting errors that can put a negative impact in predicting model. Data cleaning is used to remove any unusable columns, null values, duplicate values & repair the errors data.

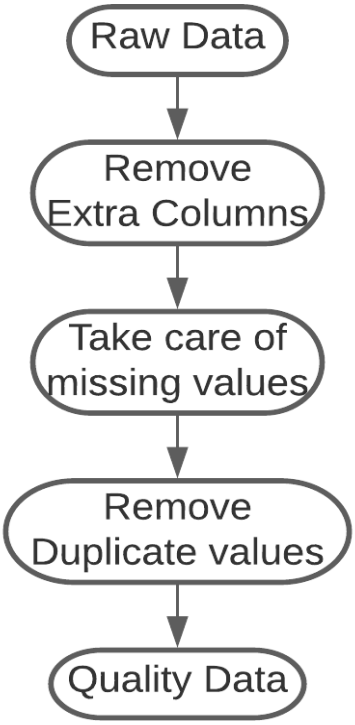


Figure 2.2.1: Data cleaning flow chart

In fig.2.2.1 firstly, remove the extra columns from dataset. Secondly, checked the null values; found 4 null values in comment columns & removed the null values. Also erased 108 duplicate values. Then finally after erasing the extra columns, null values, duplicate values; I have got quality Dataset.

CHAPTER 3
EXPLORATORY DATA ANALYSIS(EDA)

3.1 Data Analysis

Exploratory Data Analysis (EDA) refers to assaying data, probing data, recapitulating data, visualization of data by some styles. Exploratory Data Analysis is substantially shown graphically. It's a veritably important part of Machine Learning.

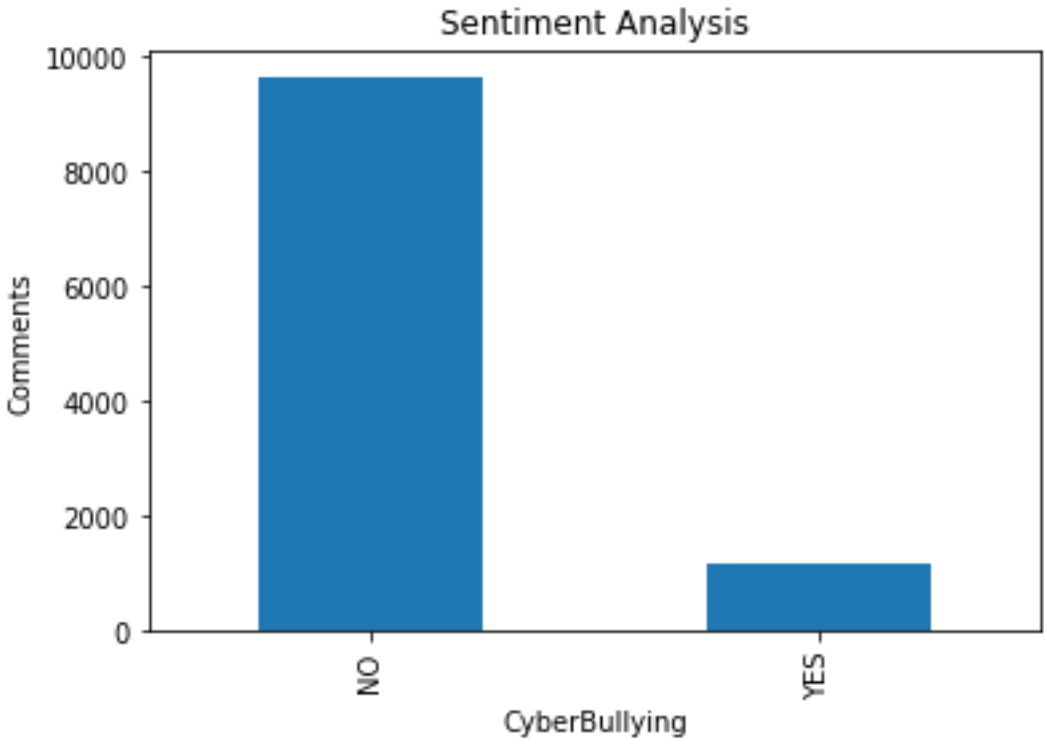


Figure 3.1.1: Sentiment analysis

At fig.3.1.1 cyber bullying “YES” or “NO” sentiment analysis is shown through bar diagram. After data cleaning 10797 data have been found in which 1165 are bullying comments & 9632 are not bullying comments.

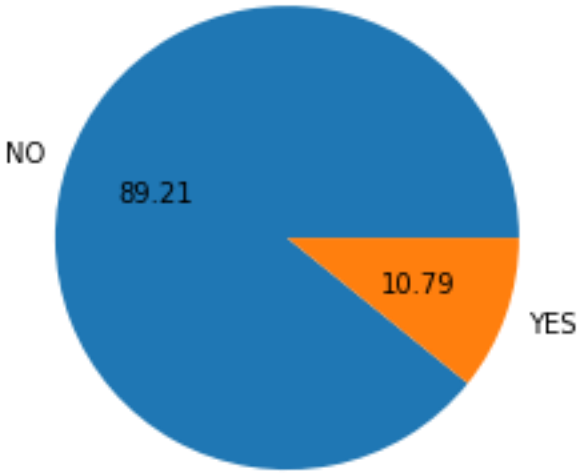


Figure 3.1.2: Percentage of analysis

In fig.3.1.2, 10.79% cyber bullying “YES” & 89.21% cyber bullying “NO” is given through pie diagram.

3.2 Natural Language Processing(NLP)

NLP (Natural Language Processing) is an Artificial Intelligence discipline. Natural Language Processing (NLP) is a computer-assisted approach to estimate language. Natural Language Processing (NLP) methods underpin Machine Learning techniques. Natural Language Processing is the ability to comprehend, scrutinize, alter, and maybe synthesize human language.

By using NLP(Natural Language Processing) the amount of words, characters, sentences, length classified through tokenization. Tokenization can partially separate sentence into small pieces. I divided the sentence of my dataset into 4 parts using tokenization through Natural Language Processing(NLP) such as words, characters, sentences length.

	comment	CyberBullying	words	characters	sentences	length
0	বাংলাদেশের পচা মাল ভারতে রপ্তানি করে বাংলাদেশ ...	NO	8	52	1	8
1	তোর মত বেশ্যা কে ত্যাগ করে ভাল আছে তাহসান।	YES	9	44	1	9
2	ভারতকে এই প্রথম কোনকিছু দিয়ে ঠকাতে পারল বাংলা...	NO	8	50	1	8
3	মিথিলা বলে সৃজিত নাকি ওর বাবা তাহলে বাবা আর মে...	NO	26	126	1	26
4	তাহসানের উচিত তার মেয়ে কে এদের সাথে মিশতে না দ...	NO	10	50	1	10

Table 3.2.1: Tokenization

In table 3.2.1, By using tokenization through NLP words, characters, sentences, length of a comment are described.

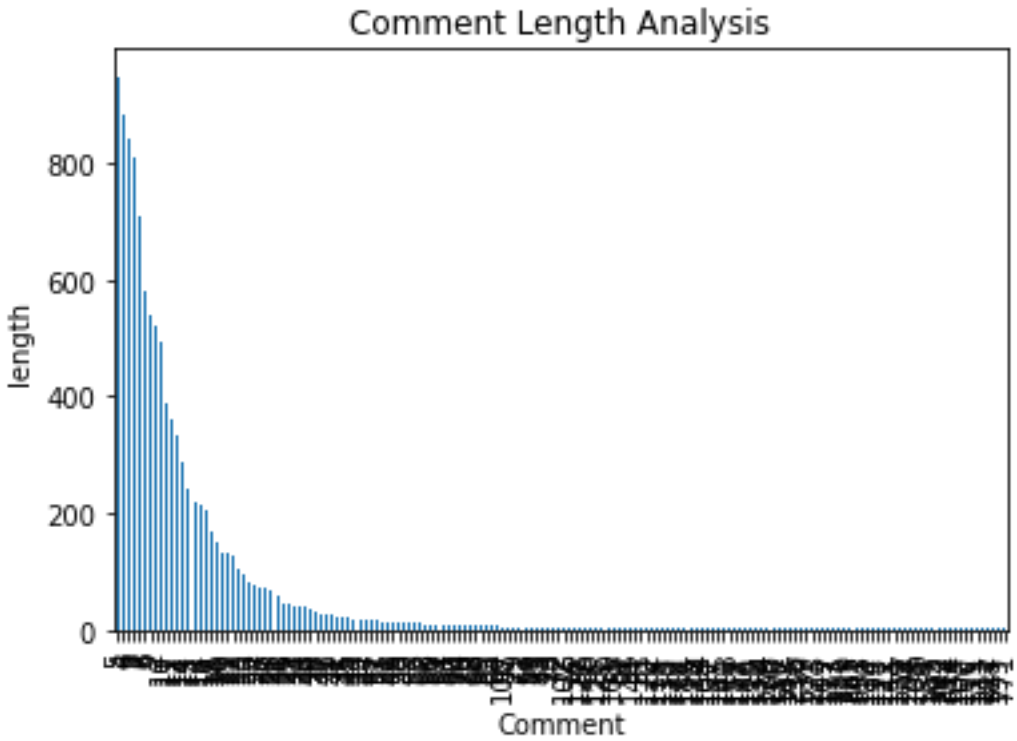


Figure 3.2.1: Comment length analysis

Fig. 3.2.1, through bar diagram comment length analysis is done.

3.3 Data Describe

Mean refers the total numbers of variable samples are divided by the number of samples. In mathematics term:

$$\bar{x} = \frac{1}{n}\sum x_i$$

here,

$\bar{x}$ = mean

n = values of x

$\sum$ = need to take sum of all the data points

Standard deviation refers the square root of the mean. In mathematics term:

$$\sigma = \sqrt{E[x^2] - (E[x])^2}$$

here,

σ = Standard deviation

$[x^2]$ = mean of the squared data

$(E[x])^2$ = square of the mean of the data

	words	characters	sentences
count	10797.000000	10797.000000	10797.000000
mean	13.884783	70.749653	1.146337
std	22.358254	117.745668	0.745294
min	1.000000	1.000000	1.000000
25%	5.000000	24.000000	1.000000
50%	8.000000	41.000000	1.000000
75%	15.000000	77.000000	1.000000
max	771.000000	3857.000000	47.000000

Table 3.3.1: Data describe

At table 3.3.1, words, characters & sentences are described of full dataset. In this figure, Words mean value 13.88, standard value 22.36, minimum words 1.00 & maximum words 771.00. Characters mean value 70.75, standard value 117.75, minimum characters 1.00 & maximum characters 3857.00. Sentences mean value 1.15, standard value 0.75, minimum sentence 1.00 & maximum sentence 47.00.

	words	characters	sentences
count	1165.000000	1165.000000	1165.000000
mean	10.300429	51.530472	1.090129
std	12.430731	65.522131	0.416103
min	1.000000	3.000000	1.000000
25%	3.000000	17.000000	1.000000
50%	7.000000	34.000000	1.000000
75%	13.000000	63.000000	1.000000
max	165.000000	960.000000	5.000000

Table 3.3.2: Cyber bullying “YES” describe

In table 3.3.2, only cyber bullying “YES” of the dataset is described. In this figure, Words mean value 10.30, standard value 12.43, minimum words 1.00 & maximum words 165.00. Characters mean value 51.53, standard value 65.52, minimum characters 3.00 & maximum characters 960.00. Sentences mean value 1.09, standard value 0.42, minimum sentence 1.00 & maximum sentence 5.00.

	words	characters	sentences
count	9632.000000	9632.000000	9632.000000
mean	14.318314	73.074232	1.153135
std	23.236662	122.360663	0.775436
min	1.000000	1.000000	1.000000
25%	5.000000	25.000000	1.000000
50%	9.000000	42.000000	1.000000
75%	16.000000	79.000000	1.000000
max	771.000000	3857.000000	47.000000

Table 3.3.3: Cyber bullying “NO” describe

In table 3.3.3, only cyber bullying “NO” of the dataset is described. In this figure, Words mean value 14.31, standard value 23.24, minimum words 1.00 & maximum words 771.00. Characters mean value 73.07, standard value 122.36, minimum characters 1.00 & maximum characters 3857.00. Sentences mean value 1.15, standard value 0.78, minimum sentence 1.00 & maximum sentence 47.00.

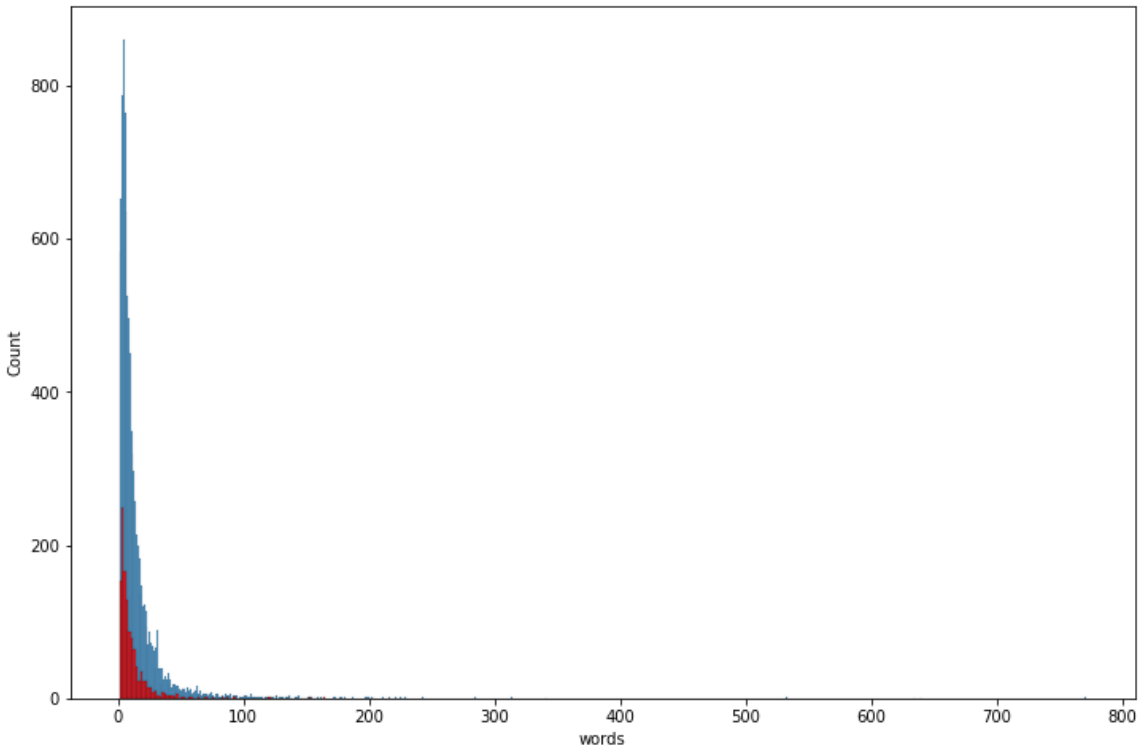


Figure 3.3.1: Words analysis

In fig. 3.3.1, a histogram is shown between bullying words and non-bullying words. Red color indicates bullying words & blue color indicates non-bullying words.

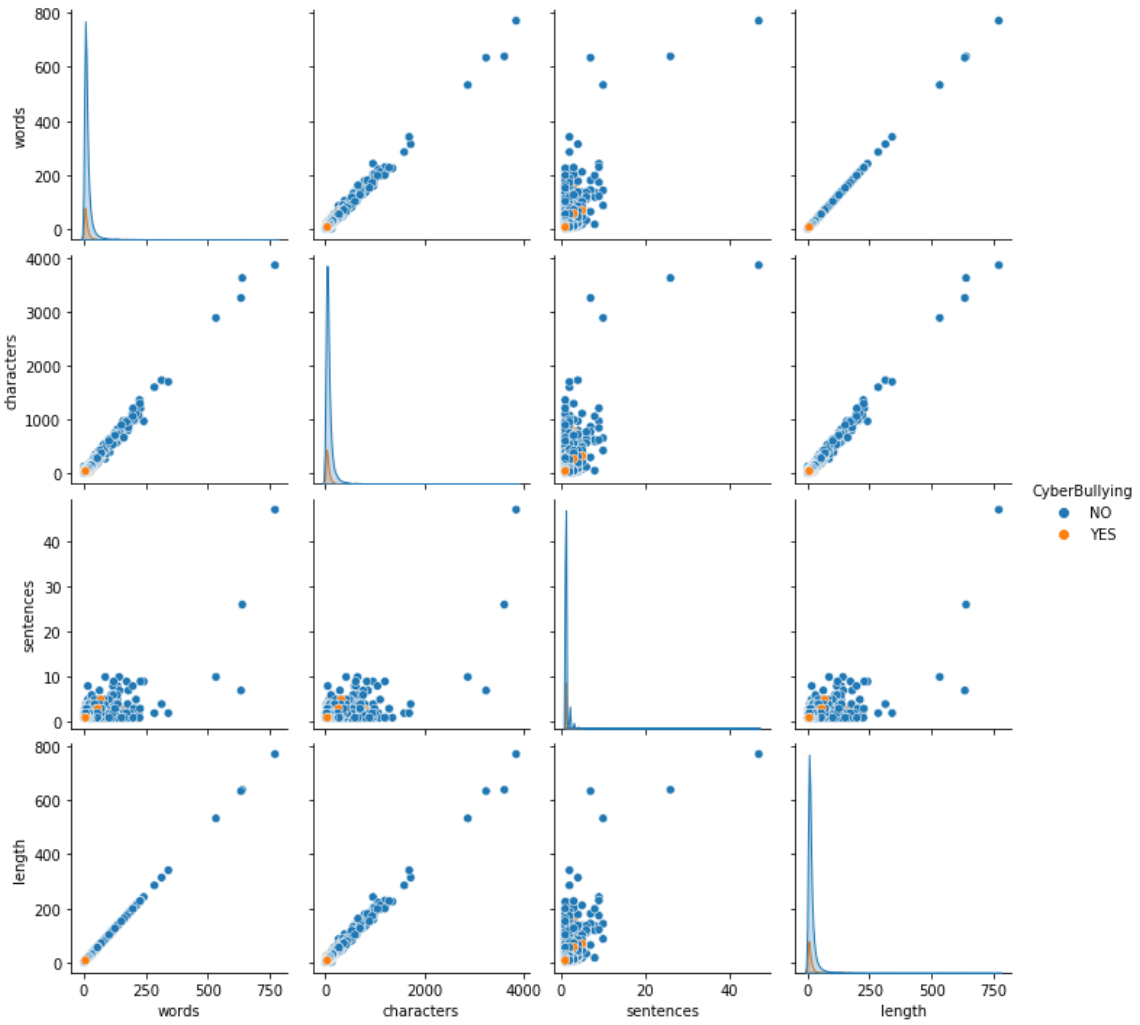


Figure 3.3.2: Relation data

Fig. 3.3.2, relation between length, sentences, characters, words of cyber bullying “YES” or “NO” analysis is shown through pair plot diagram.

By analysis table 3.3.2 & table 3.3.3 it can be said that, words, characters, sentences of mean value & standard value of comments are comparatively low than regular comments. From this analysis it is noticeable that by using less words cyber bullying is happing.

3.4 Data Correlation

Data correlation is showing how relationship is created in between one or more variables. By explaining the correlation, co means “Together” and relation means bonding between each other variables.

	words	characters	sentences	length
words	1.000000	0.992126	0.580209	1.000000
characters	0.992126	1.000000	0.553630	0.992126
sentences	0.580209	0.553630	1.000000	0.580209
length	1.000000	0.992126	0.580209	1.000000

Table 3.4.1: Data correlation

In table 3.4.1, data correlation is shown among words, characters, sentences, length. Through a heatmap diagram in fig. 3.4.1 data correlation is shown in these data.

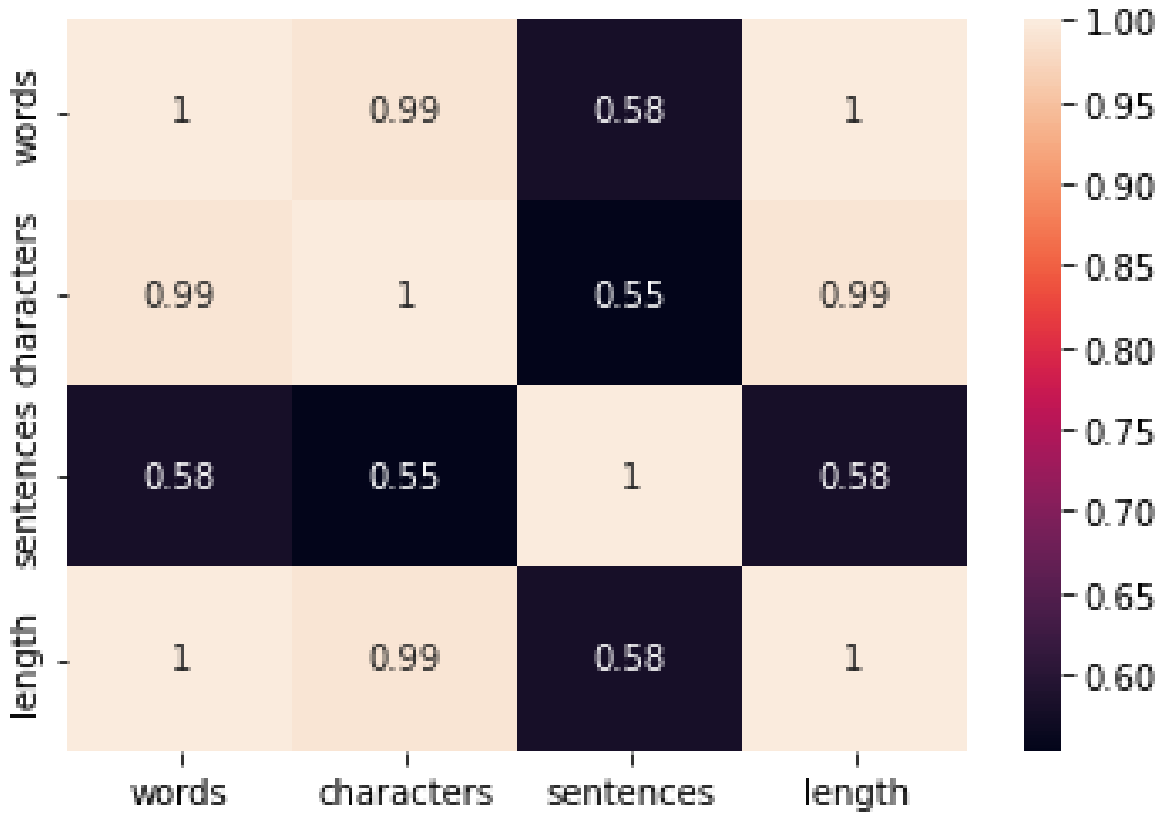


Figure 3.4.1: Data correlation

CHAPTER 4

Integration of Machine Learning

4.1 Preprocessing

The work of pre-processing is to make data readable, suitable, error free, understanding for using Machine Learning algorithms. It is basically used before Machine Leaning algorithms.

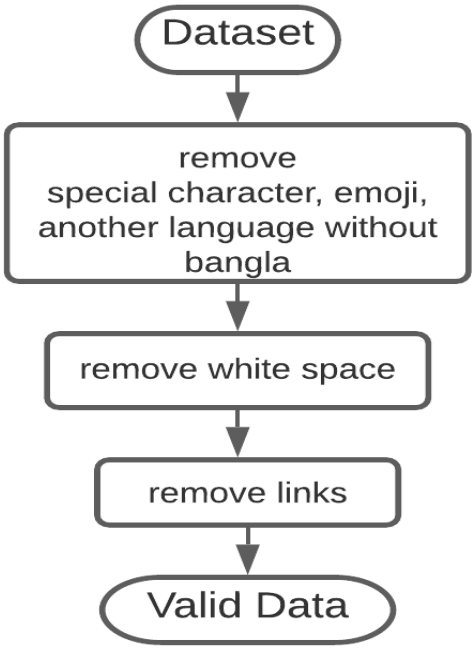


Figure 4.1.1: Data preprocessing

From figure 4.1.1, In dataset I removed special characters (@#$%&”!), emoji, all languages excluding Bangla. Then I removed all unnecessary white space , Gradually I cleared various types of links . Finally I got a valid dataset for feature exaction, training and testing.

4.2 Feature Extraction

I extracted the feature by counter vectorization and Tf-idf Transformation using a pipeline. The text are tokenization and a vocabulary of known words are created by the counter vectorization. Using that vocabulary I encoded a new document through count vectorizer. Counter Vectorization is the processing of converting into a matrix form. Transform into a normalized Tf or Tf-Idf formulation in this matrix form. Idf is for Invers Term-Frequency and Tf stands for term-frequency. In the mathematical terms:

$$\text{Term Frequency(Tf)} = \frac{Number\ of\ repetition\ of\ word\ in\ sentence}{Total\ word\ in\ sentence}$$

$$\text{Inverse Term Frequency(Idf)} = \frac{Total\ number\ of\ sentence}{No\ of\ sentence\ contain\ in\ the\ word}$$

$$\text{Tf-Idf} = \text{Tf} * \text{Idf}$$

4.3 Performance Measure

In this system, I used some evaluation criteria. These evaluation criteria are required to assess Machine Learning's performance. The most popular criteria in Machine Learning are similar to precision, recall, F1 score, support, and accuracy.

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

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

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Here,

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

In Precision, the value of true positives is divided by the aggregate of true positives and false positives. In Recall, the value of true positives divided by true positives and false negatives. The sum of recall and precision in F1 Score halves the value of recollection, 2, and precision multiplication. In Accuracy, the value of adding true positives as well as true negatives is decreased by the aggregate of true positives, true negatives, false positives, and false negatives.

4.4 Model\Classification

In this program, firstly training the dataset which is 70% of data and 30% of data are testing dataset. After training the dataset I used some classification in supervised learning. Supervised learning is supervising the data. I labeled the data in categorical value which is two parts such as "YES" & "NO". In categorical value used supervised learning classification. I used various types of classification. These types of classifications are Logistic Regression(LR), Random Forest Classifier(RFC), Multinomial Naïve Bayes(MNB), Support Vector Classification(SVC), DecisionTree Classifier, Kneighbors classifier, Ada Boost Classifier, Bagging Classifier, , XGB Classifier, ExtraTree Classifier, Gradient Boosting Classifier.

Logistic regression is the very simplest algorithm for machine learning. A supervised learning classifier is logistic regression. That classification was used to anticipate the probability of data in classification issues. A supervised learning algorithm is multinomial naive Bayes. It is the most popular algorithm of Machine Learning for analyzing text, sentimental analysis. It is usually used in text classification, sentimental analysis using NLP. These classification probabilities algorithms which are Bayes's theorem. Support Vector Classifier(SVC) is the most powerful algorithm in machine learning. It used both classification & regression. But usually used in classification problems. SVC

representation some classes of a hyperplane and multidimensional space. Random forest algorithm is supervised algorithms that are used in both regression & classification. It is the popular algorithm of machine learning. This classification is constructed from decision tree algorithms. It generates many trees for making the best decision in classification. The decision tree algorithm is a supervised machine learning technique. It employed both regression and classification techniques. But mostly used in classification problems. It generates the tree structure. The decision tree has two nodes to make a better decision. The decision tree used CART algorithms which stand both supervised learning techniques. Kneighbors classification is a supervised learning classifier. It is very simplest algorithms of supervised learning techniques. That classifier is also called the lazy learner algorithm because it does not train the data set, it performs acts on data instead. Ada boost classifier called in short adaptive boosting algorithms. It is used in ensemble method. It generally produces a strong classifier from weak classifiers. Bagging Classifier fits the random subsets in the main dataset. By taking different predictions it makes a better prediction. It is a powerful algorithm for solving supervised learning classifiers problems. extra tree classification means it generates a number of decision trees like a forest. From these multiple trees, it collects the best decision tree through multiple decorrelated & presents it as the best output result. Gradient boosting is a supervised learning technique which can be used for both classification and regression. It is based on tree-based algorithms. XBG works in large data. Its implementation of gradient boosted decision trees structure for speed and better performance. It also works in the complicated dataset. This algorithm is applied in structured and tabular data. XGB means Extreme gradient boosting techniques. It is a very lager algorithms of machine learning.

CHAPTER 5

RESULT

5.1 Result Analysis

After the performance of all the feature extraction, some important metrics, and applied some classification then I showed all the features through the tables. I got all the Precision, recall, F1 score, support value from the results. And all the accuracy values of the classification are identified.

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
classify(model, x, y)
```

```
Accuracy: 89.81481481481481
              precision    recall  f1-score   support

          NO       0.90      1.00      0.95      2890
         YES       0.76      0.08      0.15       350

    accuracy                           0.90      3240
   macro avg       0.83      0.54      0.55      3240
weighted avg       0.88      0.90      0.86      3240
```

Figure 5.1: Logistic Regression test result

In figure 5.1 shown Logistic regression found 89.81% accuracy and performance the precision, recall, F1 score and support.

```
from sklearn.naive_bayes import MultinomialNB
model = MultinomialNB()
classify(model, x, y)
```

```
Accuracy: 89.38271604938272
              precision    recall  f1-score   support

          NO       0.89      1.00      0.94      2890
         YES       0.75      0.03      0.05       350

    accuracy                           0.89      3240
   macro avg       0.82      0.51      0.50      3240
weighted avg       0.88      0.89      0.85      3240
```

Figure 5.2: MultinomialNB test result

In figure 5.2 shown Multinomial Naïve bayes(MNB) found 89.38% accuracy and performance the precision, recall, F1 score and support.

```
from sklearn.svm import SVC
model= SVC()
classify(model,x ,y)
```

```
Accuracy: 90.0
              precision    recall  f1-score   support

          NO       0.90      1.00      0.95      2890
         YES       0.80      0.10      0.18       350

    accuracy                           0.90      3240
   macro avg       0.85      0.55      0.56      3240
weighted avg       0.89      0.90      0.86      3240
```

Figure 5.3: SVC test result

In figure 5.3, shown Support Vector Classifier(SVC) found 90.0% accuracy and performance the precision, recall, F1 score and support.

```
from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier()
classify(model,x,y)
```

```
Accuracy: 89.9074074074074
              precision    recall  f1-score   support

          NO       0.91      0.99      0.95      2890
         YES       0.63      0.16      0.26       350

    accuracy                           0.90      3240
   macro avg       0.77      0.57      0.60      3240
weighted avg       0.88      0.90      0.87      3240
```

Figure 5.4: RandomForestClassifier test result

In figure 5.4, shown Random Forest Classifier found 89.91% accuracy and performance the precision, recall, F1 score and support.

```
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier()
classify(model, x, y)
```

```
Accuracy: 86.38888888888889
              precision    recall  f1-score   support

          NO       0.91      0.94      0.93      2890
         YES       0.31      0.21      0.25       350

    accuracy                           0.86      3240
   macro avg       0.61      0.58      0.59      3240
weighted avg       0.84      0.86      0.85      3240
```

Figure 5.5: DecisionTreeClassifier test result

In figure 5.5, shown Decision Tree Classifier found 86.39% accuracy and performance the precision, recall, F1 score and support.

```
from sklearn.neighbors import KNeighborsClassifier
model = KNeighborsClassifier()
classify(model,x,y)
```

```
Accuracy: 88.82716049382717
              precision    recall  f1-score   support

          NO       0.90      0.98      0.94      2890
         YES       0.45      0.15      0.22       350

    accuracy                           0.89      3240
   macro avg       0.68      0.56      0.58      3240
weighted avg       0.85      0.89      0.86      3240
```

Figure 5.6: KneighborsClassifier test result

In figure 5.6, shown KNeighbors Classifier found 88.83% accuracy and performance the precision, recall, F1 score and support.

```
from sklearn.ensemble import AdaBoostClassifier
model = AdaBoostClassifier()
classify(model, x, y)
```

```
Accuracy: 89.6604938271605
              precision    recall  f1-score   support

          NO       0.90      0.99      0.94      2890
         YES       0.61      0.12      0.20       350

    accuracy                           0.90      3240
   macro avg       0.75      0.56      0.57      3240
weighted avg       0.87      0.90      0.86      3240
```

Figure 5.7: AdaboostClassifier test result

In figure 5.7, shown AdaBoost Classifier found 89.66% accuracy and performance the precision, recall, F1 score and support.

```
from sklearn.ensemble import BaggingClassifier
model = BaggingClassifier()
classify(model, x, y)
```

```
Accuracy: 88.70370370370371
              precision    recall  f1-score   support

          NO       0.91      0.97      0.94      2890
         YES       0.45      0.19      0.27       350

    accuracy                           0.89      3240
   macro avg       0.68      0.58      0.60      3240
weighted avg       0.86      0.89      0.87      3240
```

Figure 5.8: BaggingClassifier test result

In figure 5.8, shown Bagging Classifier found 88.70% accuracy and performance the precision, recall, F1 score and support.

```
from sklearn.ensemble import ExtraTreesClassifier
model = ExtraTreesClassifier()
classify(model, x, y)
```

```
Accuracy: 89.38271604938272
              precision    recall  f1-score   support

          NO       0.91      0.98      0.94      2890
         YES       0.53      0.17      0.26       350

    accuracy                           0.89      3240
   macro avg       0.72      0.58      0.60      3240
weighted avg       0.87      0.89      0.87      3240
```

Figure 5.9: ExtraTreeClassifier test result

In figure 5.9, shown Extra Tree Classifier found 89.38% accuracy and performance the precision, recall, F1 score and support.

```
from sklearn.ensemble import GradientBoostingClassifier
model = GradientBoostingClassifier()
classify(model, x, y)
```

```
Accuracy: 89.81481481481481
              precision    recall  f1-score   support

          NO       0.90      1.00      0.95      2890
         YES       0.72      0.09      0.17       350

    accuracy                           0.90      3240
   macro avg       0.81      0.54      0.56      3240
weighted avg       0.88      0.90      0.86      3240
```

Figure 5.10: GradientBoostingClassifier test result

In figure 5.10, shown Gradient Boosting Classifier found 89.81% accuracy and performance the precision, recall, F1 score and support.

```
from xgboost import XGBClassifier
model = XGBClassifier()
classify(model, x, y)
```

```
Accuracy: 89.78395061728395
              precision    recall  f1-score   support

          NO       0.90      1.00      0.95      2890
         YES       0.77      0.08      0.14       350

    accuracy                           0.90      3240
   macro avg       0.84      0.54      0.54      3240
weighted avg       0.89      0.90      0.86      3240
```

Figure 5.11: XGBClassifier test result

In figure 5.11, shown XGB Classifier found 89.78% accuracy and performance the precision, recall, F1 score and support.

After applying all the classifications I found the best accuracy of 90.0% in Support Vector Classifier(SVC).

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

In my system, I found the best accuracy of 90.0% in Support Vector Classifier(SVC) after applying other classifications. I also used counter vectorizer and tf-idf Transformation in feature extraction. I used 11 types of classification in my system. I had shown measured performance in all classifications in the result. All the classification in my work performs very well in Bangla Language detecting. There are is limitation of correcting spelling of Bangla Language on my thesis paper. Moreover, it does not work for any specific word. Cyberbullying is being done in Bangladesh continuously for this rapid increase I have done this work. Only a bad comment can create instability. Sometimes it can also create a crime. This accuracy will help reduce cyberbullying crime in social media. Will also help reduce instability in society. The decline of society will be prevented. In society, especially women & children will feel safe in social media. It will serve as the biggest helper in the youth society. It will stop the youths from doing pornography. Abusive, sexist, racist comments, assaults or accusations of any communal or religious doctrine, instigation to criminal activity, and so on will be resisted under this threat. The intension of suicide will also decrease in the young generation.

6.2 Future Work

In my thesis paper I have done supervised learning classification. In the future, my work will be extended by applying deep learning, Neural Networks. I will also work on Bengali spelling correction. Work will be extended with the more data.

REFERENCES

[1] Learn about Wikipedia, available at << https://en.wikipedia.org/wiki/Bengalis>>, last accessed on 01-01-2022 at 03:24 PM.

[2] Learn about Wikipedia, available at << https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers>>,last accessed on 01-01-2022 at 03:24 PM.

[3] Learn about Wikipedia, available at << https://en.wikipedia.org/wiki/International_Mother_Language_Day>>, last accessed on 01-01-2022 at 03:24 PM.

[4] Learn about Datareportal, available at << https://datareportal.com/global-digital-overview>>, last accessed on 01-01-2022 at 03:24 PM.

[5] BTRC Bangladesh, available at << http://www.btrc.gov.bd/content/internet-subscribers-bangladesh-september-2021>>, last accessed on 01-01-2022 at 03:24 PM.

[6] Learn about Datareportal, available at << https://datareportal.com/reports/digital-2021-bangladesh >>, last accessed on 01-01-2022 at 03:24 PM.

[7] Banklinko, available at << https://backlinko.com/social-media-users>>, last accessed on 01-01-2022 at 03:24 PM.

[8] UNB, available at << https://www.unb.com.bd/category/Bangladesh/bangladesh-charts-9m-new-social-media-users/68129 >>, last accessed on 01-01-2022 at 03:24 PM

[9] Cyber Bullying is an alarming issue during Covid-19 situation in Bangladesh , available at << http://lawyersclubbangladesh.com/en/2021/08/03/cyber-bullying-is-an-alarming-issue-during-covid-19-situation-in-bangladesh/ >>, last accessed on 01-01-2022 at 03:24 PM

[10] New age, available at << https://www.newagebd.net/article/123926/majority-of-cyberbullying-victims-in-bangladesh-are-women>>, last accessed on 01-01-2022 at 03:24 PM.

[11] Rafael Prieto Curiel, Stefano Cresci, Cristina Ioana Muntean & Steven Richard Bishop "crime and its fear in social media," *Palgrave Communications* 6, article number: 57(2020), 02 April 2020

[12] Abdhullah-Al-Mamun, Shahin Akhter "Social media bullying detection using machine learning on Bangla text"10th International Conference on Electrical and Computer Engineering 20-22 December, 2018, Dhaka, Bangladesh

[13] Puja Chakraborty, Md. Hanif Seddiqui "Threat and Abusive Language Detection on Social Media in Bengali Language" 1st International Conference on Advances in Science, Engineering and Robotics Technology 2019 (ICASERT 2019)

[14] Estiak Ahmed Emon, Shihab Rahman, Joti Banarjee, Amit Kumar Das, Tanni Mittra "A Deep Learning Approach to Detect Abusive Bengali Text" 2019 7th International Conference on Smart Computing & Communications (ICSCC)

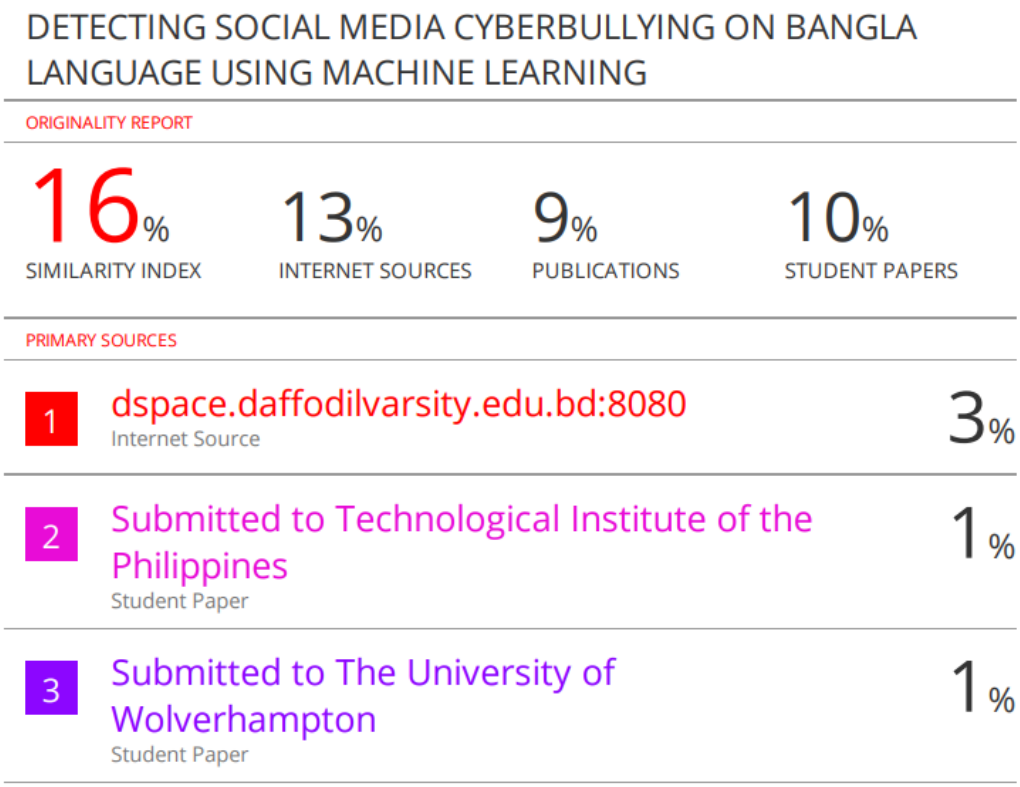
[15] Gabriel A. Leon-Paredes, Wilson F. Palomeque-Le ´ on, Pablo L. Gallegos-Segovia, Pa ´ ul E. Vintimilla-Tapia, Jack ´ F. Bravo-Torres, Liliana I. Barbosa-Santillan and Mar ´ ´ıa M. Paredes-Pinos "Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language" CHILECON 2019, October 29-31, Valpara´ıso, Chile

[16] Ricardo Martins, Marco Gomes, Jos´e Jo˜ao Almeida, Paulo Novais and Pedro Henriques, “Hate speech classification in social media using emotional analysis,” 7th Brazilian Conference on Intelligent Systems, pp. 61–66, 2018

[17] Shovon Ahammed, Mostafizur Rahman, Mahedi Hasan Niloy, S.M. Mazharul Hoque Chowdhury “Implementation of Machine Learning to Detect Hate Speech in Bangla Language” ,IEEE Conference ID: 46866; 8th International Conference on System Modeling & Advancement in Research Trends, 22nd–23rd November, 2019; College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India

[18] Batoul Haidar, Maroun Chamoun, Ahmed Serhrouchni “A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning” Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 6, 275-284 (2017)

Plagiarism Report



DETECTING SOCIAL MEDIA CYBERBULLYING ON BANGLA LANGUAGE USING MACHINE LEARNING

ORIGINALITY REPORT

16%	13%	9%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	Submitted to Technological Institute of the Philippines Student Paper	1%
3	Submitted to The University of Wolverhampton Student Paper	1%