

**DEPRESSION ANALYSIS FROM SOCIAL MEDIA BENGALI COMMENTS USING
MACHINE LEARNING TECHNIQUES**

BY

MEHEDI HASAN

ID: 191-15-2535

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

Supervised By

Ms. Naznin Sultana

Associate Professor

Department of CSE

Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam

Sr. Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY 2023

APPROVAL

This Project titled “**Depression Analysis From Social Media Bengali Comments Using Machine Learning Techniques**”, submitted by Mehedi Hasan, ID No: 191-15-2535 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 04-02-2023.

BOARD OF EXAMINERS

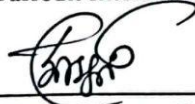
Chairman

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



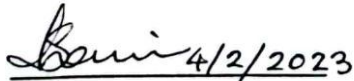
Internal Examiner

Dr. S. M. Aminul Haque
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Dewan Mamun Raza
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



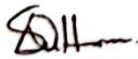
External Examiner

Dr. Shamim H Ripon
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

I hereby declare that this project has been done by me under the supervision of **Ms. Naznin Sultana, Associate Professor, Department of CSE Daffodil International University**. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

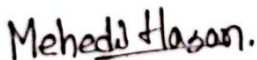


Ms. Naznin Sultana
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Mohammad Jahangir Alam
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:



Mehedi Hasan
ID: -191-15-2535
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I want to express my heartfelt thanks and gratitude to almighty God for his divine blessing, which made it possible for me to complete the final year project successfully.

I'm really grateful and wish to express my profound indebtedness to **Supervisor Ms. Naznin Sultana, Associate Professor and Co-Supervisor Mohammad Jahangir Alam, Lecturer (Senior Scale)** Department of CSE, Daffodil International University, Dhaka. Deep knowledge and keen interest of my supervisor in the field of “Machine Learning” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

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

I would like to thank my entire course mate at Daffodil International University, who took part in this discussion while completing the course work.

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

ABSTRACT

Depression is a major public health concern that can have significant negative impacts on an individual's quality of life. Early detection of depression can be crucial for facilitating timely treatment and improving outcomes. In this study, we aimed to investigate the use of machine learning algorithms for detecting depression in social media comments written in Bengali, a language spoken by millions of people around the world. We collected a dataset of social media comments written in Bengali and labeled them according to the emotional state of the person posting (e.g., happy, sad, or depressed). We describe the development and evaluation of several different algorithms, including SVM, LR, DT, KNN, CB, and LR. The results of our evaluation showed that the SVM algorithm had the highest accuracy, receiving a 75.28% score and being able to detect depression with high accuracy, suggesting that social media comments written in Bengali could be a useful source of data for detecting depression. These findings could have important implications for the development of automated tools for detecting depression in real-time and facilitating timely treatment.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgments	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-5
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	2
1.4 Research Questions	3
1.5 Expected Output	3
1.6 Report Layout	4
CHAPTER 2: BACKGROUND	5-8
2.1 Preliminaries/Terminologies	5
2.2 Related Works	5

2.3 Comparative Analysis and Summary	7
2.4 Scope of the Problem	7
2.5 Challenges	8
CHAPTER 3: RESEARCH METHODOLOGY	9-15
3.1 Research Subject and Instrumentation	9
3.2 Data Collection Procedure/Dataset Utilized	10
3.3 Statistical Analysis	12
3.4 Proposed Methodology/Applied Mechanism	13
3.5 Implementation Requirements	14
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	16-24
4.1 Experimental Setup	16
4.2 Experimental Results & Analysis	16
4.3 Discussion	24
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	25-26
5.1 Impact on Society	25
5.2 Impact on Environment	25
5.3 Ethical Aspects	26

5.4 Sustainability Plan	26
CHAPTER 6: SUMMARY, CONCLUSION & FUTURE WORK	27-29
6.1 Summary of the Study	27
6.2 Conclusions	27
6.3 Implication for Further Study	27
REFERENCES	22-23

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Non-Depressive text	10
Figure 3.2: Depressive text	10
Figure 3.3: Comparison Models	12
Figure 3.4: Workflow diagram	13
Figure 4.1: Confusion Matrix for SVM classifier	17
Figure 4.2: Confusion Matrix for multinomial NB classifier	18
Figure 4.3: Confusion Matrix for DT classifier	19
Figure 4.4: Confusion Matrix for LR classifier	19
Figure 4.6: Confusion Matrix for KNN classifier	20
Figure 4.7: Accuracy plot of all the classifiers.	20

LIST OF TABLES

TABLES	PAGE NO
Table 2.1. Comparative analysis with previous work	7
Figure 3.3: Comparison Models	12
Table 4.8: Accuracy Different Models in Percentage	22
Table 4.9: F1 Measures of six classifiers	23

CHAPTER 1

INTRODUCTION

1.1 Introduction

Depression undermines our self-assurance, optimism, and motivation, which destroys human health and is acknowledged as one of the most unbearable diseases in the world [1] [4]. Social Media is a worldwide networking system for communication. Currently, there are over 4.48 billion social media users, and each one interacts with an estimated 6.6 different social media sites. With 2.9 billion members, Facebook is the most popular social networking site among them, and YouTube has achieved 2.3 billion users worldwide. (statista0),(FB IR Q1 2020). People with depression lead lives with a wide variety of symptoms. They lose their attachments to what they enjoy & loved to do and deal with anxiety, stress, and low mood. These physical symptoms make an impact on chronic weariness, insufficient sleep, loss of appetite, and a variety of pains. It can be triggered by significant life events like a death in the family, a relationship conflict, a job loss, or the birth of a child. Based on doctors' opinions, depression's impact has three categories on a person's life. Someone who has little impact on his daily life has mild depression. People who have a significant impact on their daily lives - moderate depression.

In Bangladesh, 53.6 million people use Facebook regularly, which is an estimated 30.7% of the entire population of the country, and the majority of Facebook users are men (67.4%). The largest user group was between the ages of 18 and 24 (23.7 m). (Sept. 2022, NapoleonCat) However, as people frequently express their opinions in posts and comments on many social media sites, our main goal is to find the depressed text from there. Depression prevents students from paying close attention to their studies, and their academic or professional careers suffer. This research will help to identify the depressed person & can be ensured his/her recovery as well as helps to prevent unwilling decisions like suicide. [8,9]

1.2 Motivation

Depression is a prevalent and severe mental health condition that can greatly affect an individual's everyday life. It is a significant cause of disability globally and if not addressed can have dire consequences, including increased suicide risk. There is a pressing need for research on depression in order to better understand the underlying causes and develop effective interventions. People use social media desperately every day, revealing their issues and feelings on social networking sites. Social interaction patterns are just a few examples of data that might be used to detect depression. For instance, a person experiencing depression might stop connecting with people on social media or post less regularly. By analyzing these data, we can effectively assess whether an individual is experiencing depression or not. This analysis enables us to intervene and provide the necessary support before the individual takes any drastic action, such as attempting suicide. Identifying depression at an early stage can help prevent the progression of the condition and mitigate the risk of suicide. The data collected can include information about an individual's mood, behavior, and symptoms of depression, which can be used to make an accurate diagnosis and provide appropriate treatment. It's necessary to realize that early detection and intervention are the first actions that could be taken in the multidisciplinary and complex field of suicide prevention.

1.3 Rationale of the Study

Social media data offers a wealth of knowledge about people's ideas, emotions, and behaviors and has the potential to be utilized as a strategy to spot those who could be depressed. Systematic and quick analyses of social media data using machine learning algorithms have the potential to determine which text-to-person behaviors make the best matches. CatBoost, Multinomial NB, support vector classifiers, decision trees, and Logistic Regression are used to predictive analysis classifiers.

1.4 Research Question

- What is Depression?
- What specific machine learning algorithms are most effective for detecting depression in social media comments?

- How does the performance of these algorithms compare when applied to social media comments written in Bengali, compared to other languages?
- How accurately can depression be detected in social media comments using these algorithms, and what factors might influence the accuracy of the detection?
- How can the results of this analysis be used to better understand and address issues related to depression within the Bengali community?

1.5 Expected Output

- Improve our understanding of the causes and risk factors for depression as well as understand the effective prevention.
- Identify innovative interventions for treating depression, which could improve outcomes for individuals with the disorder.
- Explore the impact of depression on various aspects of an individual's life, such as work, relationships, and overall quality of life, which could help to identify areas that may be particularly affected by the disorder and that may benefit from targeted interventions.
- People who are depressed frequently take the time and want to think more thoroughly about their lives and life paths, which gives them the chance to make some positive adjustments.
- We can only hope that they will manage to change their life in a way that will ultimately be beneficial to them.
- To decrease the suicide rate causing depression.

1.6 Report Layout

The introduction chapter of this report covers the study's background, purpose, anticipated outcomes, research questions, project management, and report structure. The chapter also covers a comparative analysis and summary, key terms, the scope of the issue being studied, related research, and the challenges encountered during the research. In the next section, Research Methodology, we will discuss the selection of research subjects and instruments, data collection or utilization methods, statistical analysis, the proposed methodology, and implementation considerations. The experimental results and discussion section, which is a central and crucial

part of the report, includes a discussion of the experimental setup, an analysis of the results, and conclusions. The report also addresses the impact of the research on society, the environment, and sustainability, including ethical considerations and a plan for future sustainability. Finally, the report concludes with a summary of the study, conclusions, recommendations, and suggestions for further research.

CHAPTER 2

BACKGROUND

2.1 Terminologies

People of many ages, ethnic backgrounds, and socioeconomic statuses are susceptible to the prevalent and severe mental health illness known as depression. It is marked by enduring dejection, hopelessness, and a loss of enthusiasm for formerly pleasurable pursuits. Physical signs of depression might also include adjustments in eating, sleeping, and energy levels.

2.2 Related Work

Bhattacharjee, D., et al. [1] state that the purpose is to better comprehend Bengali culture, and information is being gathered from several Bangladeshi television stations and social media sites like Facebook and Twitter. A total of 35,000 messages from such platforms are gathered using the Twitter and Facebook Graph APIs, as well as a grabber script. The dataset was annotated by a highly qualified psychologist using the BDI1 and BDI2 books as references, as well as norms for human social activities and emotions and theoretical knowledge.

Sau, A., and Bhakta, et al. [2] Students from medical colleges and hospitals provided 470 data points, including sociodemographic and occupational health-related data. The data were analyzed using five different algorithms linear regression, naive Bayes, RF, SVM, and CatBoost. Among these methods, CatBoost (which enhances decision trees) had the highest accuracy.

"Uddin, A.H., et al. [6] worked to make tweet processing better. Using the Unigram model, they found that the average accuracy of different emotions were 74% for positive, 78% for sad, and 92% for a surprise. They used statistical deep learning and machine learning techniques to identify people's emotions. With a maximum accuracy of 86.3%, they were able to analyze Bengali social media depression using a long short-term memory (LSTM) network."

M.R.H. Khan et al. [9] goal is to analyze people's sentiments from the data, which is classified as happy or sad. They gather information from numerous sources, such as books, poems, and quotations, for this reason. When processing data, they tokenize their data using a count vectorizer to turn the text into a numerical value. To produce predictions, the researchers used six

different algorithms after preprocessing and tokenizing their data. The author got the highest accuracy with a precision of 86.67% using the Naive Bayes method.

Vasha, Z.N. et al. [10], the authors collected 10,000 pieces of data from Facebook and YouTube comments. Then, they separated the text into depressed and non-depressed data, which were used to compare the performance of six different algorithms, which are DT, SVM, LR, RF, KNN and NB. The best classification outcomes were obtained utilizing the TF-IDF feature combination after several words for feature extraction were used. The SVM model had a maximum prediction accuracy of 77%.

Choudhury, A.A., et al. [8] People participating in the survey should do their utmost to finish the questionnaire, which takes about 15-20 minutes. The Random Forest algorithm was identified as the best performer among the algorithms tested, with an accuracy of 75% and an f-measure of 60%. The results of the Random Forest algorithm were comparable to the Support Vector Machine algorithm in terms of accuracy and f-measure, but Random Forest showed greater precision, recall, and fewer false negatives.

Tuhin, Rashedul Amin, et al. [11] Sentiment analysis was performed at the document level using a variety of machine learning algorithms, taking into account both the type of article and the type of sentence. When using a topical approach, the researchers obtained an accuracy rate of around 70%, while using Naive Bayes resulted in an accuracy of about 50%. While there are limitations to this method, the accuracy obtained for analyzing Bengali content is considered to be fairly good.

Arora, P., and Arora et al. [12] gathered information from Twitter streams, as well as all the most recent tweets. They categorized tweets based on words and phrases used to describe anxiety, stress, and mental issues. Authors distinguished between tweets about well-being and mixed tweets using support vector regression and multinomial naive Bayes techniques. Also, they eliminated emojis and certain terms from the data during pre-processing. The complete dataset, consisting of 3754 posts, was used to evaluate the performance of three different techniques: NB, SVM and K-means clustering. The accuracy achieved by these techniques was 76%, 78.8%, and 77.17% respectively

2.3 Comparative Analysis and Summary

An analysis of the author's methods, strengths, and results in relation to earlier work is given here:

Table 2.1. Comparative analysis with previous work

SL	Author	Process	Strength	Result
1	M.R.H. Khan [9]	Multinomial Naive Bayes	Accurateness of Classification.	86.67%.
2	Bhattacharjee, D., et al [1]	BDI1 and BDI2 method	Qualitative investigation of the API based data	90%
3	Choudhury, A.A., et al [8],	Beck Depression and Depression Anxiety Stress Scales	Explanations for up-to-date text-based sentimentality analyzers	75%
4	Tuhin, Rashedul Amin, et al. [11]	Naive Bayes Algorithm	Sentiment analysis Used	70%
5	Arora, P., and Arora, et al. [12]	SVM techniques	Used Numerous Classifiers.	78.8%
6	Vasha, Z.N. et al. [10]	TF-IDF Vectorizer and SVM method	Huge amount of data	77%

2.4 Scope of the problem

This work builds a model that can distinguish between depressive and non-depressive text from any social media data using data analysis, machine learning methods. In order to solve these issues, society can make use of this model to identify those who are depressed and provide medical treatment to reduce the serious effects of depression.

2.5 Challenges

To complete the project as intended, we had to overcome a variety of challenges. In particular, it is difficult to acquire individual pieces of Bengali text from social media. These types of data contain large amounts of mistakes because different mindsets communicate their thoughts in different ways on social platforms, and sometimes they are not aware of the spelling mistakes. In Bengali text analysis, the erroneous spelling structure is a significant problem. Using English alphabets inside Bengali letters is a significant problem as well. Hence, the words' intended meaning cannot be expressed in their original meaning in Bengali.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject & Instrumentation

The process of creating a graphical flowchart involved the use of multiple tools and techniques. Six separate algorithms were utilized to analyze the data, along with Microsoft Excel and Draw.to, a website that helps to create flowcharts. To ensure the most efficient and reliable solution, we utilized well-established frameworks and a Google collaboration to run the python code and machine learning algorithms to analyze the processed data. The data used in the study was collected from various social networking websites and was then organized and stored in an Excel spreadsheet. By using these tools, we were able to effectively process and analyze the data, leading to the creation of an accurate and informative graphical flow chart.

3.2 Data Collection Procedure/Dataset Utilized

We collected a dataset of 5604 comments, with 3265 classified as depressed and 2339 classified as non-depressed, from various platforms such as Facebook, Twitter, and YouTube. The data was then stored in an Excel spreadsheet and used to train various machine learning algorithms such as Naive Bayes, Logistic Regression, K-Nearest Neighbors, Decision Tree, Support Vector Machine, and Random Forest. The implementation of these algorithms was carried out using Google Colaboratory (Colab) which is a product from Google Research. Colab is a web-based platform that allows users to write and execute Python code through their browser, making it an ideal tool for machine learning, data analysis, and education. With its user-friendly interface, Colab makes it easy for anyone to start their data science journey without the need for extensive setup. Colab also provides access to Google's computing resources, such as GPUs and TPUs, making it a powerful tool for deep learning tasks. Additionally, Colab offers different versions, such as Colab Pro and Colab Pro+, that have increased RAM limits and longer session times.

একধাপ এগিয়ে যাওয়ার গল্প কার না ভাল লাগে।	0
এক কথায় সফলতার গল্পটা সৃষ্টির সূচনা থেকেই মানুষকুলকে ঘিরে আবর্তিত।	0
যে মুহূর্ত থেকে তুমি নিজেকে অজুহাত দেখানো বন্ধ করে কাজ শুরু করবে সে মুহূর্ত থেকে তোমার স্বপ্ন আর স্বপ্ন থাকবে না- সেটি বাস্তবে রূপ নিতে শুরু করবে!	0
জীবনে সবকিছু একবার হলেও চেষ্টা করে দেখা উচিত	0
তোমার জীবনটা কিন্তু একান্তই তোমার, তুমি কী হবে কী করবে সে ব্যাপারে অন্যরা পরামর্শ দিতে পারে কিন্তু তাদের ইচ্ছা তোমার উপর চাপিয়ে দেওয়ার কোন অধিকার নেই কারো	0
পৃথিবীর সবচেয়ে আনন্দের অনুভূতিটি হচ্ছে যখন তুমি একটি লক্ষ্য ঠিক করেছিলে সেই লক্ষ্যটি পূরণ করতে পারলে!	0
জীবন হোক কর্মচাঞ্চল্যে ভরপুর, ছুটে চলার নিরন্তর অনুপ্রেরণা।	0
বিনয়ী হতে হলে প্রয়োজন অসাধারণ আত্মসম্মান এবং মানসিক শক্তিমত্তার।	0
জিততে তোমাকে হবেই!	0

Fig 3.1: Non-Depressive text.

কাউকে ছাড়া কারো জীবন থেমে থাকে নাহ কখনোই, হ্যা হয়তো অনুভূতি গুলোর অকাল মৃত্যু ঘটে।	1
এইজন্য আমি বলি তোমাকে ছাড়া আমার অনেক কষ্ট হবে,, বেঁচে তো থাকবো কিন্তু হাসতে পারবো না,	1
হ্যাঁ বেঁচে থাকে শুধু শরীরটা। জীবন কারো জন্য থেমে থাকে না কিন্তু মনটা থেমে যায় প্রিয় মানুষটার জন্য।	1
কিন্তু তাকে পেয়ে গেলে আমার আর কিছু লাগতো নাহ	1
বাঁচবো ঠিকই কিন্তু বাচার মোতো বাচবোনা	1
দিনশেষে সবাইই বেচে থাকে!	
মরে যায় আমাদের মুখে বলা সেই বড় বড় রচনা গুলো	1
ভালোবাসা বলা যতটা সহজ, হৃদয়ে তা ধারণ করা ঠিক ততটাই কঠিন।	1
সত্যিই ভালোবাসা সবার জন্য নয়!	1
যদি এই ভাঙ্গাই শেষ ভাঙ্গা হয়, তবে আমি বলবো আত্মহত্যাই শ্রেয়	1
এই জন্য একা থাকতে ভালোবাসি	1
সবকিছুর সমাপ্তি ঘটিয়ে সে অন্য কাহারো হইল	1
যখন না জানিয়ে মেয়েটি অন্য ছেলেকে বিয়ে করে নেয় তখন জীবন্ত লাশ হয়ে বেচে থাকতে হয়	1
কথা দেয় অনেকে থাকবে শেষ পর্যন্ত কয়জন পারে বল	1

Fig 3.2: Depressive text.

We collect a significant amount of data from YouTube, including movies, songs, and comments that are related to depression. Sometimes, we watch full movies for data collection purposes, and at other times, we listen to songs and memorize lyrics that pertain to depression. Additionally, we collect data from comments left by individuals who may be experiencing depression. Next, we turn to Facebook, as it allows students to interact and discuss course content at any time. Through Facebook, students can post questions, share information, and seek help from peers during their study time or when working on assignments. We collect the majority of our data from Facebook, as it is a widely used platform in our country and individuals tend to easily share personal information on it. We collect data from various posts, pages, and videos on Facebook. It's important to note that the use of data from social media platforms should always be done with caution and respect for the users' privacy and data protection laws, and the data should be collected ethically and legally. The use of the data should also comply with the terms and conditions of the social media platforms. Models for text processing, categorization, data preprocessing, and data extraction are utilized gradually. We divided the nearly 5,604 Bangla-language comments we collected from different social platforms. After that depressive data are defined with 1 and non-depressive data defined with 0. Then, we remark the whole data into training and testing groups. After that, the data was labeled in a vectorizer form for the term frequency and in inverse document frequency and trained with machine learning algorithms. So that, we can anticipate the responses are depressive or not from testing data. Here, we used six classifiers: SVM, CatBoost, Linear Regression, DT, KNN, and multinomial NB.

3.3 Statistical Analysis

We focused on the importance of accuracy in scientific experiments and its impact on the results. Precision, recall, and F1 measure are important indicators of the performance of a model in classifying positive and negative samples. The precision is the ratio of correctly classified positive samples to the total number of positive predictions made by the model. The recall measures the model's ability to correctly identify positive samples and places a high priority on reducing false negatives.

Figure 3.3: Comparison Models

Models		Precision	Recall	F1 measure
Support Vector Machine	Depression	0.75	0.85	0.80
	Non-Depression	0.75	0.63	0.69
Logistic Regression	Depression	0.75	0.84	0.79
	Non-Depression	0.75	0.63	0.68
CatBoost	Depression	0.70	0.85	0.77
	Non-Depression	0.74	0.54	0.63
Multinomial Nave Byes	Depression	0.73	0.89	0.81
	Non-Depression	0.80	0.57	0.66
K- Neighbors	Depression	0.64	0.50	0.54
	Non-Depression	0.58	0.72	0.68
Decision Tree	Depression	0.68	0.77	0.72
	Non-Depression	0.65	0.54	0.59

The F1 measure is the harmonic mean of precision and recall, which balances both values to give a better overall picture of the model's performance. Our study used different algorithms such as SVM, DT, LR, RF, NB, and KNN, to compare their performance on different datasets of depressed and non-depressed texts. The results were optimized by using the TF-IDF feature representation. Table 1 summarizes the precision, recall, and F1 scores for all algorithms. We measured the precision, recall, and F1 scores for both depressed and non-depressed texts to get a comprehensive understanding of the performance of each algorithm.

3.4 Proposed Methodology

Data mining techniques were applied in the study to evaluate the data and identify any signs of depression. These techniques involve the use of various methods to extract valuable information from large and complex datasets, such as identifying patterns, trends, and relationships in the data. In this particular study, data mining techniques were used to assess the data and determine whether or not it was indicative of depression. Finding information related to depression is our main objective at this point. After collecting data from social platforms such as YouTube, Facebook, Instagram, and Reddit, it was organized and placed into an Excel spreadsheet. The data was labeled, with depressive data being assigned a target value of 1 and non-depressive data being assigned a target value of 0, this step is important as it helps to define the problem and give the model a clear understanding of what it needs to learn. After labeling, the data was pre-processed to remove any unnecessary information such as punctuation, null values, and regular expressions, this step is crucial as it helps to improve the quality and accuracy of the data, making it more suitable for machine learning analysis. Once pre-processing was completed, the data was cleaned to remove any duplicate or irrelevant data, this step is important as it helps to prevent any inaccuracies or biases in the model's training. Finally, after cleaning the data, it was split into training and test sets. This allows for a more realistic assessment of how well the model will perform when presented with new data. This process is essential for preparing the data for machine learning analysis and it helps to ensure that the data is of high quality and ready to be used to train a model. Once the process of removing punctuation was completed, the data

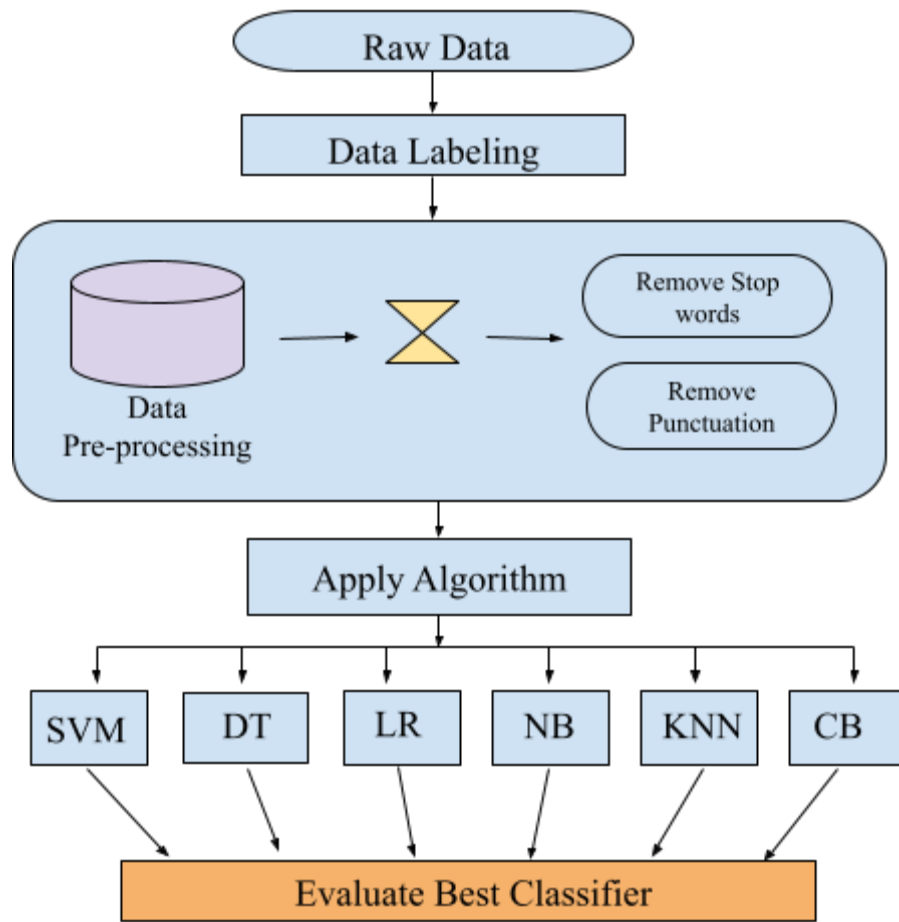


Fig 3.4: Workflow Diagram

was divided into two sets, one for training and one for testing. The training set consisted of 80% of the data, while the remaining 20% was used as a test set. The idea behind this split is to use the majority of the data to train the model, and then evaluate its performance on a smaller, unseen set of data. This allows for a more realistic assessment of how well the model will perform when presented with new data. This process is a common practice in machine learning, as it helps to ensure that the model is generalizing well and not overfitting to the training data. After applying the algorithm, we need to evaluate the best classifier based on its performance.

3.5 Implementation Requirements

The goal of this phase is to reduce noise from the data and enhance the meaningful words and patterns. To do this, we used string manipulations to remove punctuation from the data such as [!,",#,\$,%,&',(,),*,+,@,]. Stopwords are words that are often used in a language but do not have a major significance. We used Natural Language Processing (NLP) to eliminate Stop words, which include the letters

```
[ 'যথেষ্ট' : ",  
  'টি' : ",  
  'মোট' : ",  
  'সুতরাং' : ",  
  'অথচ' : ",  
  'অতএব' : ",  
  'অথবা' : ]
```

are typically relatively common terms in the Bangla language. Although these terms are frequently used in natural language, where they are used as regular expressions, they offer no relevant information and are frequently useless for tasks like information retrieval, text categorization, and machine translation. In order to build features for the ML algorithm to find patterns in the data, feature engineering was performed. The term frequency-inverse document frequency (TF-IDF) approach, which transforms the text data into numerical values, was used to label the dataset. A mathematical formula called TF-IDF (term frequency-inverse document frequency) determines how important a term is to each document in a group of documents. It has several uses, with natural language processing (NLP) being the most important, including word counting in machine learning tasks. By combining two separate metrics—the frequency of the phrase in the document and the inverse document frequency of the term in a collection of documents—TF-IDF is calculated for each term in a document. A term's frequency is calculated by counting how many times it appears in a document, whereas a term's inverse document frequency indicates how frequently or infrequently it appears in all documents. Different supervised learning classifiers were employed in this research, logistic regression, multinomial naive Bayes, Support Vector Machine, Decision trees, catBoost, and KNN.

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experimental Setup

We separated the almost 5,604 comments in Bangla that we gathered from various social media sites. Moving forward, depressive data were labeled with 1 and non-depressive data with 0. After that, we trained our model with 80 percent of the data, while the remaining 20 percent was used for testing.

4.2 Experimental Results & Analysis

The confusion matrix is used to evaluate the performance of the dataset, and the following metrics are discovered: high accuracy, recall, F1 ratings, the confusion matrix, and curves. The model's precision is a metric indicating how precise, smart, and rapid it is at predicting a given section. Recall is a metric used to determine how frequently a model is able to recognize a specific category. The F1 score or metric is composed of the average of precision and recall. In binary classification, where the computer tries to determine whether an email is a spam or not. There would be two rows (one for spam and one for not spam) and two columns in the confusion matrix for this issue (one for predicted spam and one for predicted not spam). The number of emails that were actually spam and those that were expected to be spam (true positives) would be the entries in the matrix, while the number of emails that were genuinely not spam but were expected to be spam (false positives), were spam but were expected to be not spam (false negatives), and were actually not spam and predicted to be not spam (false negatives) would also be entries in the matrix (true negatives). Accuracy, precision, recall, and F1 score are just a few of the evaluation metrics that may be computed using the confusion matrix. It is a helpful tool for figuring out a classification algorithm's advantages and disadvantages as well as evaluating how well various algorithms perform.

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

Support Vector Machine: A supervised learning technique used for classification or regression is called a support vector machine (SVM). To categorize fresh data points, they locate the hyperplane in high-dimensional space that divides several classes. High-dimensional data and imbalanced or noisy classes work well with SVMs. For improved performance on a particular dataset, they have hyperparameters that can be changed, including the regularization parameter and kernel function.

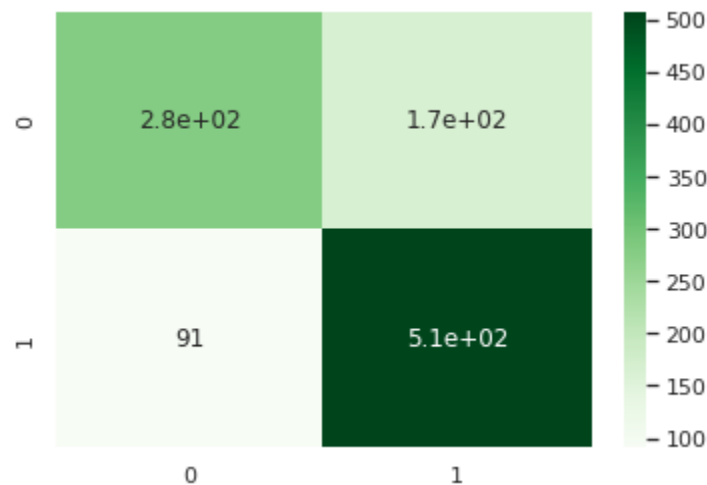


Figure 4.1: Confusion Matrix for SVM classifier

Here, the TP rate is 282, the FP rate is 172, the FN rate is 91, and the TN rate is 512.

Multinomial Naïve Bayes: When a multinomial distribution of the attributes is present, multinomial Naive Bayes is used. The Multinomial Naive Bayes procedure is the most widely used probabilistic knowledge method in Natural Language Processing. The Bayes theorem is the foundation of the approach, which determines the label of a text such as an electronic message or paper piece. It calculates the probabilities of each tag for a given sample and then generates the tag with the highest likelihood. The chance that a certain event will occur based on previously known circumstances is controlled by the Bayes theorem, which was developed by Thomas Bayes.

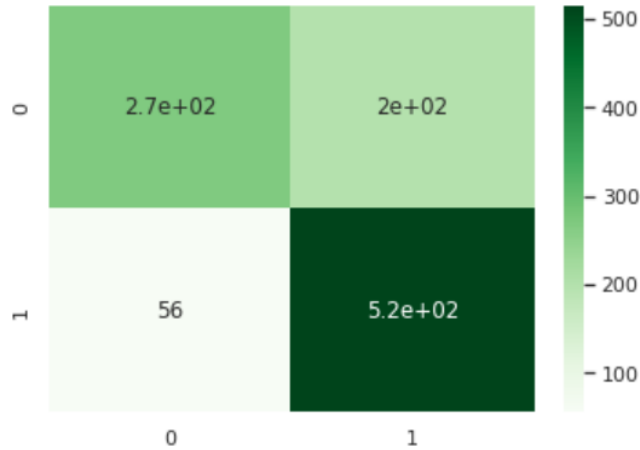


Figure 4.2: Confusion Matrix for multinomial NB classifier

Here, the TP rate is 272, the FP rate is 202, the FN rate is 56, and the TN rate is 522.

Decision Tree: The algorithm first divides the data based on the value of the most significant feature at each node in the tree, and then the data is recursively split at each child node depending on the values of the other features. The decision or prediction produced by the model is represented by the leaves, the tree's final nodes. Decision trees are a straightforward and understandable machine learning method that may be applied in a variety of situations. They can manage categorical data and complex tasks, which makes them very effective for categorization tasks.

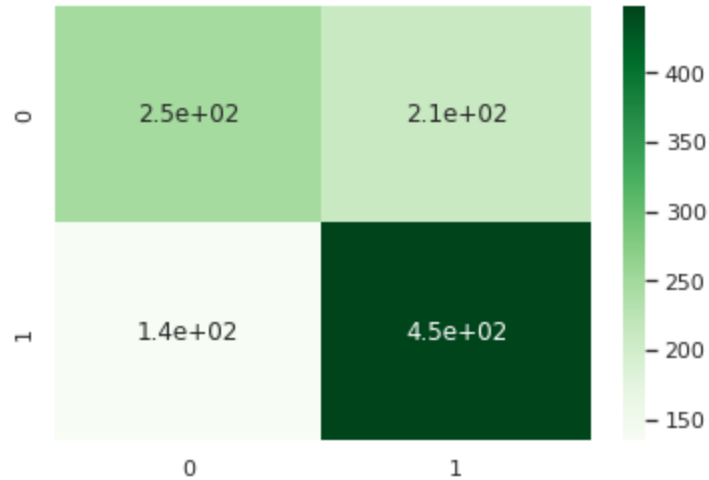


Figure 4.3: Confusion Matrix for DT classifier

Here, the TP rate is 252, the FP rate is 212, the FN rate is 142, and the TN rate is 452.

Logistic Regression Model: The outcome in a replication of logistic regression with a dual condition is often either 0 or 1. Logistic regression (LR) collects pertinent traits from text input when building a logistic regression model. Finally, we discussed how to assess the model's accuracy and forecast how effectively it will perform given fictitious data.

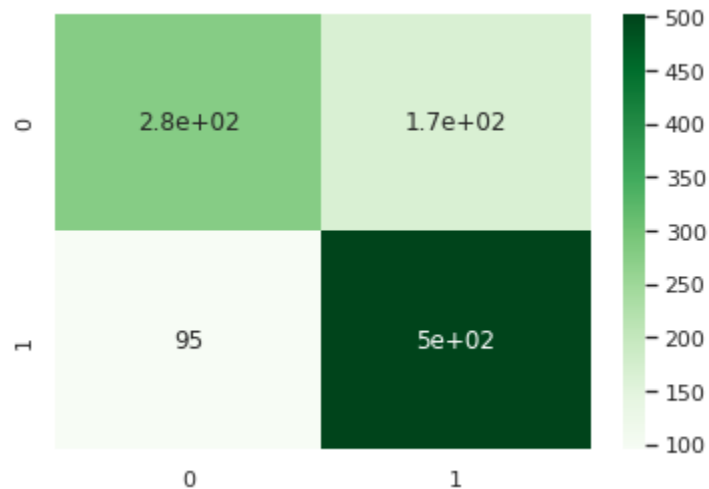


Figure 4.4: Confusion Matrix for LR classifier

Here, the TP rate is 282, the FP rate is 172, the FN rate is 92, and the TN rate is 502.

CatBoost: CatBoost is an open-source gradient boosting library that aims to be quick, scalable, and precise. Similar to other gradient-boosting methods, CatBoost creates a group of prediction trees. However, compared to conventional gradient boosting methods, it has a number of enhancements that increase its effectiveness. Additionally, it contains a number of regularization methods that can enhance the model's capacity for generalization and help avoid overfitting.

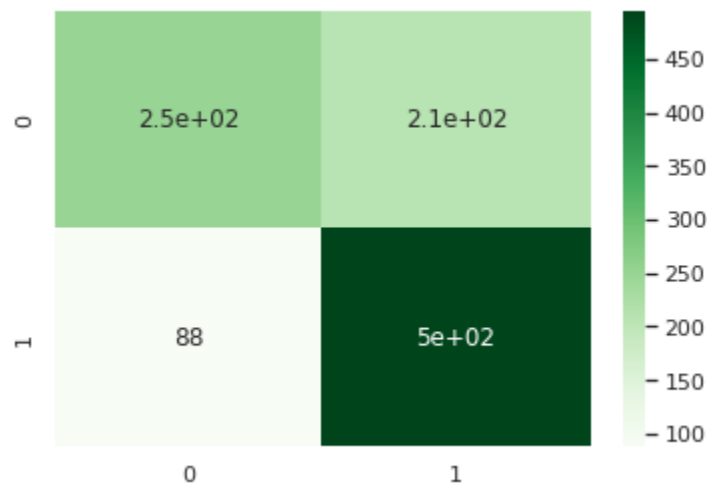


Figure 4.5: Confusion Matrix for CatBoost classifier

Here, the TP rate is 250, the FP rate is 212, the FN rate is 88, and the TN rate is 502.

KNN: The supervised machine learning method K-Nearest Neighbors (KNN) is used for classification and regression tasks. Instead of creating a model from the data, it works by keeping the training data and using it immediately to generate predictions. To classify a new case, KNN looks at the k nearest cases (determined by a distance function) and assigns the new case to the most common class among those neighbors. On small, high-dimensional, or noisy datasets, KNN is easy to use and successful, but it can be computationally expensive and sensitive to the choice of distance function and the value of k.

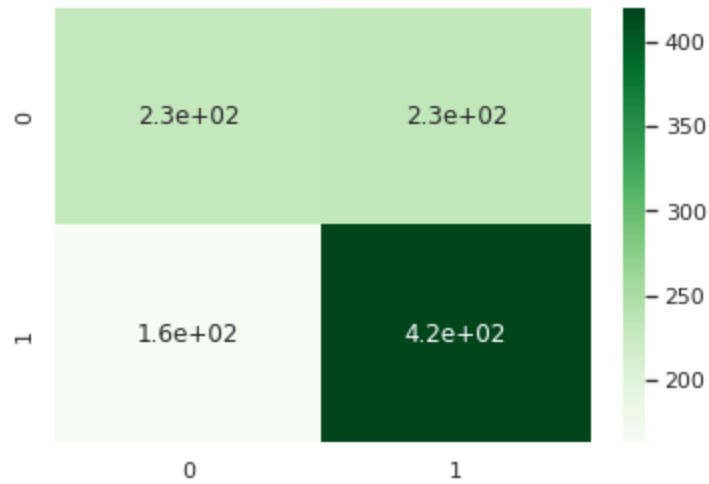


Figure 4.6: Confusion Matrix for KNN classifier

Here, the TP rate is 232, the FP rate is 232, the FN rate is 162, and the TN rate is 422.

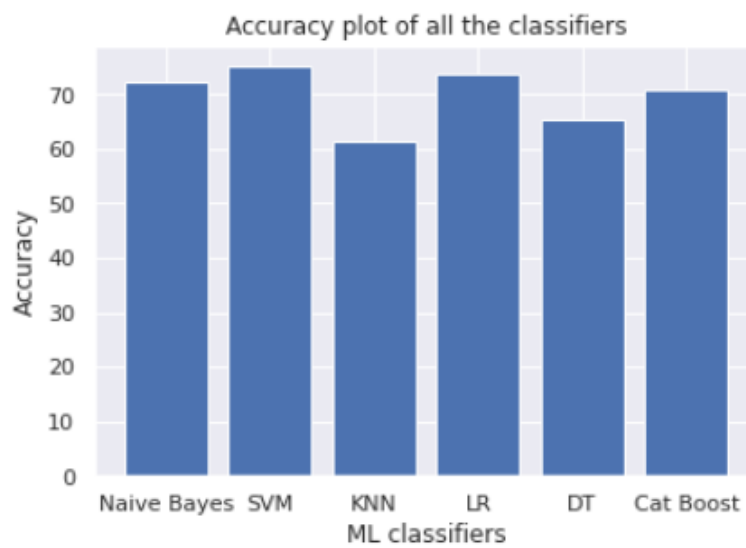


Fig 4.7: Accuracy plot of all the classifiers.

In this study, we evaluated the performance of six different machine learning algorithms: Naive Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), and CatBoost (CB). The accuracy of each algorithm was calculated as the fraction of correct predictions made by the model out of all predictions.

Table 4.8: Accuracy Different Models in Percentage

Model	Accuracy
SVM	75.28%
CB	74.42%
LR	74.33%
NB	73.18%
DT	68.00%
KNN	63.88%

Our evaluation's findings demonstrated that the SVM algorithm had the maximum accuracy, scoring 75.28% score. The LR, CB, NB, and DT algorithms had similar accuracy, respectively 74.33%, 74.42%, 73.18%, and 68%. The KNN algorithm had the lowest accuracy, at 63.88%. These results suggest that the SVM algorithm is the best choice for this particular dataset, although the other algorithms also performed reasonably well.

SVM gives the highest F1 measure of 0.80 for depressive data and 0.69 for non-depressive data is the harmonic mean of precision and recall, and it is used to balance the trade-off between precision and recall.

Table 4.9: F1 Measures of six classifiers

Models		F1 measure
Support Vector Machine	Depression	0.80
	Non-Depression	0.69
Logistic Regression	Depression	0.79
	Non-Depression	0.68
CatBoost	Depression	0.77
	Non-Depression	0.63
Multinomial Nave Byes	Depression	0.81
	Non-Depression	0.66
K- Neighbors	Depression	0.54
	Non-Depression	0.68
Decision Tree	Depression	0.72
	Non-Depression	0.59

A high F1 measure indicates that the models have a good balance between precision and recall, which means that they are good at both avoiding false positive predictions and finding all actual

cases of depression. This is important because it ensures that the models are making accurate and comprehensive predictions.

4.3 Discussion

Depression is a severe mood condition that has an impact on one's feelings, thoughts, and actions. The majority of people who commit suicide are middle-aged individuals. But serious depression can affect both young and old people. Stress, loss, family history, abuse, and heredity are some of the factors that might cause the condition. Due to a variety of barriers, including difficulty accessing therapy and the requirement for several drugs, many people do not seek treatment for depression despite the fact that it is a common condition. Prioritizing face-to-face interactions, exercise, downtime, a balanced diet, and enough sleep are essential for improving mental wellness. Large data sets can be analyzed using machine learning algorithms, which can then be utilized to make mental health-related insights and predictions.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

One of the main causes of disability worldwide and a major contributor to global sickness is depression. Compared to men, women are more likely to experience depression. Depression can lead to suicide, and there is effective treatment available for different levels of severity. All medical professionals should be educated about the numerous facets of depression due to the disorder's high prevalence. Depression has always been a problem in society and can cause social exclusion, even from close friends and family. Depression makes it more difficult to seek and receive aid since those who experience it may struggle to feel close to others and may feel like a burden to their friends and family.

5.2 Impact on the Environment

The development of depression can also be influenced by environmental variables such as pollution, natural catastrophes, and climate change. It has been demonstrated that nature has beneficial effects on mental health, including lowering anxiety and sadness. This has been proven by the use of ecotherapy, which includes participating in outdoor activities in the natural environment. Additionally, studies have shown that bright light areas, both natural and artificial, can boost mood and lessen signs of anxiety and depression. For instance, it has been found that people who are exposed to air pollution have higher rates of depression, and people who suffer natural catastrophes may have traumatic experiences that make them depressed. The effects of depression on the environment can be lessened by taking antidepressants, getting treatment, and engaging in outside activities. It has been demonstrated that therapies that involve nature, such as ecotherapy, are very successful at easing the symptoms of depression. Additionally, lowering the risk of depression and its accompanying environmental effects can be achieved through expanding access to mental health treatments in environmentally susceptible locations.

5.3 Ethical Aspects

Many ethical concerns regarding patient safety, effective illness management, and the restoration of individual autonomy are raised by the treatment of depression, which includes crises and deeply ingrained feelings of sadness and misery in patients. The ethical issues of confidentiality, empathy, fairness, non-discrimination, expertise, trust, and other imprecise ideas commonly come up in psychiatrists' day-to-day work. Study in psychiatry must be conducted ethically, keeping the rights and privileges of research subjects in mind, much like research in other areas of medicine. Ethics provides guidelines for professional conduct and decision-making.

5.4 Sustainability Plan

To make sure that the research findings are effectively applied and sustained, collaborate with key stakeholders such as mental health groups, healthcare providers, and governmental organizations. Disseminate the study's findings to a large audience, including those in the mental health field, those in the medical field, decision-makers in politics, and members of the general public, using a variety of media channels like scholarly publications, conferences, and public forums. Keep a close eye on how the research findings are affecting procedures and results in the field of mental health, and make changes as needed. Evaluate and quantify the research's effects throughout time, and adjust as necessary to increase sustainability. Looking for sources of funding that have an emphasis on sustainability and encourage research methods that are respectful of the environment.

CHAPTER 6

SUMMARY, CONCLUSION & FUTURE WORK

6.1 Summary

Gathering the Bengali dataset is a significant challenge also, people share their opinions, including a lot of spelling mistakes, and while we can effectively construct a model that yields quite excellent outcomes, our model has a few drawbacks too. The likelihood of accuracy could be significantly increased if we can gather a lot of datasets. Because there are no Bengali datasets available online, it is very time-consuming to collect them manually from social media.

6.2 Conclusion

The method we've outlined in the research paper focuses on analyzing and assessing Bengali text from social networks to classify it into two categories: depressive and non-depressive. To do this, six different types of machine learning classification methods are utilized, including SVM. According to the evaluation's findings in the research report, SVM had the highest accuracy among the six machine learning algorithms employed to categorize the Bengali text, scoring 75.28%. Although a few algorithms had lower accuracy scores of 74.33%, 74.42%, 73.18%, and 68%, the LR, CB, NB, and DT algorithms still did well. With a score of 63.88%, the accuracy of the KNN method was the lowest. The SVM algorithm is the best option for this particular dataset, according to these results, even though the other algorithms performed quite well. The SVM algorithm's high accuracy score enables it to accurately classify a large portion of the text samples as depressive or non-depressive, offering important insights into depression in the Bengali-speaking community. The use of this model is expected to provide better predictions and higher precision in the classification of the text. By leveraging the power of machine learning, the research aims to accurately categorize the text and gain deeper insights into depression in the Bengali-speaking community.

6.3 Implication for Further Study

The use of machine learning algorithms, which can be trained to spot patterns in the data that are indicative of specific genders and emotional states, is another way that gender and depression levels may be quantified using social media data. A large collection of social media postings that have been classified as coming from men or women, for instance, might be used to train an algorithm, which would then use this training to determine the gender of fresh, unseen posts. Similar to this, an algorithm may be trained on a dataset of social media postings that have been annotated with the emotional state of the author (such as happy, sad, or depressed), and then utilize this training to anticipate the emotional condition of a fresh, unread messages.

Reference:

- [1] Bhattacharjee, D., Kawsher, J., Labib, M.S. and Latif, S., 2020. Machine Learning Techniques for Depression Analysis on Social Media-Case Study on Bengali Community. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE.
- [2] Sau, A. and Bhakta, I., 2019. Screening of anxiety and depression among the seafarers using machine learning technology. *Informatics in Medicine Unlocked*, 16, p.100149.
- [3] Cunningham, S., Hudson, C.C. and Harkness, K., 2021. Social media and depression symptoms: a meta-analysis. *Research on child and adolescent psychopathology*, 49(2), pp.241-253.
- [4] Uddin, A.H., Bapery, D. and Arif, A.S.M., 2019, July. Depression analysis from social media data in Bangla language using long short term memory (LSTM) recurrent neural network technique. In *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)* (pp. 1-4). IEEE.
- [5] Biradar, A. and Totad, S.G., 2018, December. Detecting depression in social media posts using machine learning. In *International Conference on Recent Trends in Image Processing and Pattern Recognition* (pp. 716-725). Springer, Singapore.
- [6] Islam, M.R., Kabir, M.A., Ahmed, A., Kamal, A.R.M., Wang, H. and Ulhaq, A., 2018. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6, pp.1-12.
- [7] Aldarwish, M.M. and Ahmad, H.F., 2017, March. Predicting depression levels using social media posts. In *2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS)* (pp. 277-280). IEEE.
- [8] Choudhury, A.A., Khan, M.R.H., Nahim, N.Z., Tulon, S.R., Islam, S. and Chakrabarty, A., 2019, June. Predicting depression in Bangladeshi undergraduates using machine learning. In *2019 IEEE Region 10 Symposium (TENSYP)* (pp. 789-794). IEEE.
- [9] Khan, M.R.H., Afroz, U.S., Masum, A.K.M., Abujar, S. and Hossain, S.A., 2020, July. Sentiment analysis from bengali depression dataset using machine learning. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.
- [10] Vasha, Z.N., Sharma, B., Esha, I.J., Al Nahian, J. and Polin, J.A., 2023. Depression detection in social media comments data using machine learning algorithms. *Bulletin of Electrical Engineering and Informatics*, 12(2), pp.987-996.
- [11] Tuhin, R.A., Paul, B.K., Nawrine, F., Akter, M. and Das, A.K., 2019, February. An automated system of sentiment analysis from Bangla text using supervised learning techniques. In *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)* (pp. 360-364). IEEE.
- [12] Arora, P. and Arora, P., 2019, March. Mining twitter data for depression detection. In *2019 International Conference on Signal Processing and Communication (ICSC)* (pp. 186-189). IEEE.
- [13] Giuntini, F.T., Cazzolato, M.T., dos Reis, M.D.J.D., Campbell, A.T., Traina, A.J. and Ueyama, J., 2020. A review on recognizing depression in social networks: challenges and opportunities. *Journal of Ambient Intelligence and Humanized Computing*, 11(11), pp.4713-4729.

- [14] AlSagri, H.S. and Ykhlef, M., 2020. Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems*, 103(8), pp.1825- 1832.
- [15] Tadesse, M.M., Lin, H., Xu, B. and Yang, L., 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7, pp.44883-44893.
- [16] Feightner, J.W. and Worrall, G., 1990. Early detection of depression by primary care physicians. *CMAJ: Canadian Medical Association Journal*, 142(11), p.1215.
- [17] Clarke, F.M., Morton, H. and Clunie, G.J., 1978. Detection and separation of two serum factors responsible for depression of lymphocyte activity in pregnancy. *Clinical and Experimental Immunology*, 32(2), p.318.
- [18] Chatterjee, R., Gupta, R.K. and Gupta, B., 2021. Depression detection from social media posts using multinomial Naive theorem. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1022, No. 1, p. 012095). IOP Publishing.
- [19] Wu, C.T., Dillon, D.G., Hsu, H.C., Huang, S., Barrick, E. and Liu, Y.H., 2018. Depression detection using relative EEG power induced by emotionally positive images and a conformal kernel support vector machine. *Applied Sciences*, 8(8), p.1244.
- [20] Smith, M.V., Rosenheck, R.A., Cavaleri, M.A., Howell, H.B., Poschman, K. and Yonkers, K.A., 2004. Screening for and detection of depression, panic disorder, and PTSD in public-sector obstetric clinics. *Psychiatric services*, 55(4), pp.407-414.
- [21] Smys, S. and Raj, J.S., 2021. Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. *Journal of trends in Computer Science and Smart technology (TCSST)*, 3(01), pp.24-39.
- [22] AlSagri, H.S. and Ykhlef, M., 2020. Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems*, 103(8), pp.1825-1832

Mehedi Hasan

ORIGINALITY REPORT

21 %
SIMILARITY INDEX

17 %
INTERNET SOURCES

9 %
PUBLICATIONS

12 %
STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7 %
2	Submitted to Daffodil International University Student Paper	5 %
3	Md. Zia Uddin. "Depression Detection in Text Using Long Short-Term Memory-Based Neural Structured Learning", 2022 International Conference on Innovations in Science, Engineering and Technology (ICISSET), 2022 Publication	1 %
4	Submitted to Oklahoma State University Student Paper	1 %
5	"Proceedings of the International Conference on Cognitive and Intelligent Computing", Springer Science and Business Media LLC, 2022 Publication	1 %
6	beei.org Internet Source	<1 %
7	Submitted to Jacksonville University	