Shahjalal University of Science and Technology

Department of Computer Science and Engineering

Analyzing the Reasons for Depression from Bangla Social Media Contents and Delivering Automated Islamic-Based Counseling

Md. Takrimul Hasan
Reg. No.: 2018331025
4th year, 1st Semester

Rafiqul Ala Mahid
Reg. No.: 2018331092
4th year, 1st Semester

Department of Computer Science and Engineering

Supervisor

M. Jahirul Islam, PhD., PEng.
Professor
Department of Computer Science and Engineering

9th November, 2023

Analyzing the Reasons for Depression from Bangla Social Media Contents and Delivering Automated Islamic-Based Counseling

A Thesis submitted to the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering.

By

Md. Takrimul Hasan
Reg. No.: 2018331025
4th year, 1st Semester

Rafiqul Ala Mahid
Reg. No.: 2018331092
4th year, 1st Semester

Department of Computer Science and Engineering

Supervisor

M. Jahirul Islam, PhD., PEng.
Professor
Department of Computer Science and Engineering

9th November, 2023

Recommendation Letter from Thesis Supervisor

The thesis entitled *Analyzing the Reasons for Depression from Bangla Social Media Contents and Delivering Automated Islamic-Based Counseling* submitted by the students

1. Md. Takrimul Hasan
2. Rafiqul Ala Mahid

is under my supervision. I, hereby, agree that the thesis/project can be submitted for examination.

Signature of the Supervisor:
Name of the Supervisor: M. Jahirul Islam, PhD., PEng.
Date: 9th November, 2023

Certificate of Acceptance of the Thesis

The thesis entitled *Analyzing the Reasons for Depression from Bangla Social Media Contents and Delivering Automated Islamic-Based Counseling* submitted by the students

1. Md. Takrimul Hasan
2. Rafiqul Ala Mahid

on 9th November, 2023, hereby, accepted as the partial fulfillment of the requirements for the award of their Bachelor Degrees.

Head of the Dept.	Chairman, Exam. Committee	Supervisor
Md Masum	Dr. Sadia Sultana	M. Jahirul Islam, PhD., PEng.
Professor	Associate Professor	Professor
Department of Computer Science and Engineering	Department of Computer Science and Engineering	Department of Computer Science and Engineering

Abstract

Depression is a prevalent and significant medical condition that has negative effects on our emotions, thoughts, and behaviors. It can impair our ability to perform at work and at home and result in a variety of mental and physical issues.Depression is neither a sign of weakness nor a personal shortcoming. As a number of efficient therapies that can treat this medical condition. Medication, psychotherapy, or a combination of the two may be utilized to treat depression.Depression analysis is the process of gathering and interpreting information regarding depression. This information can be used to identify risk factors for depression, develop new treatments, and enhance the quality of care for those with depression. Depression is now one of the leading causes of disability in the globe, as its prevalence continues to rise. Depression also has a substantial economic impact, costing billions of dollars annually in delayed productivity and healthcare expenses. Depression analysis can assist us in gaining a deeper understanding of the causes of depression and in developing more effective treatments. In addition, it can help us identify individuals who are at risk for depression so that they can receive early treatment. We collected Bangla depressive data from social media namely facebook and X. In this paper, we trained our dataset with four multi-class classification algorithms to analyze the reasons for depression.The present work is the first to investigate the causes of depression in Bangla social media data and to offer Islamic-based counseling. The results of this study indicate that counseling based on Islamic principles may be an effective means of addressing the causes of depression among Bangla speakers. This study makes a significant contribution to the field of Bangla NLP by developing a new method for analyzing Bangla social media data for depression.

Keywords: Depression analysis, Reasons For Depression(RFD), BNLP, Social Media, Naive Bayes, RNN, LSTM, BERT, DistilBERT.

Acknowledgements

We would like to offer our profound gratitude to M. Jahirul Islam, PhD., PEng., our supervisor, for his leadership and assistance during this study. We are appreciative of his willingness to share his knowledge with us because his knowledge and insights have been priceless to us.

We also acknowledge Mst. Fazilatun Nessa and Md. Meraz Hossain, two psychologists who provided assistance to us in conducting this study. We are appreciative of their willingness to share their knowledge with us because their insights into the psychology of depression have been really helpful to us.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The world's fourth most common cause of death is now considered to be depression.[1].According to the World Health Organisation (WHO), 4.4% of people around the world are suffering from depression, and various types of depression are more common among females compared to males.[2].Depression is often difficult to diagnose because it manifests itself in different ways[3]. Depression is a multifaceted and significant mental health condition that impacts an individual's cognitive processes, affective states, behavioral patterns, and overall physical health. It transcends the typical oscillations in effect that are universally encountered. Depression would be the primary cause of disability in high-income countries like the US in 20 years (Mathers and Loncar 2006).[4]. Individuals who experience depression frequently exhibit enduring feelings of sadness, hopelessness, and a lack of interest in previously appreciated activities. The aforementioned condition can exert a substantial influence on individuals' capacity to engage in normal daily activities, impacting relationships, occupational performance, and general well-being. Depression is a widespread and dangerous mental condition that affects millions of individuals around the world. It is distinguished by recurring emotions of melancholy, hopelessness, and worthlessness. Changes in eating, sleep, energy levels, and focus can all be symptoms of depression. Depression is becoming more common, and it is now one of the top causes of disability worldwide. Depression also has a significant economic impact, costing billions of dollars in lost productivity and healthcare costs each year.

Our thesis work can't say exactly that a person is depressed. But it can detect possible depressed people. As depression is related with a period of time we can't say that a people is must depressed.For a depression diagnosis to be made, the symptoms must have been present for a period of at least two weeks.[5]

A WHO study [Depression and Other Common Mental Disorders, WHO - 2017] estimates that 4.1% of Bangladeshi population (6,391,760) has depression, and 4.4% (6,900,212) has anxiety disorder.[6]

According to the statistics, the number of depressed individuals in Bangladesh is increasing daily. By 2030, depression is likely to be the most prevalent disease worldwide.[7]

Our research is to face the upcoming challenges of depression among the Bengali community people. We are collecting data from social media and trying to improve the accuracy of our model. In BNLP, the task is first where the reasons for depression will be detected from depressive Bangla text.

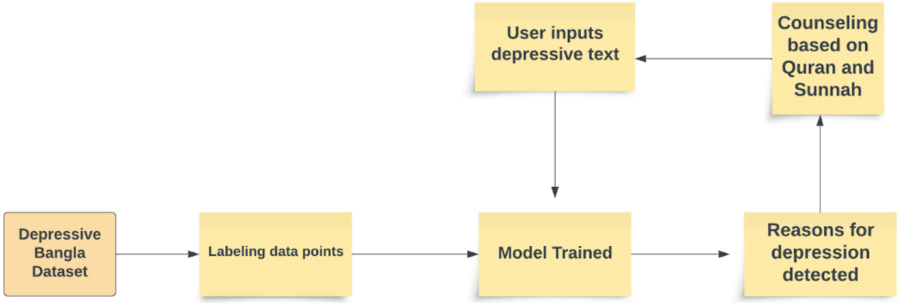

User inputs depressive text
Counseling based on Quran and Sunnah
Depressive Bangla Dataset
Labeling data points
Model Trained
Reasons for depression detected

Figure 1.1: An overview of our work plan

1.1 Motivation

Our thesis work is Analyzing the Reasons for Depression from Bangla Social Media Contents and Delivering Islamic-Based Counseling. The Quran is the holy book, and it contains numerous verses that offer guidance and comfort to those who are afflicted. Quranic counseling is a form of therapy that employs the Quran and Sunnah to assist individuals in overcoming their difficulties. An increasing body of evidence indicates that the teachings of counseling can be an effective treatment for depression. According to a study published in the Journal of Religion and Health, Quranic counseling is just as effective as cognitive-behavioural therapy for reducing the symptoms of depression. According to a second study published in the Journal of Islamic Mental Health, Quranic counseling is effective for enhancing the mental health of depressed individuals. we're interested in the potential of Quranic counseling to assist depressed individuals, which is why we're working on this topic. We believe that the Quran can offer hope and guidance to those who are struggling, and we wish to investigate how this can be utilized to assist individuals in overcoming their difficulties. We also believe that this research is significant because it has the potential to enhance the lives of depressed individuals. Depression is a severe illness that can have detrimental effects on individuals' lives. If Quranic counselling can be shown to be effective in treating melancholy, it could provide those in need a novel and effective treatment option. Our dissertation may make a substantial contribution to the field of mental health. It could help raise awareness of the potential of Quranic counselling and provide depressed individuals with a fresh approach to therapy.

1.2 Research Objectives

The primary objective of our research is to design and implement a system that offers Islamic-oriented counseling services to individuals experiencing depression, by accurately identifying the specific form of depression exhibited by the user. The method will offer immediate alleviation to anyone experiencing depression. Depressed individuals commonly seek guidance from numerous Facebook therapy groups in order to obtain advice on achieving relief from their emotional distress, as shown within the realm of social media. In our nation, the current availability of psychologists and psychiatrists is insufficient to adequately address the counseling and guiding needs of the sizable population affected by mental disorders. Consequently, our proposed approach holds the potential to provide a transformative shift within this domain.

In addition to our thesis work, we have identified the following objectives:

- This study aims to develop the initial framework for a depression detection system specifically designed for the Bengali population, while also providing Islamic-based therapy tailored to different types of depression.
- The approach aims to assist individuals experiencing depression by offering guidance derived from the teachings of the Quran and Sunnah, with the intention of aiding them in overcoming their condition.
- The initial use of BNLP involves the identification of depression subtypes and the provision of Quranic therapy tailored to individuals experiencing depression.

1.3 Novelty of our work

This study is groundbreaking within the realm of Bangla Natural Language Processing (NLP) due to its pioneering utilization of NLP methodologies for the identification of causal factors of depression within Bangla social media material. This work is a challenge owing to the informal nature of the language used. Nevertheless, the utilization of Natural Language Processing (NLP) procedures presents a more effective and adaptable approach in detecting individuals who may be encountering depression. This research endeavor has the potential to address the insufficiency of resources in the field of Bangla NLP.

Chapter 2

Literature Review

The purpose of this literature review is to investigate the various causes of depression and the ways in which they manifest throughout time. This review compiles the literature to shed light on the many factors—biological, psychological, environmental, and social—that contribute to the prevalence of depression.

2.1 Depression Analysis of Bangla Social Media Data

This paper presents a Gated Recurrent Neural Network (GRU)-based method for analyzing melancholy in Bangla social media data. The authors collected data from Twitter, Facebook, and other sources to compile their dataset. They focused on optimizing four hyper-parameters to increase the accuracy of GRU models on a smaller dataset. The study aims to support psychologists and authorities in identifying melancholy among Bangla-speaking social media users and aid researchers in implementing Natural Language Processing tasks using Deep Learning techniques. The paper emphasizes the importance of focusing on GRU model performance over a small dataset and provides insights into tuning hyper-parameters for enhanced performance.This work is the first attempt to utilize a Deep Learning approach for depression analysis in Bangla. Therefore, they cannot directly compare their results with those of other relevant works.[8]. However, the study's hyper-parameter tuning strategy is not dependent on a particular hyper-parameter optimization technique, suggesting that more advanced optimization techniques could enhance the GRU model's performance. The paper's limited data and lack of comparisons with other depression analysis methods or models may limit the generalizability of the findings.

2.2 Sentiment Analysis from Bengali Depression Dataset

The paper discusses sentiment analysis in Bengali text, specifically determining if a paragraph expresses happiness or sadness. Data is collected from social network sites and Bengali blogs to create a Bengali depression dataset. The authors discuss challenges in preprocessing Bengali text and the use of CountVectorizer for tokenization. Six machine learning classification algorithms are applied, with Multinomial Naive Bayes achieving the highest accuracy of 86.67%[9]. An automatic sentiment analysis method is presented, detecting people's sentiment and classifying it as happiness or sadness. However, the research has limitations, including the uncertainty of consistency across Bengali text types and the small dataset used for training, potentially limiting the model's applicability to larger and more diverse data sets. Collecting Bengali data is a significant challenge for research.[9].

2.3 Depression Analysis from Social Media Data in Bangla Language using LSTM

This article explores the use of Long Short Term Memory (LSTM) Deep Recurrent Network to analyze sadness in Bangla social media data, specifically tweets. It demonstrates how hyperparameter adjustment can be beneficial in depression analysis, even with limited data sets. The study also highlights the importance of properly preparing small datasets, such as Bangla tweets, for analyzing sadness and emotions.They gathered 5,000 Bangla tweets through repetitive sampling in which tweets were permitted to be repeated arbitrarily.[10]. However, the paper's limited dataset, consisting of a subset of Bangla tweets, may limit the generalizability of the results and the adaptability of the model. Additionally, the manual labeling process by a single sociology student may increase subjectivity and bias. The quality of the dataset could be improved with a more comprehensive and stringent labeling process involving multiple annotators or experts. The research does not compare alternative methods or approaches to analyzing depression in Bangla social media data, making comparisons challenging. The benefits and drawbacks of the LSTM method would have been clearer with a comparison examination. Additionally, the article does not address the moral and privacy issues that may arise from using social media data to predict depression, such as consent, data privacy, and potential harm to individuals.

2.4 Machine Learning Techniques for Depression Analysis on Social Media

This paper uses machine learning methods to examine depression in the Bengali community using social media data. The authors used Facebook, Twitter, and chat app data to identify Bengali speakers who had posted or commented on depression-related topics. They used various machine learning methods, including Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbors, Naive Bayes, and Logistic Regression, to predict depression. Although the accuracy of these algorithms varied, the overall level of accuracy remained consistent. The paper's novelty lies in its use of social media data to detect depression in the Bengali community, a rare area of research. In this research work, depression on two categories were classified and applied among total of 6 algorithms and every algorithm gives the best accuracy and that was on an average of 90% of test data and 91% for train data[11]. However, the study's limitations include the potential underrepresentation of people with depression in the dataset, which could affect machine learning algorithms. Additionally, the complexity of sadness and its underlying causes may not be fully captured by social media. The study's results are not yet transferable to other languages and cultures.

2.5 Detecting depression and mental illness on social media

The paper reviews recent studies on using social media to predict mental illness, specifically depression. It discusses methods like screening surveys, Twitter diagnosis sharing, and online forums. The authors suggest automated detection methods on social media can help identify depressed or at-risk individuals through passive monitoring, potentially leading to early detection and intervention[12]. The paper also highlights the potential of automated analysis to target individuals with elevated depression scores for further assessment, resources, support, and treatment. The paper discusses the limitations of using social media as a screening tool for mental illness, including low sensitivity and high false positive rates. It also discusses the ethical and legal questions surrounding data ownership and protection, as well as the integration of social media-based screening into care systems. The paper does not address potential biases or limitations in the methods used to detect mental illness on social media, such as the representativeness of the sample or the accuracy of self-report surveys. The paper does not provide a comprehensive analysis of these issues.

2.6 The Reasons for Depression Questionnaire (RFD)

The paper investigates the reliability and validity of the Reasons for Depression Questionnaire (RFD), a self-report instrument designed to evaluate explanations for depression causes. Recent research on depression explanations has demonstrated the significance of client beliefs regarding the cause of their depression.[13]. It provides preliminary normative data for both clinical and non-clinical samples from the United Kingdom, indicating high reliability for all subscales, including a newly added biological subscale. The study finds significant correlations between specific subscales, depression severity, and specific aspects of self-esteem, supporting the validity of the measure. The paper emphasizes the significance of client beliefs regarding the etiology of their depression and how these beliefs can influence the therapeutic process and treatment outcomes. It suggests that a shared explanation of depression can impact the therapeutic alliance and treatment strategy. The RFD is considered a valuable instrument for therapists to understand the causes of depression in their clients. However, the study acknowledges potential weaknesses, such as the small sample size for the additional biological scale and the lack of investigation into test-retest reliability and long-term stability. Additionally, the study does not discuss confounding variables and alternative explanations for the relationships revealed between RFD subscales and depression and self-esteem.

2.7 Congruence between reasons for depression and motivations for specific interventions

This research advances the understanding of the connection between depression causes and the impetus for targeted therapies. The Reasons for Depression (RFD) questionnaire is used to gauge how much clients attribute their depression to mental health issues. The resulting RFD questionnaire consists of 44 items and 8 subscales that can be categorised into two higher-order factors.[14].

The study supports the theory that patients are more motivated to engage in treatment if their motives and therapies are consistent. It isolates individual causes of depression and the corresponding impetuses for treatment, such as unresolved childhood difficulties or biological predispositions. The report emphasizes the value of adjusting interventions to meet the unique needs of each client to boost treatment success rates. However, the findings cannot be generalized beyond the current study due to the small sample size and varied nature of the sample. The study's cross-sectional design and lack of thorough diagnostic evaluations limit its ability to draw firm conclusions about what led participants to seek treatment for depression. The study also acknowledges that its ability to examine the efficacy of reason-matching interventions is constrained because it did not include treatments that would have matched the exact reasons supplied by clients.

Chapter 3

Background Study

In a thesis, the background of study is the part where you lay out the scene and introduce the reader to the subject at hand. It provides a synopsis of what we already know about your thesis topic in terms of research and theory. As the first section of your paper, the background of study should introduce your topic and explain why it merits further investigation. Your readers will gain a better appreciation for the bigger picture and the significance of your research with this information.

3.1 Reasons for Depression Scale Understanding

Depression scales are crucial for accurate diagnosis and treatment planning, providing a standardized method for assessing the presence and severity of depressive symptoms. They help clinicians and researchers differentiate depression from other conditions with similar symptoms, reducing subjective biases and providing a consistent framework for assessment. Accurate assessment of depression severity aids in tailoring appropriate treatment plans, guiding healthcare professionals in determining the most suitable interventions. Regular assessments help clinicians gauge the effectiveness of treatment plans and make necessary adjustments. Depression scales also serve as standardized tools for measuring depression prevalence and severity across different populations, contributing to better understanding of depression's prevalence, risk factors, and outcomes. They also aid in patient communication and contribute to public health planning by identifying trends, disparities, and high-risk groups.

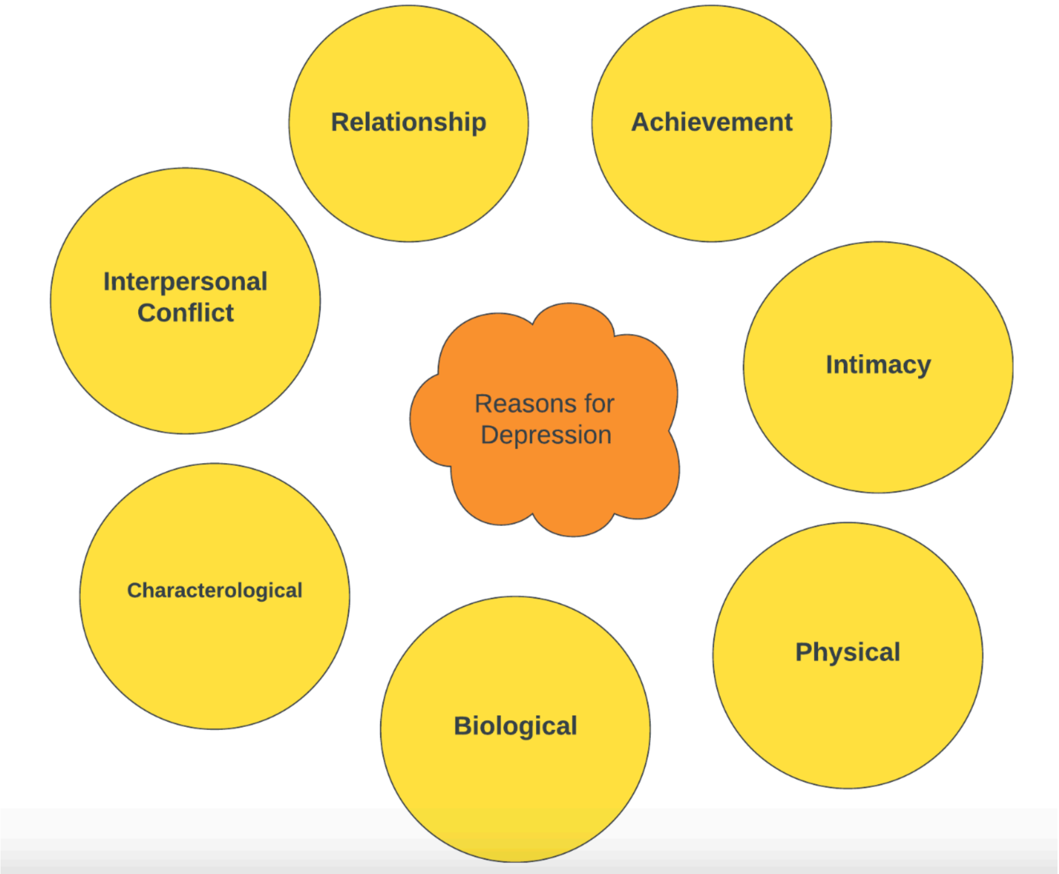


Figure 3.1: RFD Scale

3.2 Natural Language Understanding

Natural language understanding (NLU) is the study of how computers can understand human language, including context and meaning. It is a challenging field due to the ambiguity of human language and its various interpretations. However, NLU has potential applications in automated translation, chatbots, and question-answering programs. Challenges include ambiguity, context, and insufficient information. Natural language varies based on the speaker, context, and subject. Despite these, NLU is a rapidly expanding field with numerous applications. As NLU technology advances, computers will increasingly understand human language, making it a valuable tool in various fields.

3.3 Artificial Neural Network

An Artificial Neural Network (ANN) is a machine learning system that uses input data to create a network of interconnected nodes. The input layer receives raw data, such as numerical values, images, or text. Hidden layers, between the input and output layers, consist of multiple neurons that process the input data. These neurons are connected to form a network of interconnected nodes. Each connection between neurons has an associated weight, representing the strength of that connection. The resulting weighted sum is passed through an activation function, introducing non-linearity to capture complex relationships in the data. Following propagation is the process of passing data through the network, applying weights, activation functions, and producing an output. Artificial Neural Networks (ANN) can optimise a wide variety of coefficients. Consequently, it can accommodate considerably more variability than conventional models[15]. Backpropagation is the key process for training the neural network, adjusting weights based on the calculated error to minimize error. The output layer produces the network's prediction or classification result, with the number of neurons depending on the task. The neural network is trained using a labeled dataset, and its performance is evaluated using a separate testing dataset to ensure it can generalize to new, unseen data.

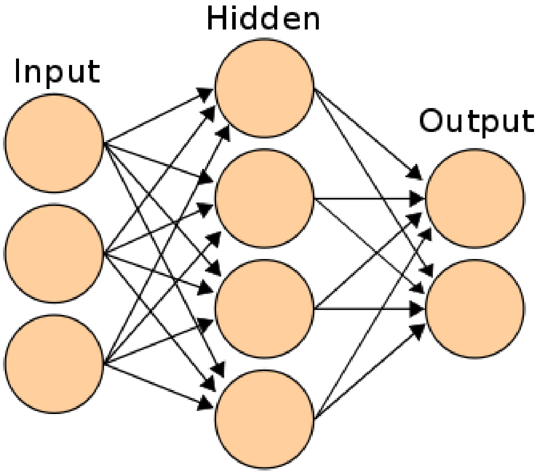


Figure 3.2: Work process of ANN

3.4 Long Short Term Memory (LSTM)

LSTMs are machine learning algorithms that handle sequences by incorporating memory cells and gating mechanisms to retain and selectively update information over time. At a high level, LSTM functions very similarly to an RNN cell[16]. They have a "cell state" that acts as a conveyor belt, allowing information to flow with minimal change. Three gates control the flow of information: the forget gate, input gate, and output gate. These gates determine what information to keep, update, or discard from the cell state. The forget gate decides which information from the previous cell state to forget or retain based on the current input. The input gate determines what new information to store in the cell state, considering the current input. The output gate selectively exposes the cell state's content based on the input and current state. LSTMs use sigmoid and hyperbolic tangent (tanh) activation functions to control gate values and the cell state's content.

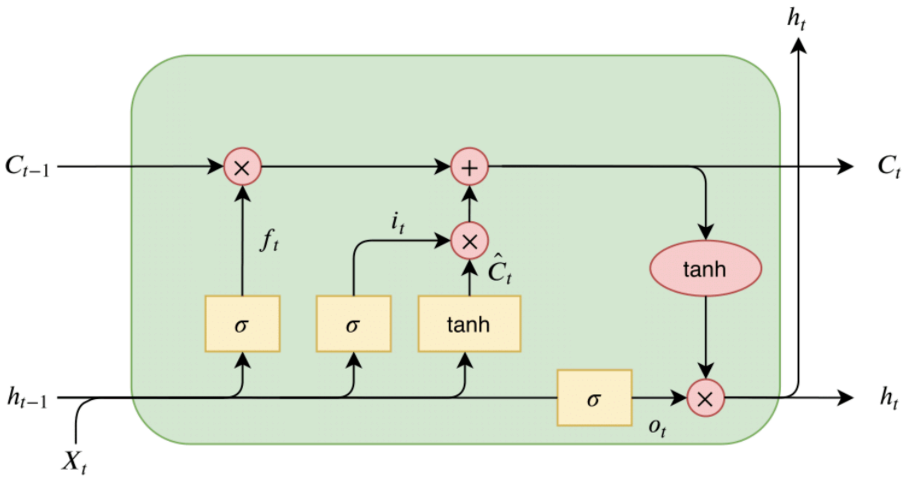


Figure 3.3: LSTM architecture

3.5 Recurrent Neural Network(RNN)

RNNs are a type of machine learning algorithm that processes data in a sequential manner, one element at a time, while simultaneously keeping an internal hidden state that carries knowledge from the steps that came before it. Because they are able to capture interactions and dependencies between items over distinct time steps, they are particularly useful for undertaking activities in which the order in which elements occur in a sequence is important. RNNs make use of weights that are shared across several time steps, which enables them to learn patterns and relationships that are consistent across the entire sequence. An RNN acts as a memory by retaining information from earlier steps by combining the input from each time step with the hidden state from the previous time step. This process occurs at each time step. The combination is then put through an activation function, which results in the production of a new hidden state for the current time step. This state incorporates information from the most recent input as well as information that has been acquired over the course of earlier time steps.

3.6 Bidirectional Encoder Representation from Transformers(BERT)

The Bidirectional Encoder Representations from Transformers (BERT) model is a natural language processing (NLP) system that was developed by Google and is considered to be state-of-the-art. It is a member of the Transformer architecture family and was developed to comprehend the context and meaning of linguistic expression through the use of huge amounts of textual data in its pre-training. The results that BERT has obtained on a wide variety of NLP tasks, such as sentiment analysis, question answering, text categorization, and more, have been nothing short of spectacular. It uses WordPiece tokenization to break words into smaller subword units, and embeds each token into an embedding vector. BERT employs the Transformer architecture, which consists of multiple encoder layers with self-attention mechanisms and feedforward neural networks. It captures contextual information by assigning different weights to tokens based on their relevance. BERT also performs Masked Language Model (MLM) and Next Sentence Prediction (NSP) pretraining, predicting missing words in sentences and learning relationships between sentences. After pre-training, BERT can be fine-tuned on specific tasks by adding task-specific layers and training on datasets.

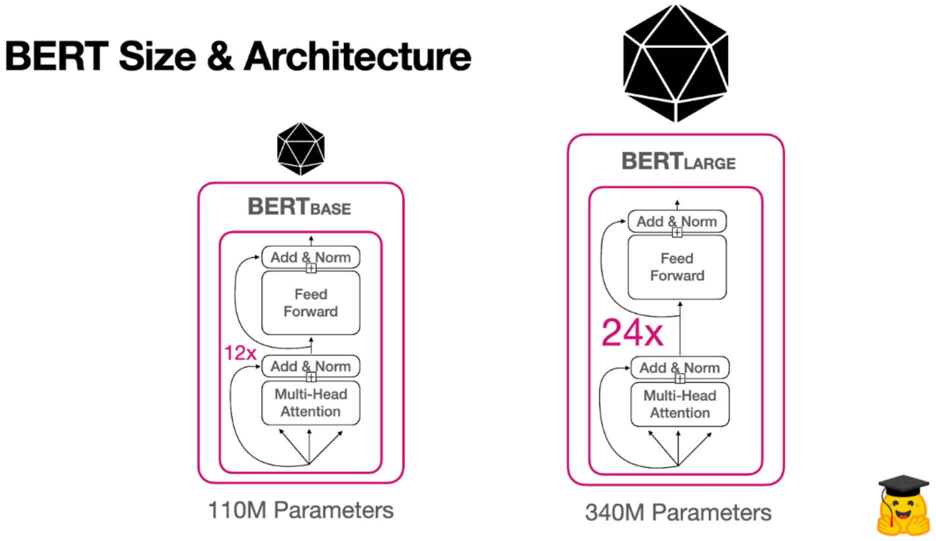


Figure 3.4: BERT size and architecture

BERT is a highly complex and advanced language model that helps people automate language understanding. Its ability to accomplish state-of-the-art performance is supported by training on massive amounts of data and leveraging Transformers architecture to revolutionize the field of NLP[17].

3.7 Tools

3.7.1 TensorFlow

TensorFlow is an open-source machine learning framework developed by Google. TensorFlow is a Python-friendly library for numerical computation that accelerates and simplifies machine learning and neural network development[18]. It is a versatile machine learning framework that uses a computation graph to represent mathematical operations, enabling efficient execution on CPUs, GPUs, and TPUs for parallel and distributed processing. It features an automatic differentiation framework for optimizing model parameters during training, a multi-layer API for building neural networks, and a visualization tool called TensorBoard. TensorFlow also offers high-level APIs like Keras for easier interfaces. Its community contributes to a variety of pre-built models, tools, libraries, and resources. TensorFlow is used in various applications, including image and video analysis, natural language processing, speech recognition, recommendation systems, healthcare, autonomous vehicles, and game AI.

3.7.2 Keras

Keras is an open-source high-level neural network API written in Python. It is a user-friendly API for creating and configuring neural networks with minimal code. It supports various types of neural networks, provides high-level abstractions, and allows advanced customization using low-level TensorFlow functions.It cannot perform low-level computations, so it uses the Backend library to rectify the issue[19]. Keras integrates with TensorFlow, a backend that supports multiple libraries, and includes built-in datasets for training and testing. It also integrates with tools like TensorBoard for visualizing training and performance metrics.

3.7.3 Jupyter Notebook

Jupyter Notebook is an open-source interactive web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Jupyter Notebook offers an interactive environment for writing and executing code in individual cells, allowing for immediate testing and experimentation with data. It supports multiple programming languages, including Python, R, Julia, and more, and supports rich text, including Markdown. Jupyter Notebook also allows for visualizations, such as plots and charts, and can be shared with others. It integrates with popular data science libraries like NumPy, pandas, Matplotlib, and Seaborn, making it a powerful tool for data analysis and manipulation.

Chapter 4

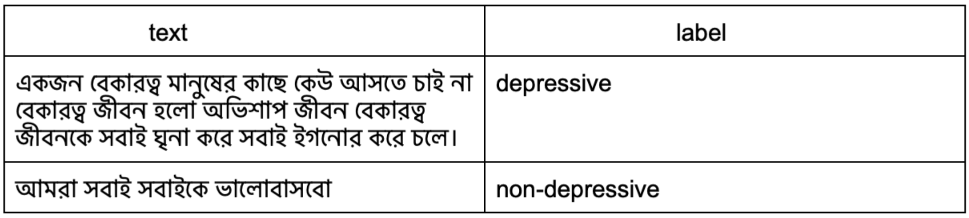
Dataset

4.1 Datasets in Bangla

For effective analysis and providing Islamic-based counseling, a large dataset is needed to train the deep learning models.That's why we need to acquire the essential Bangla text data in order to analyze the causes of depression more effectively and to improve our model accuracy. However, there isn't enough Bengali-language research on relevant datasets at this time. In order to make progress in acquiring better datasets connected to depression in the Bengali language, We are trying to enrich our datasets gradually.

It is very challenging to label Bangla social media depressive content with proper RFD scale. It is very time consuming and the data annotator needs a good visionary sense. First of all, we had to discuss with a psychologist to understand the labeling approach and then we had labeled data following the psychologist's guidance.

In Bangla, have some analysis on depression analysis. But those works are on detecting data entities whether depressive or not. But our thesis work is analyzing those depressive data to find out the reasons behind the depression of individuals. The existing dataset sample is :



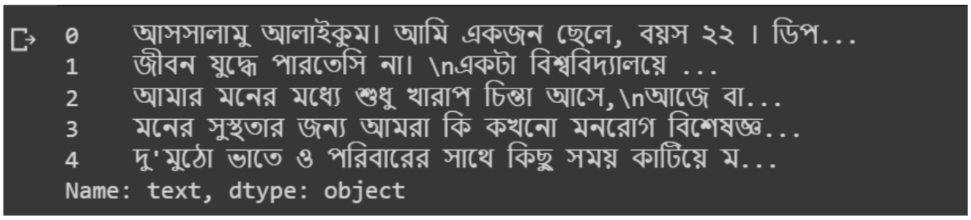
text	label
একজন বেকারত্ব মানুষের কাছে কেউ আসতে চাই না বেকারত্ব জীবন হলো অভিশাপ জীবন বেকারত্ব জীবনকে সবাই ঘৃনা করে সবাই ইগনোর করে চলে।	depressive
আমরা সবাই সবাইকে ভালোবাসবো	non-depressive

Our dataset sample is :

Depressive text	Reasons for depression scale
একজন বেকারত্ব মানুষের কাছে কেউ আসতে চাই না বেকারত্ব জীবন হলো অভিশাপ জীবন বেকারত্ব জীবনকে সবাই ঘৃনা করে সবাই ইগনোর করে চলে	Achievement
আমার একজন এর সাথে রিলেশন আছে তাকে নিয়ে ওভারথিংকিং করি, দুশ্চিন্তা করি	Relationship

4.2 Dataset collection

Dataset collection is one of the challenging phases of our thesis work. To analyze the reasons for depression using deep learning models we need a consistent and large dataset. We have collected data from a facebook group related to depression and preventing suicide. We scraped data from facebook groups. Also collected some data from existing works on depression analysis. But we have to take only the depressive data from the dataset. From 17k data points we were able to take only 3k data points as rest of the data points were related to non depressive data. So we have to create a dataset manually. We are just working with depressive Bangla text data. And collecting only depressive data is not an easy task. It takes a lot of time to create a dataset for our thesis work. We have 5k data points on our dataset. Our goal is to make a dataset with 35k data points. Example of our dataset before preprocessing and labeling :



```
0    আসসালামু আলাইকুম। আমি একজন ছেলে, বয়স ২২ । ডিপ...
1    জীবন যুদ্ধে পারতেসি না। \nএকটা বিশ্ববিদ্যালয়ে ...
2    আমার মনের মধ্যে শুধু খারাপ চিন্তা আসে,\nআজে বা...
3    মনের সুস্থতার জন্য আমরা কি কখনো মনরোগ বিশেষজ্ঞ...
4    দু'মুঠো ভাতে ও পরিবারের সাথে কিছু সময় কাটিয়ে ম...
Name: text, dtype: object
```

Figure 4.1: Dataset labeling

Reasons for Depression (RFD)	Examples
Relationship	• My family treated me poorly as a child. • I haven't worked through things that happened to me as a child. • My spouse/partner treats me poorly. • My spouse/partner doesn't understand me.
Achievement	• I've failed to achieve a specific goal I set for myself. • I am not fulfilling my potential. • I can't get done the things I should be able to. • I haven't done anything important in my life.
Interpersonal Conflict	• Other people criticize me. • Other people don't like me. • I can't make friends. • People don't give me the respect I deserve.

Table 4.1: Reasons for Depression and Examples

4.3 Dataset labeling and preprocessing

After collecting data points from facebook public groups we have checked every data point manually if a data point is related to depression or not. If not we have discarded the data from our dataset. Among existing dataset we have taken only depressive data points and the rest of the data points get dropped out. Then we have annotated our final dataset following the guidance of two psychologists. We have annotated each and every data point manually. Annotating data is not an easy task so it takes a lot of time. We labeled our data according to RFD scale following the research paper[#]

Reasons for Depression (RFD)	**Examples**
Characterological	• I think about things in a depressing way. • No one really understands me. • This is the way I respond when things get tough. • I pay more attention to the bad things in my life than the good things.
Biological	• My nervous system is just wired this way. • I inherited it from my parents. • I have always been this way. • It's a biological illness.
Physical	• I don't get enough exercise. • I'm not active enough. • I don't take care of myself physically. • I don't eat well enough.
Intimacy	• Loss of a loved one. • Going through a breakup or divorce. • I have been hurt by past friends who betrayed my trust. • My partner is emotionally unavailable. • Friendship breakdown.

Table 4.2: Reasons for Depression and Examples

Chapter 5

Methodology

Multi-class text classification models

5.1 Training Naive Bayes Model

Data Loading and Preprocessing

- Using the `pandas` library, load the dataset from an Excel file.
- To assure data quality, remove rows with blank values in the ”text” or ”label” columns.

Data Preparation

- Distinguish the feature (text) and target (label) variables from the dataset. The target variable (’y’) contains the labels that go with the text data in the feature variable (’X’).

Data Splitting

- Use the `train_test_split` function to split the dataset into training and testing sets. This makes it easier to assess how well the model works with unknown data.

Using TF-IDF to Vectorize Text

- Textual data is transformed into numerical vectors using the TF-IDF (Term Frequency-Inverse Document Frequency) approach.

- Each text document is converted into a vector, with each dimension representing the TF-IDF score for a particular phrase.

- TF-IDF takes into account a word's relevance in a document in relation to its frequency across the entire corpus.

Initializing a Multinomial Naive Bayes Classifier for Training the Naive Bayes Model

- Use the training data and accompanying labels to train the model.

- Naive Bayes makes use of the Bayes theorem to estimate the likelihood of a label given a set of features (in this case, words).

Model Evaluation and Classification Report

- Using the trained model, forecast labels for the testing data.

- Assess the model's effectiveness based on a classification report.

- The classification report includes metrics for each class, including precision, recall, F1-score, and support.

The accuracy was : 44%

5.2 Training SimpleRNN

Data Collection: The first step is to collect a dataset of Bangla social media posts that have been labeled with their corresponding depression class.

Data Preprocessing: The next step is to preprocess the data. This includes tokenizing the text and padding the sequences to a fixed length.

Model Training: The SimpleRNN is then trained on the preprocessed data. The training process involves adjusting the weights of the SimpleRNN so that it can accurately predict the class of a given text sequence.

Model Evaluation: The performance of the SimpleRNN is evaluated on a held-out test set. This is done by feeding the test set to the SimpleRNN and calculating the accuracy of the predictions.

Hyperparameters Used:

Hyperparameter	Value
Embedding Dim	64
Hidden Units	64
Number of Classes	7
Epochs	20
Loss Function	Sparse Categorical Crossentropy
Validation Split	0.2
Optimizer	Adam

Table 5.1: Hyperparameter vs Value in SimpleRNN

Model Accuracy with Different Number of Epochs:

Num of Epochs	Test Loss	Accuracy
3	1.5423	0.3993
5	1.5788	0.40257
10	1.5722	0.4154
15	1.9276	0.36876
20	1.5763	0.43800

Table 5.2: Comparison Accuracy between Number of Epocs

We achieved the highest accuracy of approximately 44% using 20 epochs.

5.3 Training LSTM-RNN Model

Data Loading:

- Load data from an Excel file.

Handling Missing Values:

- Remove rows when the ”text” or ”label” columns have missing values.

Tokenization and Padding:

- Tokenize the text data and ensure consistent length by padding sequences.

Data Conversion:

- Convert text sequences and label data into numerical format.

Train-Test Split:

- Divide the data into training and testing sets following standard rules.

LSTM Model Architecture:

- Build a sequential model with an embedding layer to convert words into vectors.
- Stack LSTM layers to capture sequential patterns in the data.
- Introduce non-linearity with dense layers.
- Use dropout to prevent overfitting.
- Final output layer provides class probabilities.

Model Compilation:

- Compile the model with an appropriate optimizer and loss function.
- Define evaluation metrics.

Model Training:

- Train the model using training data.

Hyperparameters Used in LSTM-RNN Model:

Hyperparameter	Value
max_words	10000
maxlen	200
embedding_dim	64
Number of LSTM Layers	2
LSTM Units per Layer	64
Dense Layer Units	64
Dropout Rate	0.5
Number of Classes	7
Number of Epochs	10
Batch Size	16
Optimizer	Adam
Loss Function	Sparse Categorical Cross-Entropy

Table 5.3: Hyperparameter vs Value in LSTM-RNN

LSTM Model Accuracy with Different Epochs and Batch Size:

Number of Epochs	Batch Size	Accuracy
5	32	40.58%
10	32	40.74%
5	16	39.61%
10	16	40.74%

Table 5.4: Compparison Accuracy vs Num of Epochs

The highest accuracy achieved from LSTM is 40.74%.

- Validate the model's progress using a subset of the training data.

Model Evaluation:

- Assess the trained model's performance on unseen testing data.
- Measure loss and accuracy metrics.

5.4 Training BERT Model

The initial step in this procedure involves the loading and preprocessing of the data.

- The dataset is loaded from an Excel file utilizing the `pandas` package.

- To proceed with the analysis, it is necessary to extract the text data (`X`) and the corresponding label data (`y`) from the dataset that has been loaded.

- The dataset should be divided into training and testing sets using the `train_test_split` method provided by the `sklearn.model_selection` module.

In the second step, the pre-trained BERT model and tokenizer are loaded.

- To load a pre-trained BERT tokenizer ('bert-base-uncased'), the `BertTokenizer` from the `transformers` library is employed.

- To ascertain the count of distinct classes inside the label data, it is necessary to identify the number of unique labels for the purpose of classifying the data.

- To perform sequence classification, the `TFBertForSequenceClassification` class from the `transformers` library is utilized to load a pre-trained BERT model.

The third step involves the process of tokenization and dataset preparation.

- The training and testing text data should be tokenized using the tokenizer that has been loaded.

- To ensure the tokenized sequences have a consistent length of 128 tokens, truncation and padding techniques are applied.

- To create TensorFlow datasets, namely `train_dataset` and `val_dataset`, the tokenized data needs to be organized. This may be achieved by constructing a `tf.data.Dataset` from tensor slices.

- To regulate the number of instances processed every iteration during training and assessment, it is advisable to group the datasets into batches.

In this stage, the model is compiled and trained.

- The BERT model is compiled using the Adam optimizer, employing a learning rate of 2×10^{-5} (0.00002).

- The loss function is specified as `'sparse_categorical_crossentropy'` and the metric being monitored is `'accuracy'`.

- The model should be trained by utilizing the `fit` method, which requires the inclusion of both the training dataset and the validation dataset.

- The training epochs have been configured to a value of 3, signifying that the model will undergo three complete iterations over the training data.

In this phase, the model's performance will be assessed and the results will be documented.

- Once the BERT model has been trained, it may be utilized to make predictions on the validation dataset by assigning labels to the respective data points.

- The `np.argmax` function can be employed to extract the predicted labels by choosing the index that corresponds to the highest logit value.

- To generate a classification report, the `classification_report` function from the `sklearn.metrics` package can be utilized.

- The classification report offers a thorough assessment of the model's performance, encompassing precision, recall, F1-score, and support for each individual class.

In this section, we will present the conclusion and results of our study.

- The outputs of the classification report should be presented, with a focus on key metrics such as accuracy, precision, recall, and F1-score.

- Analyze the findings within the framework of the particular classification task, examining the efficacy of the model and identifying potential areas for enhancement.

- In conclusion, the methodology section encapsulates the approach employed, the training process of the model, and the evaluation outcomes.

The given code effectively utilizes a pre-trained BERT model for sequence classification by adhering to a systematic methodology. This methodology encompasses many stages like data loading, preprocessing, model setup, training, assessment, and result reporting.

Hyperparameter Statistics:

Hyperparameter	**Value**
test_size	0.2
max_length	128
batch_size	16
learning_rate	0.00002
num_epochs	3
BERT Pre-trained Model	'bert-base-uncased'
BERT Tokenizer	'bert-base-uncased'

Table 5.5: Hyperparameter vs Value in BERT

The accuracy of the model was : 42.40%

5.5 Training DistilBERT model

This code employs TensorFlow and Hugging Face's DistilBERT for text classification. Data is split into training and testing sets, tokenized with DistilBERT's tokenizer, and transformed into TensorFlow datasets. The model is compiled using an Adam optimizer with a learning rate of 2e-5 and trained for 3 epochs. However, the model's accuracy at 47% suggests potential for hyperparameter tuning or model complexity adjustments to enhance performance.

Hyperparameter	Value
Test size	0.2
Batch Size	16
Maximum Sequence Length	128 tokens
Learning Rate	0.00002
Number of Epochs	3

Table 5.6: Hyperparameters for the Model

Chapter 6

Experimental Results Comparison

In order to analyze the RFD and obtain the findings given in the table, we trained 5 models using our dataset. DistilBERT's transformer model provided us with the most accurate results. According to the comparison, the transformer-based paradigm is better suited for NLP-based activities.

Model Name	Highest Accuracy
Naive Bayes	44%
SimpleRNN	43.8%
LSTM	40.74%
BERT	46.00%
DistilBERT	47%

Table 6.1: Model vs Accuracy

Chapter 7

Future Work

- In order to enhance the precision of the results, it is important to gather a greater quantity of high-quality data. Additionally, employing a range of sophisticated and contemporary models for training the dataset is recommended.
- The objective is to develop an online platform that offers text counseling services rooted in Islamic principles to its users.
- Once the system has been established, we will seek feedback from users in order to enhance the quality of our dataset and service.
- The objective is to establish a comprehensive database for Islamic-oriented counseling services, facilitated by esteemed and highly skilled Islamic experts.
- The accuracy of our system will be enhanced based on user feedback over a specified duration.
- In order to enhance the applicability of depression categories within the context of Bangladesh, we intend to conduct surveys and engage in discussions with a broader range of psychologists. Through these efforts, we aim to establish more comprehensive and applicable generalizations pertaining to depression.
- We will use transformer based architecture.

Chapter 8

Conclusion

To conclude, in our work, we discussed multi-class text classification, a well-known NLP topic. Analyzing the reasons for depression in Bangla social media content is a novel work in the field of Bangla NLP. Although we get some research works on depression analysis, which are binary classifications like a Bangla text, whether depressive or not, our work is analyzing a depressive Bangla text to find out the reason behind the depression. We have trained our dataset with five models, namely, Naive Bayes, SimpleRNN, LSTM-RNN,pre-trained BERT, and DistilBERT. We get the highest accuracy from the DistilBERT model. We also applied tuning to improve accuracy. However, we are still collecting data. We need more data points to improve our model accuracy, and then we will add our Islamic-based counseling system to our thesis work.

References

[1] A. U. Hassan, J. Hussain, M. Hussain, M. Sadiq, and S. Lee, “Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression,” in *2017 international conference on information and communication technology convergence (ICTC)*. IEEE, 2017, pp. 138–140.

[2] P. V. Narayanrao and P. L. S. Kumari, “Analysis of machine learning algorithms for predicting depression,” in *2020 international conference on computer science, engineering and applications (iccsea)*. IEEE, 2020, pp. 1–4.

[3] L. He and C. Cao, “Automated depression analysis using convolutional neural networks from speech,” *Journal of biomedical informatics*, vol. 83, pp. 103–111, 2018.

[4] F. T. Giuntini, M. T. Cazzolato, M. d. J. D. dos Reis, A. T. Campbell, A. J. Traina, and J. Ueyama, “A review on recognizing depression in social networks: challenges and opportunities,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 4713–4729, 2020.

[5] JabRef Development Team, *JabRef*, 2016. [Online]. Available: https://www.nimh.nih.gov/health/topics/depression

[6] *JabRef*, 2016. [Online]. Available: https://www.who.int/bangladesh/news/detail/28-02-2017-number-of-people-with-depression-increases

[7] *JabRef*, 2016. [Online]. Available: https://www.who.int/bangladesh/news/detail/28-02-2017-number-of-people-with-depression-increases

[8] A. H. Uddin, D. Bapery, and A. S. M. Arif, “Depression analysis of bangla social media data using gated recurrent neural network,” in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT).* IEEE, 2019, pp. 1–6.

[9] M. R. H. Khan, U. S. Afroz, A. K. M. Masum, S. Abujar, and S. A. Hossain, “Sentiment analysis from bengali depression dataset using machine learning,” in *2020 11th international conference on computing, communication and networking technologies (ICCCNT).* IEEE, 2020, pp. 1–5.

[10] A. H. Uddin, D. Bapery, and A. S. M. Arif, “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).* IEEE, 2019, pp. 1–4.

[11] D. B. Victor, J. Kawsher, M. S. Labib, and S. Latif, “Machine learning techniques for depression analysis on social media-case study on bengali community,” in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA).* IEEE, 2020, pp. 1118–1126.

[12] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, “Detecting depression and mental illness on social media: an integrative review,” *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017.

[13] R. Thwaites, D. Dagnan, D. Huey, and M. E. Addis, “The reasons for depression questionnaire (rfd): Uk standardization for clinical and non-clinical populations,” *Psychology and Psychotherapy: Theory, Research and Practice*, vol. 77, no. 3, pp. 363–374, 2004.

[14] B. Meyer and L. Garcia-Roberts, “Congruence between reasons for depression and motivations for specific interventions,” *Psychology and Psychotherapy: Theory, Research and Practice*, vol. 80, no. 4, pp. 525–542, 2007.

[15] [Online]. Available: https://www.analyticsvidhya.com/blog/2014/10/ann-work-simplified/

[16] [Online]. Available: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/

[17] [Online]. Available: https://huggingface.co/blog/bert-101

[18] [Online]. Available: https://www.infoworld.com/article/3278008/what-is-tensorflow-the-machine-learning-library-explained.html

[19] [Online]. Available: https://www.javatpoint.com/keras

Appendices

Appendix A

Title of Appendix A

Appendix A goes here.....

Appendix B

Title of Appendix B

Appendix B goes here.....