GREEN UNIVERSITY OF BANGLADESH

Machine Learning Techniques For Detecting Human Depression Using Social Media Text Data

Submitted by

Estiak Hasan Emon (201002059)
Shamima Aktar (201002265)
Dev Sarkar (201002026)

A thesis paper is submitted to the Department of Computer Science & Engineering for the partial fulfillment of the degree of Bachelor of Science in Computer Science & Engineering

Supervised by

Md. Noyan Ali
Lecturer, Department of CSE

Department of Computer Science & Engineering
Green University of Bangladesh
Purbachal American City, Kanchan, Rupganj (Narayanganj-1461, Dhaka, Bangladesh)
February, 2024

Declaration

The work that we, under the guidance of **Md. Noyan Ali**, Lecturer in the Department of Computer Science and Engineering at the Green University of Bangladesh, has presented in this research project, entitled **Machine Learning Techniques For Detecting Human Depression Using Social Media Text Data**, is legally declared. Research materials sourced from other investigators are referenced where appropriate. This work has never before been submitted, in whole or in part, for a degree competition.

Estiak Hasan Emon	Shamima Aktar	Dev Sarkar
ID: 201002059	ID: 201002265	ID: 201002026

Certificate

This is to certify that the thesis entitled **Machine Learning Techniques For Detecting Human Depression Using Social Media Text Data**, has been prepared and submitted by **Estiak Hasan Emon, Shamima Aktar & Dev Sarkar** in partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering in February 2024.

Md. Noyan Ali
Supervisor

Accepted and approved in partial fulfillment of the requirement for the degree Bachelor of Science in Computer Science and Engineering.

Dr. Muhammad Aminur Rahaman
Chairperson

Mr. Tamim Al Mahmud
Assistant Professor & PC (Eve)

Mr. Abdullah Al Farhad
Lecturer

Acknowledgments

Firstly, we would like to thank almighty God for his wonderful gift, which has enabled us to successfully complete the final year research. We sincerely thank and express our great gratitude to **Md. Noyan Ali**, Lecturer at the Green University of Bangladesh, Purbachal American City, Kanchan, Rupganj (Narayanganj-1461, Dhaka, Bangladesh), Department of Computer Science and Engineering. To complete this research, our supervisor has a deep understanding and a strong interest in the subject of machine learning. It has been made possible to finish this research by his unending patience, scholarly advice, persistent encouragement, energetic supervision, constructive criticism, insightful advice, reading numerous subpar versions and editing them at every stage. His foresight gave rise to the initial suggestion that the sensitivities of the entire research project be re-examined. Throughout the study and dissertation writing, he has supported us through some very trying times, and for that, we are truly grateful. We would like to extend our sincere gratitude to the CSE department chair, **Dr. Muhammad Aminur Rahaman**, as well as to the other faculty members and personnel of Green University of Bangladesh, for their gracious assistance in helping us to complete our research. We express our gratitude to everyone of our Green University of Bangladesh course participants who participated in this conversation during the course of the assignment. Lastly, we must respectfully thank our parents for their unwavering support and patience.

Abstract

Depression is one of the most important problems facing modern human society. It is among the modern era's well-known mental health problems. In this research, we examined the mental health condition known as depression. Numerous individuals of all racial, gender, and age groups have been impacted by this. Furthermore, some individuals find it forbidden to discuss, which adds to its gravity. But because of advancements in technology, we now live in a civilization that is referred to as the "age of modern communication." We can now communicate more effectively and share our thoughts especially our emotions thanks to the era of digital communication. Social media is now another way that people express their ideas about depression. It should be noted that many people do not even recognize they are depressed, despite the fact that their social media posts indicate they are. We decided to collect our data from multiple social media platform, such as: Facebook, Instagram, Twitter, You-tube comments, because of this kind of the social media platforms where individuals share their opinions. Because of the recent coronavirus, our society has seen a wide spectrum of challenges arise. For a considerable amount of time, we had to stay behind closed doors. These problems are leading to an increase in depression rates in the general population. The daily increase in the statistics suggests that more people are utilizing the internet to access social media. Since social media allows individuals to communicate their personal thoughts with one another through private messaging or publicly shared postings, it's a great medium for communication. Our study's goal is to use data from publicly accessible social media platforms to analyze information and use that knowledge to identify depression. Due to our enormous data requirements, we used tools to scrape data from social media and categorized it based on the emotional perspective of the user.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Modern this period about human history, “Depression” is among the most well-known terms. It would be difficult to locate someone who utilizes contemporary communication techniques who has never heard of depression. This topic gained increased attention recently, particularly during the COVID-19 pandemic. However, there is a potential question: what exactly is this depression? Depression is classified by the World Health Organization (WHO) as a mental disorder that is largely characterized by continuous melancholy. Other symptoms of depression include a general lack of excitement in something and decreased enjoyment from formerly relaxing pursuits. It is also possible to experience other symptoms, such as appetite loss or insomnia. Fatigue and difficulty concentrating might also be symptoms.

Our society has witnessed a wide range of issues emerge as a result of the recent Coronavirus. Numerous interruptions in our lives have been caused by the ongoing Corona pandemic lockdown. In addition, we were required to remain behind closed doors for an extended period of time. Depression rates are on the rise in the general population as a result of these issues. Anybody, regardless of age, gender, or circumstance, has been impacted by the coronavirus. A great deal of people experienced job loss, which left them alone and afraid of dying, among other things. Anxiety and sleeplessness were considerably elevated. According to one study, the pandemic increased drug usage by 36 percent and alcohol consumption by 18 percent. Depression increased for all the previously listed reasons. The data indicates that it increased by over 25 percent.

According to data, there are over 5.07 billion internet users worldwide, or roughly 72 percent of all persons with access to or use of the internet. Based on this, 4.74 billion individuals, or 59.3 percent of the global population, utilize social media. The fact that the numbers are increasing every day suggests that more and more people are utilizing online resources for using social platforms. Social platforms is an excellent method for communication since it allows people to post publicly or exchange private messages with each other about personal views. They exchange information like voice recordings, photos, videos, and much more. Additionally, this data indicates the user's sentiments, feelings, and moods.

As previously indicated, depression is classified by the WHO as a mental illness characterized by symptoms such as loss of interest in once-interesting activities and feeling progressively sadder than before. It leads to physical issues like decreased productivity at work and lethargy as well as mental health issues. Mild to severe symptoms are possible. Patients may also experience worthlessness or unjustified guilt.

1.1.1 Five Prevalent Forms of Depression

Deb Fulghum Bruce, PhD degree, claims that there are 5 prevalent forms of feeling depressed:

- **Manic-depressive psychosis:** Manic depression is another name for it. There are several mood episodes associated with this type of depression, ranging from high energy to low depressive times. When someone is at their lowest, they may exhibit the symptoms of major depression-a topic we shall cover later.

- **Severe mood disorders:** It should go without saying that one of the most well-known forms of depression is severe depression. Severe depressed individuals refer to it as "global gloom." Individuals experiencing major depression may exhibit symptoms such as loss of interest, suicidal thoughts, difficulty making decisions, difficulty concentrating, feelings of worthlessness and guilt, insufficient energy to perform daily tasks, difficulty sleeping or insomnia, agitation, and a languid or lethargic mental and physical state.

- **Chronic Depression:** The term dysthymia was once used to refer to persistent depression disorder. It is usually defined, and its title suggests, and continuous, ongoing symptoms that can last as long as two years. Individuals diagnosed with persistent depression can handle daily tasks with ease, but they hardly ever show signs of joy. It's possible to experience changes in a sense of self food intake, mental state, and quality of rest.

- **After giving birth Depression:** One out in every seven new mothers experiences this kind of depression. It can be challenging for new moms because it causes anxiety and tiredness in them. It makes completing their everyday tasks more difficult.

- **Seasonal Affective Disorder:** This title also speaks for itself because the majority of cases of this kind of disease occur in the winter. Less sunlight and shorter days became the norm. In the spring or summer, its impact disappears.

1.1.2 Covid Depression Spike

The fast spread of the corona virus during the COVID-19 pandemic has caused significant worry and anxiety worldwide. Numerous decisions were made to deal with the epidemic and deaths, one of which was the lockdown, which required individuals to remain inside their homes for months at a time without ever exiting. Because of this, a large number of the public possess experienced unhappiness beyond surely being aware of it.

Signs and manifestations of depression:

- A sense of annoyance.

- Sleep disturbed often.

- Behavioral changes, such as a decrease in appetite.
- Depression can also manifest as physical aches, pains, stomach discomfort, and other kinds of ailments.
- Having trouble staying motionless.
- Difficulty focusing on tasks or other activities.
- Depression can also cause weight to fluctuate.

Machine Learning: We are aware that machine learning is a unique type of data analytics automation. It is sometimes referred to as an artificial intelligence subset. It makes decisions automatically, without the need for human intervention, by using the data that is given to it to identify patterns and build its own.

1.2 Motivation

The World Health Organization (WHO) website states that there are approximately 280 million depressed individuals worldwide. This indicates that 3.8 percent of people worldwide suffer from depression. Additionally, depression is the cause of 700,000 suicides. Furthermore, these are the figures that have been disclosed; yet, many cases remain unreported.

Social media is one of the many helpful additions to our list that the twenty-first century has brought about. The majority of people on the planet utilize social media. Social media is a fantastic platform for people to communicate and exchange ideas. People enjoy using social media to express their emotions online. They also talk about their stories and melancholy. where hints on their sadness may be found.

Our objective is to use this social media data to build a model that will enable us to automatically identify depression from text data using machine learning. As many people are unaware that they are depressed, depression is still stigmatized in society.

Therefore, they are unable to even seek depression treatment. However, if social media platforms are able to identify their users' mental health, they may be able to assist them without making them feel uncomfortable.

We select multiple social media platform, such as: Facebook, Twitter, Instagram, You tube comments, from among all the social media platforms to compile our data. Because a diverse spectrum of users use it, and most of them express their feelings with the public through a narrow path. Additionally, it is simple to gather a vast amount of information about the efforts from here. Certainly so, ourselves are aware that a particular person feeling particularly depressed is an issue for the majority of people. We continued collecting data based on how simple it was for us to sort it. In order for it to function using simple emotion detection.

1.3 Objectives

- **To develop a robust dataset:**

First, we create a reliable dataset from text data from social media platforms to identify depression in people. We gather information from a variety of social media sites, including YouTube comments, Facebook, Instagram, and Twitter. To minimize model bias and enhance generalization, the dataset should include a balanced proportion of postings with both depressive and non-depressive language. Maintaining the relevance and efficacy of the dataset in accurately detecting depression also requires regular updates and refining based on feedback loops and continuing research discoveries.

- **To retrieve pertinent attributes:**

This procedure entails taking the text data from social media platforms and extracting several linguistic properties like word frequency, sentiment polarity, and grammatical structures. Furthermore, metadata elements that provide useful insights into individuals' online behavior and emotional states include posting frequency, time of day, and user engagement metrics.

- **To assess the functionality of the model:**

An evaluation method needs to be carried out in order to determine how well the model works for identifying depression from text data from social media platforms. To improve accuracy, we employ a variety of models, including term frequency–inverse document frequency, bag-of-words, SVC, AdaBoost, forest of decision trees, multinomial logit, and XGB classification algorithm. It is possible to compute a number of assessment measures, including accuracy, precision, and recall, to determine how well the model performs in accurately detecting depressive episodes.

- **To evaluate and assess literary material for indications of depression:**

Analyzing the text's many elements, such as language use, ideas, and character development is crucial. First, a quick assessment of the work's general tone and mood can provide early indications of the existence of melancholy features. Second, recurrent concepts like sadness, loneliness, and observant questioning might point to early stages of depression. Thirdly, a fuller comprehension of lonely narratives can be obtained by examining the representation of exponent and their internal conflict.

- **To creating a predictive model capable of correctly identifying whether textual input is suggestive of depression or not:**

It is crucial to have a different and easy to understand dataset including examples of both depressed and non-depressive material. Taxation, stemming, and stop-word removal are examples of processing techniques that can assist standardize the textual input for analysis. Pattern detection within the text data can then be aided by interesting machine learning algorithms like logistic regression, support vector machines, or deep learning frameworks like recurrent neural networks.

1.4 Rationale of The Study

Recently, there has been a lot of study using AI and ML to automate processes across a variety of fields. It is also heavily utilized in research on text type and emotion recognition. This is the direction taken by our work, where we use ML to detect depression in social media data. Numerous studies have been conducted in this topic in the past. Different approaches were taken while utilizing the same concepts. Various works produced various results based on different data. Our work focuses on utilizing newly gathered information via online communities, combining it added to older information, also building the personal extraction.

When depression sets in, a person may experience difficulty focusing and making judgements; it also impairs their agility. Depression has an impact on the brain as well, which may lead to problems remembering things. Not only may depression impact a person's physical structures, but it can also impact their mentality. It's also possible that depression could have an impact on a person's central nervous system. It may also result in long-term brain damage, which is why many depressed people experience problems remembering things. Notable is the fact that about 20 percent of those who experience this may never fully recover.

1.2 billion individuals, or 16 percent of the world's population, are considered youthful, defined as those between the ages of 15 and 24. In addition, the majority of depressed individuals are between the ages of 18 and 29. Additionally, the age group of adults between 45 and 65 is the one most severely affected. Notably, ninety percent of adults between the ages of eighteen and twenty-nine utilize social media.

Because of this, it is more probable that we will be working with data from the younger generation if we use the social media platforms that are available today. This means that if we are successful in our work, we will be able to collect more data on the diagnosis of depression in young people. This is because young people make up the majority of social media users, and they are more likely than older people to share their emotions online.

1.5 Research Question

There are certain questions that must be answered before we can begin our investigation. Our work will be organized according to these questions. The questions we came across are as follows:

- Why is it necessary to recognize depression?
- Why is it necessary for us to collect data through social media?
- How will we get the information we need?
- What kinds of data collection techniques should be applied?
- Why is machine learning going to be used?
- Which type of machine learning shall we employ?
- Which standards will we use to gather data?

Before we get to the outcomes, we will talk about a few quantitative findings from earlier research. We searched through social media posts pertaining to our work to determine keywords for it. More signs of frustration became apparent as a result of its improved performance in our ability to recognise depression. We made an effort to include emotions like dissatisfaction, fury, happiness, and anger in the word for despair.

Therefore, this research will provide insight into how to swiftly and effectively diagnose depressed symptoms in written texts. In the past, depression has been linked to changes in language use. We have also read a variety of articles in an effort to find excellent methods that are more accurate than the others at detecting depression.

For our research, there are two primary data sources. One source is a dataset that is accessible online and includes information from multiple social media platforms. It contains over 36,000 depression-related data points. It contains three different types of data: neutral data, not depressed data, and sad data. Using technologies, we have gathered data from multiple social media platform, such as: Facebook, Instagram, Twitter. The period covered by this data is July 2023-January 2024. The data that we manually labeled from Facebook was gathered. We additionally labeled our data based on our web dataset according to several parameters.

1.6 Gantt Chart of The Research

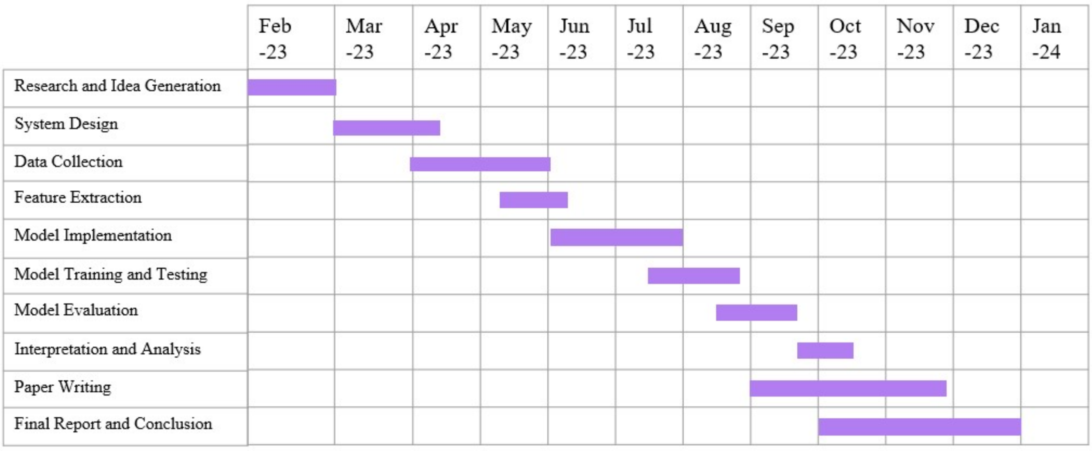

Figure 1.1: Gantt chart for the research current progress and the remaining time plan

We can see from the above graph that it took us till February 23, 2023, to conduct our research and generate our ideas. System design took until March 23, 2023, to complete. After that, we began gathering data, which lasted from April 23, 2023, to September 25, 2023. We performed some feature extraction following data collection, which took place between May 30, 2023, and June 27, 2023. Then, between June 25, 2023, and July 30, 2023, we employed five distinct models to improve accuracy. Then we began the process of training and testing our models, which lasted from July 28, 2023, until August 29, 2023. And from August 30, 2023, to September 29, 2023, we evaluated the model. Our time for the interpretation and analysis phase was from September 27, 2023, until October 27, 2023. After that, we got to work composing the paper, which took us from September 23, 2023, until December 30, 2023. Finally, it took us from

October 23, 2023, until December 30, 2023, to complete the final report and conclusion.

1.7 Budget of The Research

We have our budget table below. Here we have broken down our budget into fifteen subcategories in addition to seven primary areas. We'll go into more detail about our budget now. First off, we have a fixed budget of 60,000 tk for our computer resource, which covers the cost of the hardware, software, and licences. After that, there is a fixed price of 12,000 tk for data access and collection, which covers both data collecting and data access. A budget of 22,000 tk has been allocated for personal expenses such as hiring data analysts and researchers in the third category. Next, we decided on a 10,000 tk budget for outside services, which includes transcribing and ethics approval. Then, we have set aside 18,000 tk for travel and conferences, which covers lodging, meals, and conference registration. Next, we decided that 20,000 tk would cover publication fees, printing, and distribution as part of our budget for publication and distribution. And lastly we have budgeted 8,000 tk for miscellaneous expense, which includes communication and internet, documentation and training. And our total budget is 1,50,000 tk.

Table 1.1: Annual funding allocation for our study research

Category	Estimated Cost (BDT)
Computing Resources (Hardware, Software and License)	60,000
Data Collection and Access (Data Acquisition, Data Access Fees)	12,000
Personal Expenses (Researcher's, Data Analyst)	22,000
External Services (Ethics Approval, Transcription Services)	10,000
Travel and Conferences (Conference Reg, Travel and Accommodation, Meals)	18,000
Publication and Dissemination (Publication Fees, Printing and Dissemination)	20,000
Miscellaneous Expenses (Communication and Internet, Documentation and Training)	8,000
Total	1,50,000

1.8 Report Layout

The introduction to our study effort and a brief summary of our thesis are contained in Chapter 1. We talk about the relationship between social media and its users and how it greatly influences the data collection for our research. This chapter also discusses our motivation. We have also discussed the basic actions and efforts we are doing to complete this research. We have also briefly discussed several possible outcomes.

Our work's background study is covered in Chapter 2. In this chapter, we discussed the idea that motivated our effort and the reasons behind it. Additionally, we have made an effort to present a fundamental synopsis in worldwide online communities examination using data that got acquired online. We have also talked about the difficulties we still need to overcome.

The primary goal of the experimental analysis section of Chapter 3 is to outline and establish our study plan. We have talked about this, which is our key point. Because the type of data we use is crucial to our job and holds the most significance, that is our first concern. Because the kind of data we select will determine our result and outcome. In this chapter, it receives the most attention. We have discussed the features of our two datasets and their respective topics. We discussed and looked at the dataset.

We have also discussed the studies we conducted throughout our experiment in this particular chapter, which is the experimental outcome and discussion in Chapter 4. We talked about the algorithms we employed and the mathematical methods we employed for this study. We have discussed the ideas, results of the experiment, and explanations of them.

A summary of our entire study research, a conclusion about our current work, recommendations, and potential future applications are included in Chapter 5, along with an idea of potential consequences and areas for implementation.

1.9 Conclusion

Depression detection is a technique that can be used to text, image, or video data. We employed social media text for depression identification for our specific task. The purpose of our study is to analyze data from social media platforms that are accessible to the general public in order to identify depression from the information acquired. Our goal is to identify depression using various machine learning techniques as a reliable and effective approach.

Chapter 2

Literature Review

2.1 Introduction

The most recent pandemic brought about a lot of changes in our lives. Most individuals felt the effects of these developments everywhere in the world. For most individuals in the world, it has made them more tense and anxious.

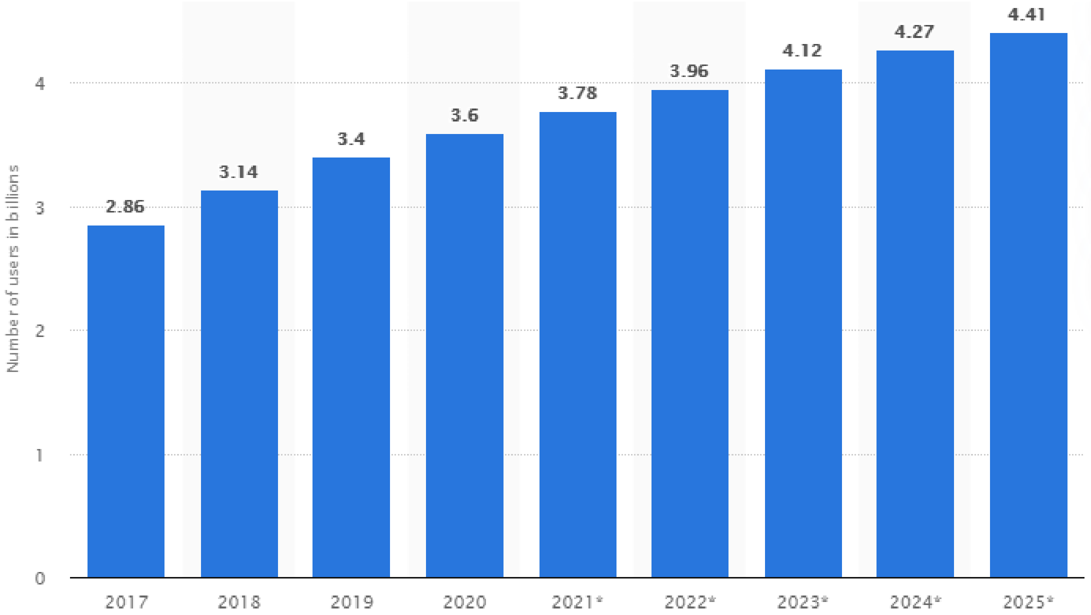

Figure 2.1: Growth in social platform users between 2017 to 2025

The fact that so many individuals had to stay at home due to the pandemic is one of the most significant elements; as a result, many of them developed sadness. Thus, it is imperative to keep an eye on public health for this reason. The shutdown has resulted in numerous modifications. Due to a lack of activities during the lockdown, a lot of people started using social media. Throughout the past two years, there has been a noticeable increase in the number of people using social media. During this time, the number of users on social media sites like Facebook, Twitter, Tiktok, Reddit, and so on increased significantly.

This study concentrated on individuals who express their emotions on social media since, among all the emotions, there are feelings of melancholy, loneliness, and other signs that indicate the person may have been affected by depression.

Using the data that regular internet users have contributed to the public domain is the aim of this study. Based on the data that is stored on the user's post, each input serves as data for our study. Since most data on the internet is text-based, there are many different sorts of data available, information such as preference, gender, age, and many more things. It tells us a great deal about an individual. Nonetheless, posts or text data can reveal an individual's emotional state. We intend to gather and apply this data that we find on social media. The figure is taken from statista.com site.

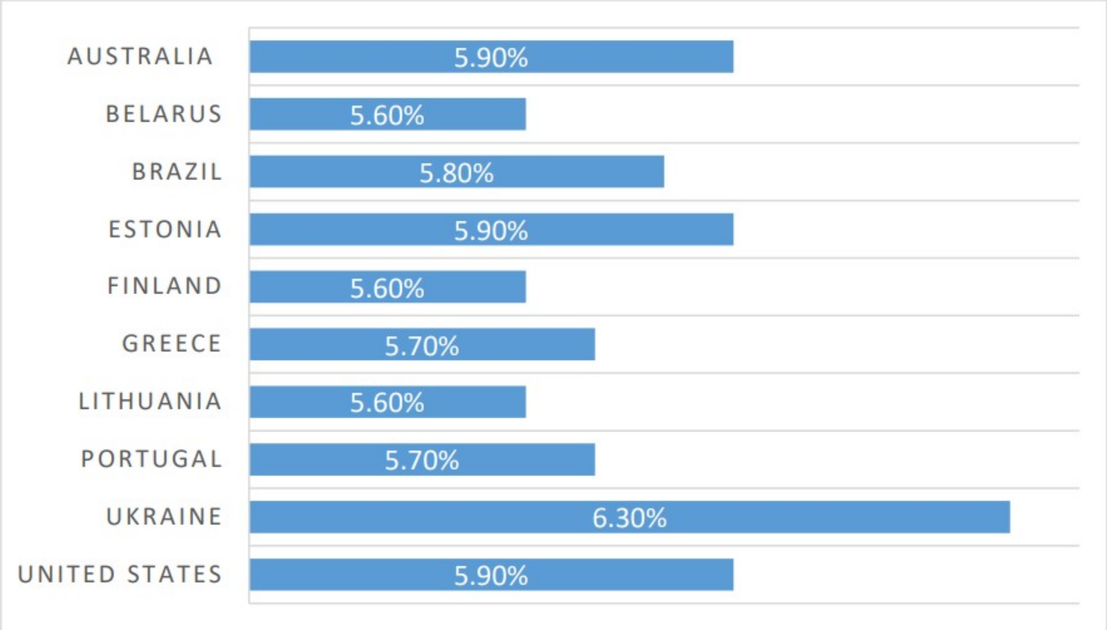

Figure 2.2: The top ten nations with the highest depression rates

If this volume of data had to be expressed numerically, 527,800 pictures were switched upon converse per minute. Over four million clips were viewed on YouTube, and 456,000 posts were sent on Twitter. On Facebook alone, almost 16 million SMS are written, about 500,000 comments are made, and 293,000 statues are updated.

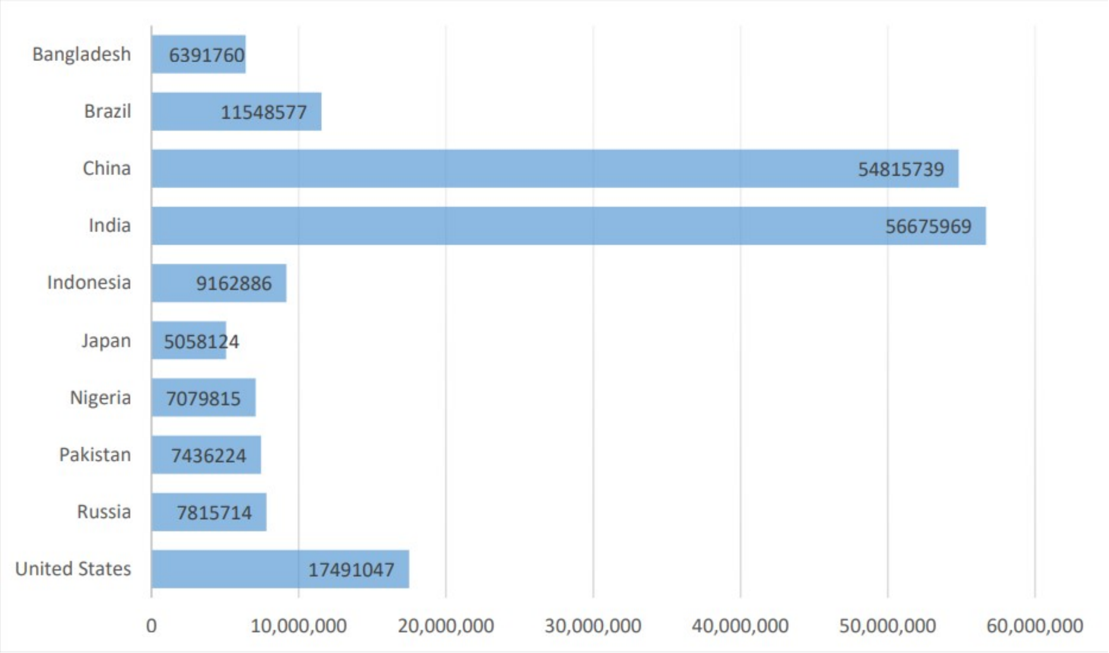

Figure 2.3: The top ten nations with the greatest rate of depression

2.2 Related Works

The purpose of this part is to summaries the current level of knowledge about systems, awareness, and a few work steps. There are numerous study papers discussed that deal with identifying depression in people.

In [1], the authors assert that people who fit a particular profile in terms of personality or demographics are more likely to post details regarding the personal psychological well-being assessed upon online platforms. Our study's findings indicate that the MLP classifier performed the best in identifying and comprehending the existence of sorrow on the social media platform Reddit. It had a 91 percent accuracy rate. The machine learning system for Multilateral limited partnerships demonstrated a potency also efficacy about an merged characteristics by achieving a score of 91 percent.

In [2], this article outlines a process for getting usernames from social media users who post in order to assess a person's risk of depression. Fifty individuals received an invitation on Facebook requesting them to submit their latest postings. These user-generated posts were subsequently integrated to the artificial intelligence system. That was demonstrated to sadness may and may not result in severe mental illness or even suicide, and that machine learning techniques can be used to identify depression. Furthermore, it has been demonstrated that depression can be identified using machine learning techniques.

In [3], the recommended framework has capable of identifying when and whether this consumer has melancholy by using machine learning techniques. Based on the words that the user submitted, the algorithm reads the emotion contained in the text before deciding whether or not there is any indication of depression. The system's ability to be accessed in the privacy and comfort of one's own home protects the user from the social stigma that permeates his or her surroundings.

In [4], the suggested approach's feature set diversity and richness have enabled it to reach a greater accuracy than the prior method suggestions. Words with greater frequencies are selected based on the user's perception of what they ought to be. Using the feeling of the sentences approach, the sentiments expressed in each tweet are ascertained. The general feeling across everyone's messages seems then ascertained to be known as consumers combined phrases.

In [5], major catastrophes like illnesses prompted for a covid virus 2019 (Covid-19) would cause people's mental health problems to escalate even more, instead of just due to about this crisis actually however additionally due to regarding a societal issues which subsequently, such as unemployment, to inadequate finances, additionally the economic downturn.

In [6], their analyses provide insight into the sentiment of society over time and problematic themes covered in Weibo posts. Therefore, it could be beneficial to do study on unfavorable social media posts within arrange in acquire the superior knowledge about what this Chinese people went through during the COVID-19 outbreak and to set an example for other countries. The conclusions drawn from the Weibo posts provided helpful guidance on public health, and it's feasible that openness and evidence-based recommendations will allay public concerns.

In [7], by utilizing a range of techniques, machine learning may identify depressive states in individuals by listening to or watching recordings of sounds. Classifying and diagnosing a wide range of disorders, including neuro degenerative diseases, it has been applied to medical diagnostics. It is possible to diagnose depression by extracting facial feature data from images and videos and analyzing the results using artificial intelligence techniques.

In [8], specifically, user-generated tweets from Twitter are utilized as the main source of data for analytical purposes in this article. The quantity of space needed for text, audio, and video content is far less than that needed for storage. Using emotional artificial intelligence to identify depression has been most successful on Twitter. The reason behind this is that Twitter sets the maximum character count that can be used in a tweet.

In [9], to investigate the topic of mental health, the authors created the EmoCT dataset, which they used to classify tweets about COVID-19 into distinct feelings. On April 7, 2010, they performed the individual specific grouping task upon an a million randomly chosen specified posts information, utilizing a Bidirectional Encoder Representations from Transformers emulate, that proved developed to forecasting a sentiment indicating about simply a single name.

In [10], the scientific method to as certain people's level of frustration with the messages they get via social media. Because the system extracts data from tweets using keywords instead of from Facebook postings, it is unable to ascertain an individual's level of displeasure. A six-point rating system is available for use with the depression criterion according to the machine learning model.

In [11], on the review of image and video-based depression detection using machine learning. That is the technique of data analysis that encompasses a computer to learn to classify or predict the given input to produce a smart decision or result as output.

In [12], mental Health Analysis on Tweets Using Natural Language Processing.

Outbreak of coronavirus disease 2019 (COVID-19) recently has affected human life to a great extent. Besides direct physical and economic threats.

In [13], bio politics, the Current Pandemic and A Sociological Approach to Global Disorder. World in deep crisis and the end of the current global disorder is unforeseeable. Life is the most common utterance when people talk about the pandemic, whereas any explanation of its global-political effects by referring to deaths.

In [14], text-based Depression Detection on Social Media Posts: A Systematic Literature Review, Psychiatrists found difficulties in identifying the existence of mental illness in a patient because of the complicated nature of each mental disorder, thus making it hard to give the appropriate treatment to the patient before it's too late. It using text-based approach, most studies use deep learning models such as RNN on the early detection of depression cases.

In [15], bridging Emotion Role Labeling and Appraisal-based Approaches, Emotion analysis in text subsumes various natural language processing tasks which have in common the goal to enable computers to understand emotions. Most popular is emotion classification in which one or multiple emotions are assigned to a predefined textual unit.

In [16], machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features. Social media channels, such as Facebook, Twitter, and Instagram, have altered our world forever. People are now increasingly connected than ever and reveal a sort of digital persona. Although social media certainly has several remarkable features, the demerits are undeniable as well. Recent studies have indicated a correlation between high usage of social media sites and increased depression.

2.3 Research Summary

The effort of that investigation concentrates to various community-accessible methods. A total of seven distinct algorithms, each with its own unique approach, have been used using our dataset. Facebook has been our main source of data in this case. As we previously indicated, our dataset includes both recently acquired personal data from us and data that was previously used. It will enable us to check for things like the effect of freshly added data from the same source and obtain the correctness of the seven algorithms we employed. The new data is the same as the old dataset that we combined with. This indicates that the labels are of the same type and class. Our primary language of choice was Python, and the feature extraction processes we employed made use of machine learning approaches.

2.4 Scope of The Problem

Gaining a thorough understanding of the problem's scope is essential to finding effective solutions:
Influence analytics platforms must be able to identify melancholy in textual content in order to be used in a wide range of scenarios. Those who are kept inside during the lockdown may end up experiencing some level about psychological happiness issues to be the outcome. In order to instance, young people possibly experience anxiety about their professions or careers, business-people may worry about their enterprises, and students may experience depression or unhappiness related to their course work.

This article concentrates on the section where a text will be the input and a suitable output that can identify a suitable emotional state will result. One will be able to determine whether or not the person whose data was input is depressed with the use of this suggested technique. Because the system has been trained by numerous user inputs, it may customize recommendations and provide a suitable response that tackles the primary issue of the subject.

One specific goal of our research is to develop a notion. A notion that can accommodate the needs of everyone who is experiencing or may experience depression. a theory that uses a computationally efficient text-based paradigm. An excellent objective whereby this effort can benefit everyone.

Our primary goal is to identify depression in individuals before it progresses to an incurable state or to the point where making poor decisions may become necessary. If the idea is successful, many individuals' bad decisions, like suicide, can be avoided.

2.5 Research Challenges

Textual data is the most commonly utilized type of communication, thus using it to analyze data to detect sadness is tremendously helped.
We encountered the following difficulties when doing this research:

- Proper data collection (Appropriate information gathering)
- Selection of algorithms
- Searching for sources of data collecting
- Verification of data

In this study, we try to offer a framework that may assess users' emotional distress levels based on the information they share on social media. Because of the intricacy of language, depression detection is one of those tasks that has a reputation for being challenging, much like text summarization and machine translation. It can be highly complex and difficult to identify indicators about sadness within conversation produced from different phrases. Supervised machine learning seems not a widely used or embraced technique as there has difficult regarding get sufficient labelled instructions information. This is called difficulty in diagnosing a depressed illness utilizing several forms of online social media, which was also fairly high during the COVID-19 crisis.

2.6 Conclusion

For a long time, depression was considered a singular illness with a set of diagnostic standards, similar to other prevalent mental health disorders. It affects how affected people feel and behave and frequently co-occurs with anxiety or other psychological and physical illnesses. Depression is under-diagnosed and untreated in many nations, which can have major consequences for one's self-perception and, in the worst case scenario, lead to suicide.

Chapter 3

Proposed Methodology

3.1 Introduction

Machine Learning technique approaches to diagnosing depression by looking at user-generated content or online posts. We have taken into consideration Facebook as a social media platform for our specific job. Because we are employing the English language, which broadens the scope of our data gathering, our study includes a wide spectrum of individuals. This will enable us to cover a range of age groups. Then there are members of various categories, such as regular people, students, or people in other professions. Two classes are included in our work model: one is sad and the other is not.

The fundamental paradigm that we have employed for this study endeavor is depicted in Figure 3.1. Searching the internet for data is the first phase, followed by data extraction from the source. Next, the unlabeled data will be obtained. But for our task, labeled data is necessary. Thus, we must classify the information that we have gathered. After that, the machine can be fed with this data. Then, we may apply our preferred ML techniques to that information in training besides testing data. Then, as the objective here, this algorithm will provide us with precision and enable us to identify depression. The raw data must be processed, also known as data preparation, in order for machine learning to be applied. For the ML model, it is a critical step and an important procedure. In the process, removing unnecessary data is also essential.

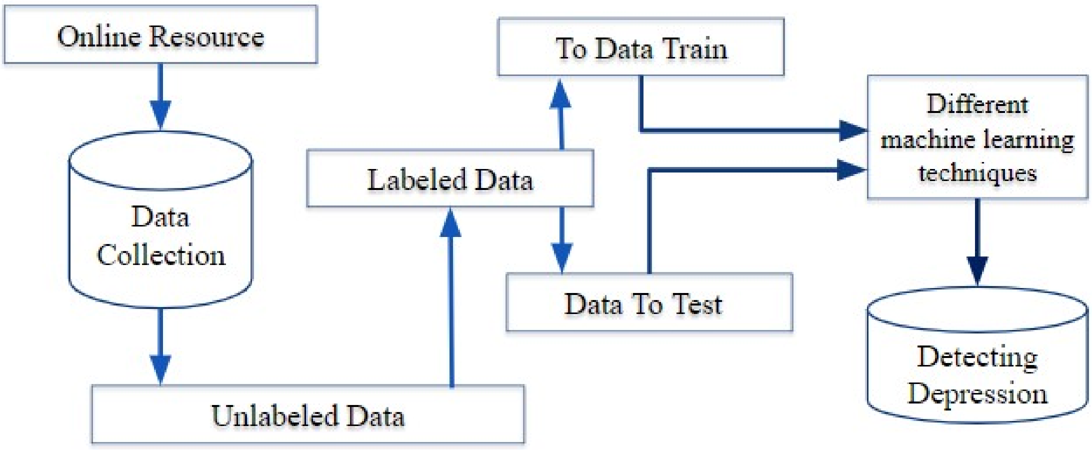

Figure 3.1: Recommended methodology for our exploration

3.2 Data Collection Procedure

The procedure, which is basically self-explanatory, is gathering data. Even so, there exist numerous methods for collecting data. When collecting data, there are a lot of requirements that need to be fulfilled. Factors such as the sort of data, the time it was generated, if it satisfies our work expectations, and so forth.

The internet has made a vast array of data readily available for usage, in an easily navigable format. There are tools available that make gathering data simple. The Face Pager Tool is the other from a instruments. This was the instrument it applied to get our data.

FacePager: Designed in 2019, Jakob Junger and Till Keyling created this autonomous data retrieval tool.

Data can be retrieved via the variation of multiple facebook and other social sites, including Facebook and others. It can be configured to pull information from several sources. One can adjust the parameters to suit their needs.

Various parameters able to employed regarding gather information to the research; people have utilized elements such as:

- Messages
- Label

In addition, we gathered information such as: whereas our data collection was contingent upon the parameters.

Categories of information	Number
Total amount of information	45000
Positive or non-depressing facts	25000
Negative or depressing facts	20000

Table 3.1: Distribution of total, positive, and negative data

3.2.1 Labelling Encoder

Within the domain of machine learning, datasets alongside multiple labels in one or more columns are frequently worked with. These designations can take the form of words or numerals. For the purpose of making the training data more comprehensible or human-readable, the data are frequently labeled with words. Label encoding is the procedure of translating labels into a language that a computer can comprehend. To begin this process, the labels must be converted to a numerical format. In the end, ML algorithms might decide how such labels should be used after doing some research. This stage of pre-processing the structured dataset is crucial for supervised learning.

3.2.2 Multiple Proposed Model

- **XGB Classifier:** The process of developing software that can learn without instruction is known as machine learning. This area of artificial intelligence creates and develops computer programs using statistical and mathematical techniques. Many machine learning algorithms are based on a type of data categorization system called neural networks. This XGB neural network categorization is the industry standard in machine learning. The purpose of that is recognize speak language the recorded sound. To identify speech, an XGBoost classification employs normal 2 steps: recognition and training. In order to build a model, the classifier must be fed the audio source together with a collection of target words during training. Next, the model is applied to the audio recordings to detect unfamiliar terms. Following training, the classifier may be applied to voice recognition tasks; in noisy circumstances, it can precisely and accurately identify words. Numerous applications, including voice-activated cars and dictation services, employ XGB for automated speech recognition. It has also been implemented by some gaming platforms for automatic text-to-speech features. In addition, an XGBoost voice felicitation classification was employed to uses in medicine to analyze heartbeats, breathing noises, additionally other data. Additionally, XGB is used by military organizations for automatic voice translation; only high-priority translations can be translated by humans.

- **Random Forest:** Use the tree-based method of the RF Classifier for both regression and classification. It's a method used in machine learning to build a hierarchy based on trees. An artificial intelligence method creates a hierarchy of "decision trees." The Random Forest Classifier creates many decision trees as an ensemble approach, then averages them. In this way, the over-fitting problem is mitigated. Machine learning is currently one of the most talked-about topics in business since that could being used in some scenario anywhere where that a lot of information's. An incredibly popular ML technique with several advantages over other algorithms is the RF Classifier. The technique was developed to handle big datasets and was originally published in 1997. A period that here were classes information initiates the labeling to every potential group additionally that minimum two examples per set, it can be used to solve any classification-related problem. Because it requires training data sets provided by a supervisor, the Random Forest Classifier is a member of the supervised classifier family. When there are sufficient datasets, this method proceeds to pick, split, and estimate out-of-bag using those datasets as the basis for its predictions.

- **Logistic Regression:** One type of categorization technique is called logistic regression. It chooses between two options by considering a number of factors that

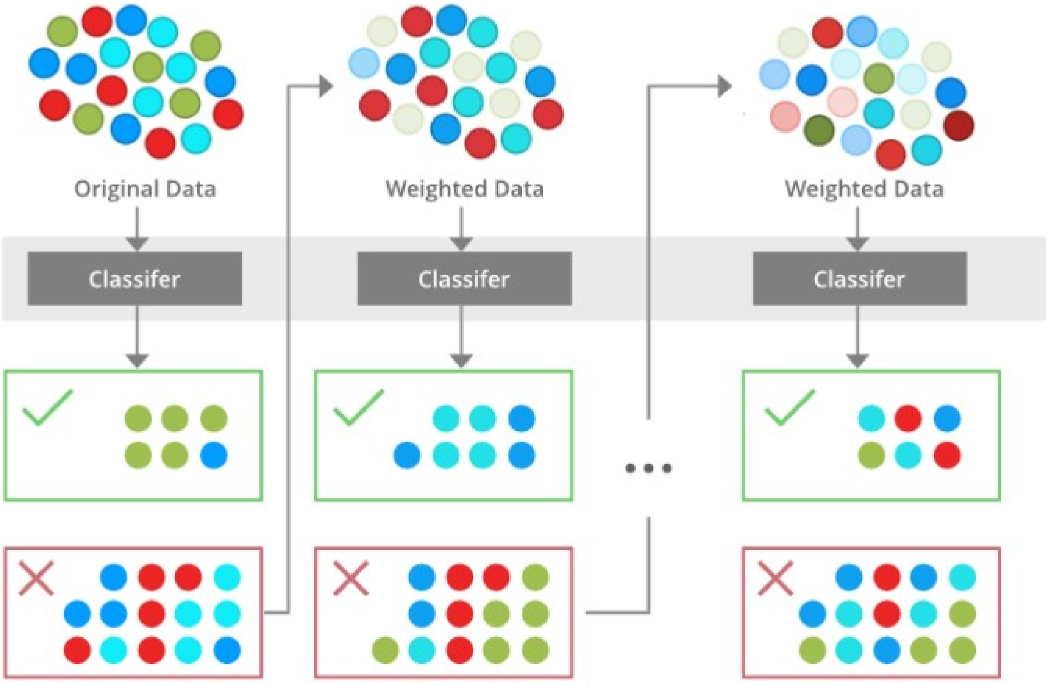

Figure 3.2: Workflow of Random Forest algorithm

are considered individually. What does this mean, then, in that scenario? A binary result can be defined as one in which there are only two possible outcomes: it will either occur (1) as well as that don't occur during every individual (0). Towards placed that only, everything other than the dependent variable that could have an impact on the study's ultimate result is considered an independent variable (or dependent variable). Therefore, a logistic regression is the suitable analysis method if you are working with binary data. You are working with binary data if the result or deciding factor is binary or categorical; to put it another way, if the data fits into one of two categories, you are working with binary data.

- **SVC:** Support Vector Classifier that the type of multivariate grouping technique, by definition, doesn't assume any information about the underlying structure of the data, such as that quantity about groupings and its corresponding dimensions. Previous studies have demonstrated that it functions best with low-dimensional data; hence, if your data have a high dimensionality, you will typically need to

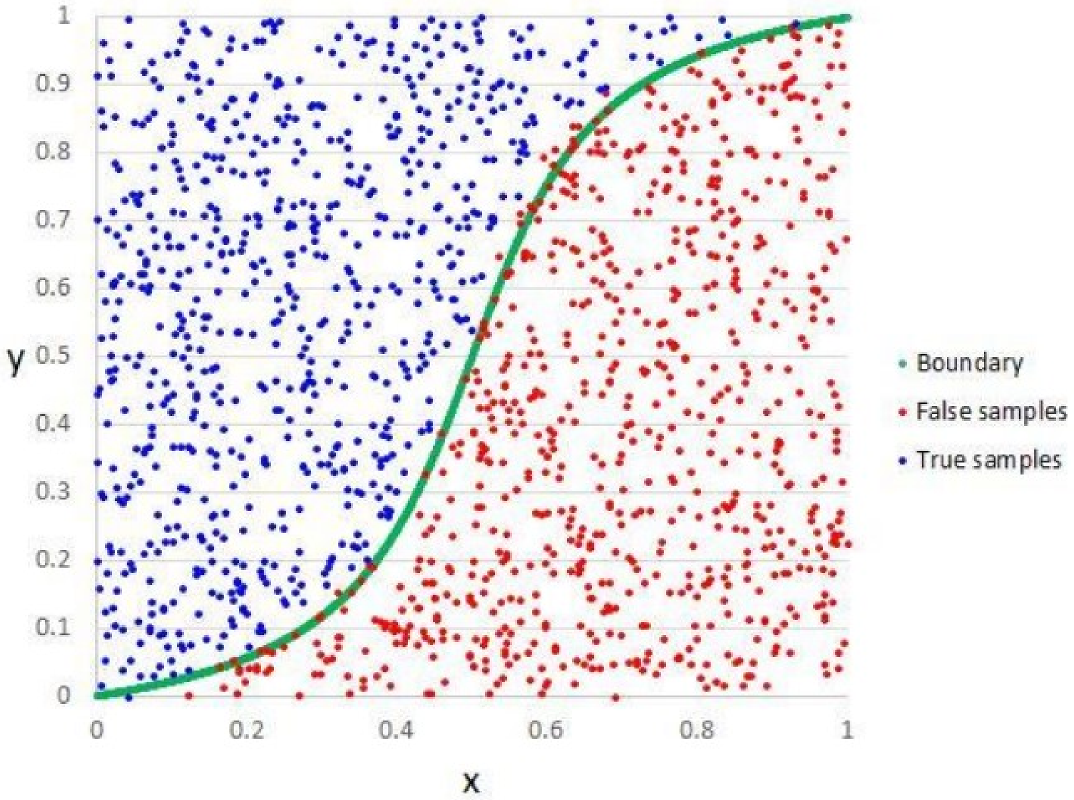

Figure 3.3: Example of Logistic Regression model visualization

do a preprocessing phase, such as principal component analysis. A few enhancements have been made to the original method, which yields specialized methods to calculating this groupings through figuring out just the portion to this adjacency matrix's sides. Numerous alterations have been suggested.

- **AdaBoost:** We are increase that effectiveness of our current ML method alongside AdaBoost. It works best when used with kids that are less capable. When these models are used for a classification task, their performance is just marginally better than that of chance. Since they are the most appropriate, single-level making choices structures have been via long way and a majority widely utilized kind on method when combined alongside Adaptive Boosting. "Boosting" combines and results of numerous weaker classifiers in an effort to create a robust classifier. In order to accomplish this, a model is built by connecting multiple more basic models. First, a model can be constructed by feeding in the training data.

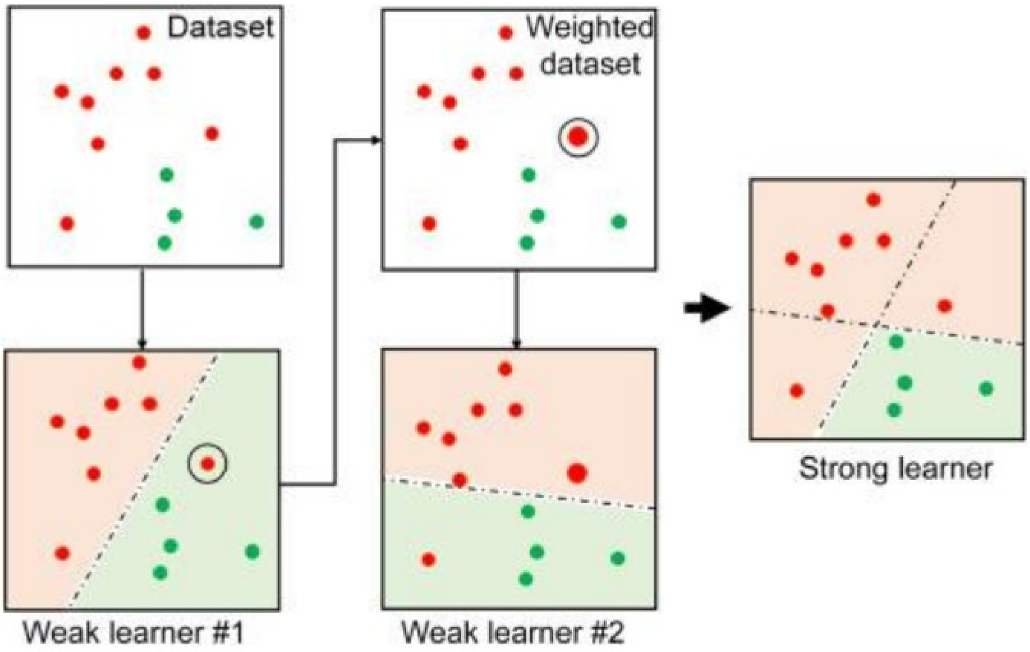

Figure 3.4: Workflow illustration of AdaBoost algorithm

- **TF-IDF:** To quantify a word's importance in relation to a collection of texts, statisticians employ that contrary prevalence record, (sometimes called term frequency-inverse document frequency). To obtain this conclusion, the term frequency-inverse document frequency that a term beyond a group and records is multiplied by two measurements: the first measures that regularity alongside that the phrase appears to the article. It is particularly helpful for word evaluation in NLP-related machine learning algorithms.

  TF and IDF are calculated with the following formulas:

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d} \tag{1}$$

$$IDF(t) = log\,\frac{N}{1+df}$$

$$TF-IDF(t,d) = TF(t,d)*IDF(t) \tag{2}$$

It may be used for a wide range of purposes, chief among them being computerized text analysis. They developed TF-IDF to do this task of searching for papers and retrieving data. This is accomplished by decreasing inversely proportional to the number of articles containing the term, while growing conversely to be an prevalence alongside that the phrase happens to this original document. Thus, even though they are frequently used in the work, common phrases like this, that, if, and what receive low marks because they are not significant there. On the other hand, if the term “Bug” appears more than once in one publication but not in another, it’s probably because the material is highly significant. For example, since most replies containing the word “Bug” will be concerning the category “Reliability,” we may assume that the word is most closely related with that category that the objective has identify this groups from that specific Net Promoter Score responses abide.

- **BOW:** Bag of words (BOW) is a text modeling technique used in natural language processing (NLP). If we want to talk in more technical terms, we could call it a feature extraction approach employing text data. This method makes it possible to extract features from papers in an easy and adaptable way. A “bag of words” is a text visualization that displays the frequency with which particular keywords occur in a given manuscript. An example would be: We don’t focus on the grammar or word order; we just count the words. Since the structure or order of the text’s contents are not preserved, this word “bag of words” applied to describe the obtaining to phrases. The sequence in which words appear in the text is not taken into consideration by the model; instead, it is primarily concerned with the existence or absence of specific phrases. Why therefore use such a random selection that phrases? Which specifically to incorrect alongside the straightforward, understandable fragment on composing? One of the biggest challenges when working with text is its messiness and lack of structure. This is because ML methods perform finest while given set up, mended distance ingredients which is precisely-defined. Additionally, the BOW method helps ourselves change documents and different sizes onto the graph regarding perpetual dimensions. Furthermore, since numerical data can be parsed into more precise details than textual data, machine learning models concentrate on it. To be more exact, we convert a phrase into a vector of integers using the bag-of-words, or BOW, method.

3.3 Statistical Analysis of Dataset

3.3.1 Data Analysis

Data analysis is done to extract significant insights, develop reliable conclusions, and offer helpful advice for decision-making. Numerous techniques with a wide range of names are used in data analysis, which is complex and multi-method and found in many business, scientific, and social science domains. Data analysis has the potential to improve operational efficiency and lead to more data-driven decisions, which would be advantageous for the modern business world.

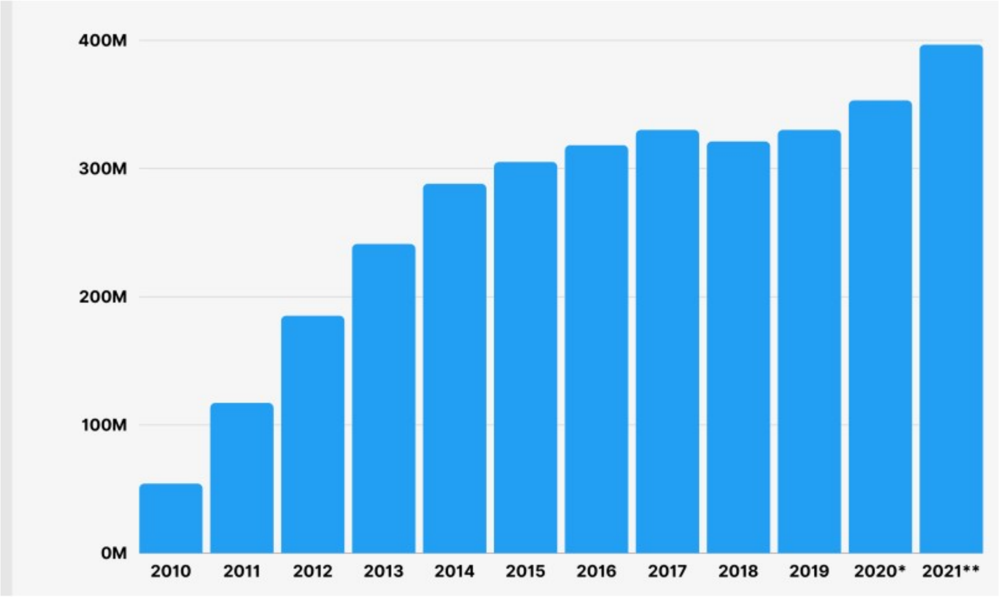

Figure 3.5: Global Twitter user count (per million) from 2010 to 2021

In the figure 3.5 we can see the twitter user count 2010-2021 (in millions). We can see the twitter users are increasing day by day. Similarly they publish all their types of posts through this social media through twitter.

From 2010 to 2021, the user's activity on Twitter increased significantly, and they continued to share posts on the network. Their history is a voyage of ideas, encounters, and social connections. They have probably made contacts, exchanged ideas, and added to a number of Twitter discussions through their tweets.

In figure 3.6 below, we observe that the data of Facebook users from 2012 to 2023 has been reviewed in detail. Today, Facebook is the most popular social networking platform. The daily count of Facebook users is rising. In a similar vein, people's addiction to these social media has increased. Girls in particular are posting depressive types as a result. Even more shocking is the fact that they are committing horrible crimes like suicide for no apparent reason. It now poses a serious risk to our civilization. Thus, we ought to confront it head-on and investigate its solutions in great depth.

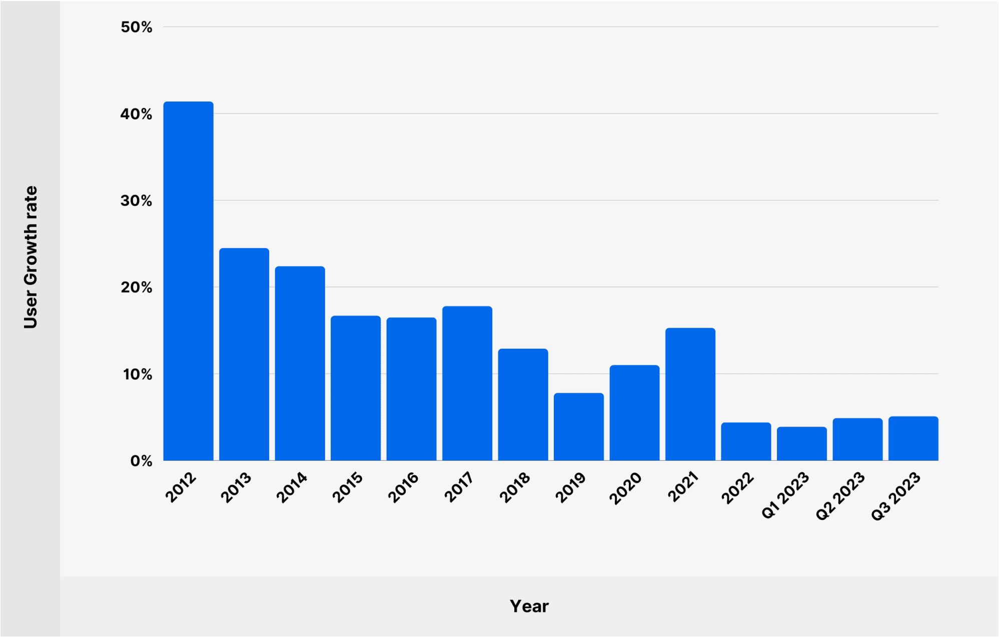

Figure 3.6: Global growth rate of facebook users between 2012 to 2023

According to Facebook's most recent investor report, Facebook currently has 3.049 billion monthly active users (MAUs). The investors' report from the previous quarter indicates a 3.08 percent year-over-year growth in MAUs. Every day, 68.38 percent of the monthly customers will check in on their desktop or mobile devices. Every month, 57.53 percent of the world's active internet users visit Facebook. Every month, 37.75 percent of people on the planet use Facebook.

The Facebook user growth rate is now visible.

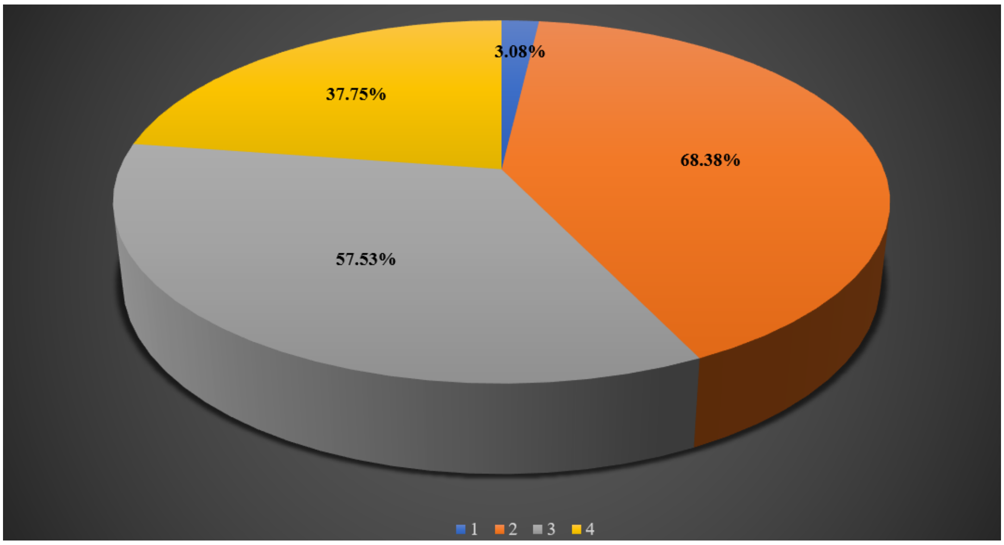

Figure 3.7: Pie chart of global growth rate of Facebook users

3.3.2 Data Processing

The process of data analysis involves data cleaning, conversion, examination, and modeling in order to achieve the following goals: finding useful information, assisting with decision-making, and supplying information for conclusions. Data analysis is an umbrella term for a wide variety of methods that can be approached from different directions. It is useful in many domains, such as science, business, and the social sciences. Data analysis is essential in today's professional environment since it makes decisions more objective and makes business operations function more smoothly.

We started by gathering raw and unlabeled data from Facebook, which was not categorized in any way. Subsequently, we manually tagged the data. The cleaning and preprocessing steps were carried out using the subsequent methodology.

- Every stop word was eliminated.

  A text processing approach used in information retrieval and natural language processing (NLP) is stop word elimination. Remove the punctuation and non-alphanumeric characters from the stop word.
- Eliminate every link that was included in the comment.
- Identifying and eliminating redundant data.
- Occasionally, an item is left empty after the top has been removed.

3.3.3 Building Dataset

As is widely known, a dataset is an assortment of identically typed data. Similar types of data, such as: matrix data, images, music, and text. Utilizing the most recent or real-time data available to us is ideal. Since things can change at any time, it is best to handle the data carefully. However, text data was required for our dataset, and that data was subsequently gathered from the appropriate sources. Since our name implies, we require data from social media, and we have data that we have gathered from multiple types of social media platforms, such as: Facebook, Twitter, Instagram, YouTube comments. Next, we needed to properly label the data that we had obtained:

Now we can see the Unlabeled non processed data examples:

For this dataset, there are just two data columns: real data, where users have indicated their own emotions; and data, where we have classified users as either depressed or not.

Since it is widely known that using labels in their current form is improper, we substitute decimal for the text. For example, assigning a number to depression (1 or 0), and so forth. With our dataset, where 0 represents depression and 1 does not, we proceeded in the same way.

No	Not labelled information
01	finished$tothejustsessiongym!feelinghappinessobstacles-_{}gooddeal$
02	some struggle affects me very dangeriously() ‘** so many tunnel of occasiond
03	i am the best weekend camping find out the label of damages
04	felling exhausted cz the sleep dont come onto my eyes fore ever conclude
05	enjoyin the birthday celebrtion of my taddys good days
06	??? why man why the smile feels numb get beeter depresed
07	finaly cooking the delicious food for my family that was amazing day gd dayz
08	these days i lost my favourite person $_{thatisthecrodycitydontanybodyhelpanyone}$
09	felling so hppy bcz of this dayz i complt my project and get the very gd grade in acamic rslt-$_{}$
10	every monday and every morning i feel like depreesed bcz of his lonelinesscryingggg

Table 3.2: Examples of unlabeled and unprocessed data and problems with labelling

In the given table we labeled processed the data:

No	Example	Label
01	I just got out of a fantastic exercise at the gym! I’m energised.	1
02	One more day, one more challenge. There may seem to be no hope for the future at times.	0
03	I can’t wait to go camping with pals this weekend. Nature, meet us here!	1
04	I’m so tired and depleted, but I can’t seem to fall asleep. Everything just weighs too much.	0
05	I’m spending my birthday today with family and friends. Happy to have another year!	1
06	Putting on a smile, yet deep down, I just feel numb. Will things ever improve?	0
07	Tonight I made a great dinner from scratch. I adore the gastronomic explorations!	1
08	feeling disoriented and isolated in a crowded place. I feel like no one can see me.	0
09	Proud to have finished my most recent do-it-yourself job. I’m eager to display it!	1
10	Every morning feels like I’m facing an insurmountable mountain. Will there ever be light again after this darkness?	0

Table 3.3: Examples of real-world use cases of labeled data

3.3.4 Dataset Distribution

The dataset distribution comprises a varied range of textual information gathered from different social media platforms, including Facebook, Twitter, Instagram, and YouTube comments, with the aim of detecting human depression through text data from social media.

This dataset includes user-shared messages, postings, and comments from a variety of social backgrounds, geographies, and demographics. The distribution reflects the inherent variety of language usage in social media platforms, with a mix

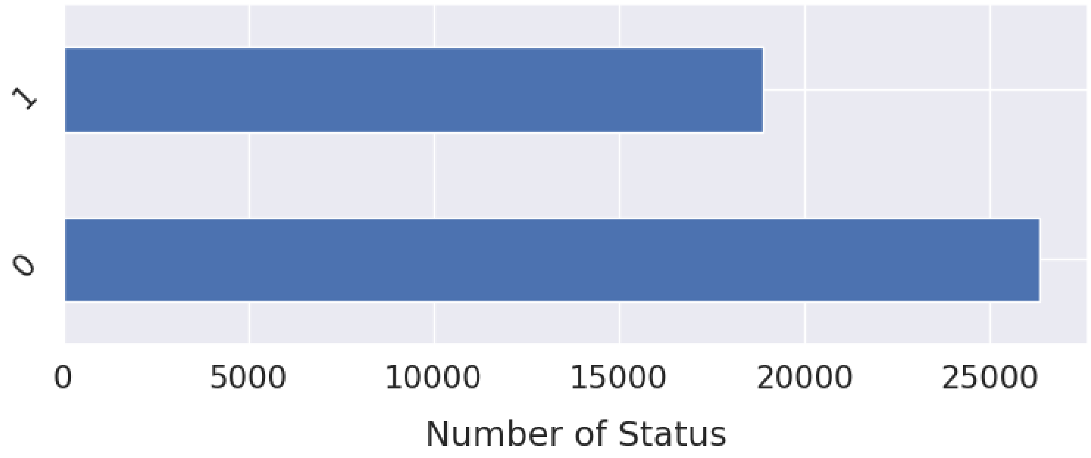

Figure 3.8: Distribution of positive and negative datasets

of texts with positive, neutral, and negative sentiment. We've gathered 45,000 data points here. Data on depression is positive for 20,000 people and negative for 25,000 people.

3.3.5 Feature Extraction

Characteristic extracting to the machine learning, design grouping, additionally picture recognition starts alongside the first decide to determined principles also building efficiency to the characteristics created regarding perceptive to irredundant, with the aim of facilitating the subsequent learning and generalization steps and, in some cases, leading to better human interpretations. Feature extraction and dimensionality reduction go hand in together.

Provided this information has transformed entering the comparable tractable collecting in feature, subsequently broadly related can being achieved on any occasion this area about this information filled through the algorithm surpasses that's handling ability within this whole additionally that the assumed these few of the information is reproduction (instances consist of same items of evaluation composed on each metres additionally yards of this repetition to distorted picture exhibits). That process of selecting a few essential traits from a wider range is re-

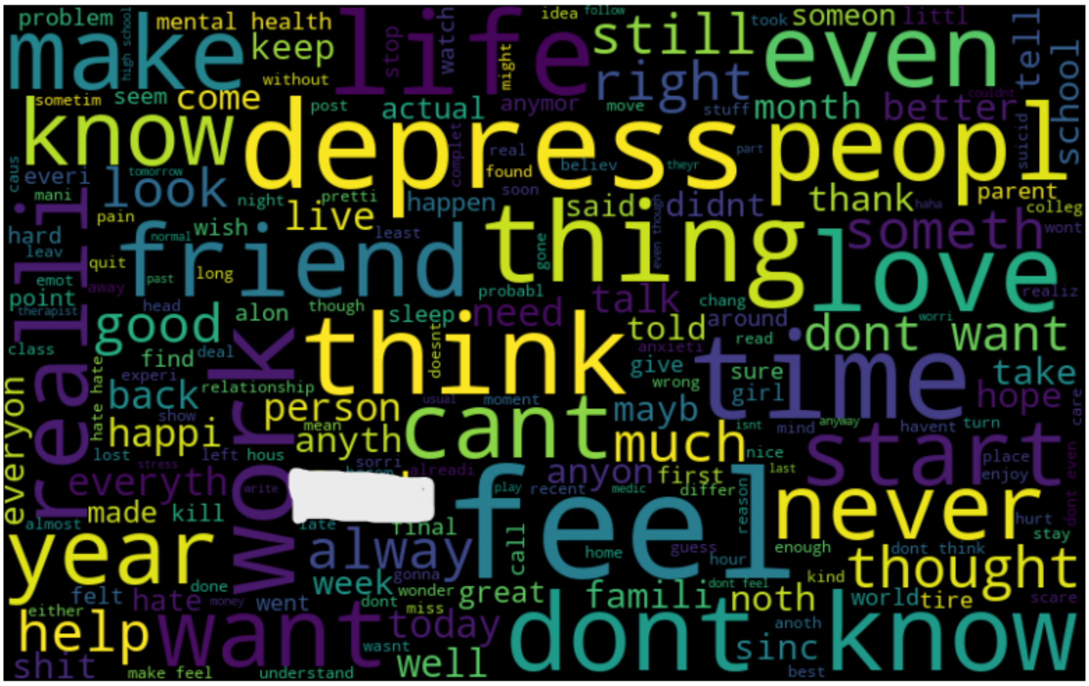

Figure 3.9: Total word cloud from analysis to identify depression

ferred to as feature selection. It is anticipated that the selected characteristics will contain information of interest from the supplied data. Consequently, utilising that decreased portrayal as an alternative to this entire source information will allow this desired job to be completed.

3.3.6 Word Cloud For Depression Positivity

We have used feature extraction to express the terms of depression positive words in figure 3.10 above:

Depression-positive words that represent the ability to bound back from challenges. It summarizes the capacity and perseverance individuals substantiate in Subdue Calamity. Embracing flexibility means conceding one's capacity to navigate difficult emotions and situations with courage and conviction. By promoting flexibility, individuals can find hope and build a foundation for their mental well-being.

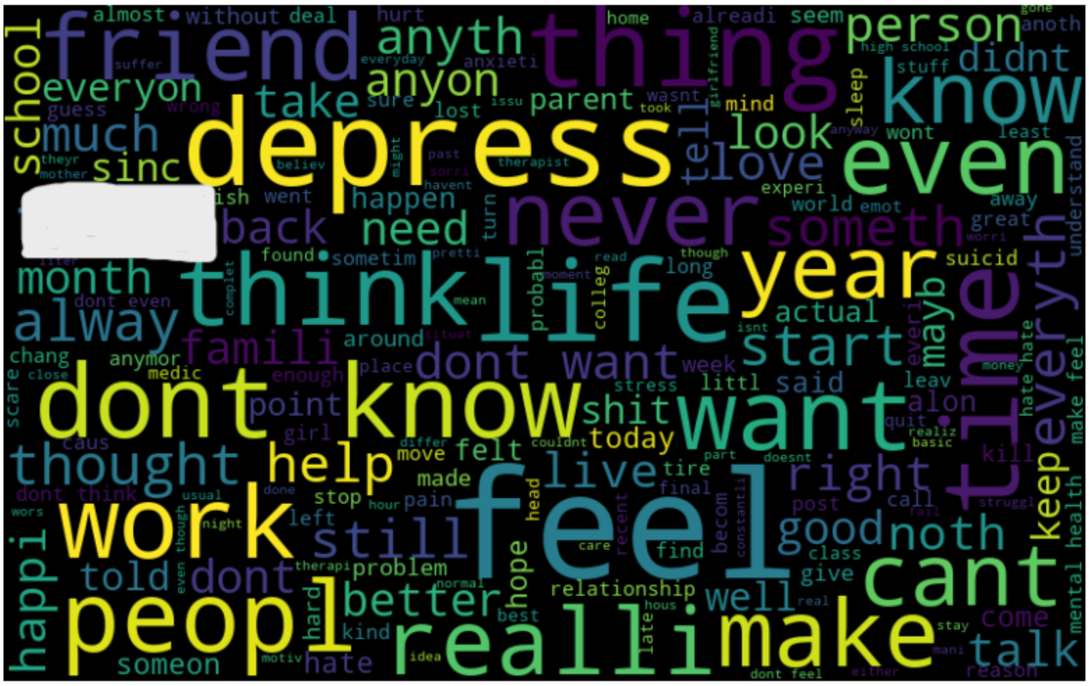

Figure 3.10: Positive word cloud from analysis to identify depression

We found some uncensored words or buzz words during our feature extraction of total and depression positive words. Which can cause obstacles in our academic activities. That's why we have blurred those specific words.

3.3.7 Word Cloud For Depression Negativity

We have used feature extraction to express the terms of depression negative words in figure 3.11 above:

In our total dataset we have found some negative data, that can be referred to as peoples are not depressed because of their positive type of comments. In our whole dataset we maintain our every model accuracy and got this result.

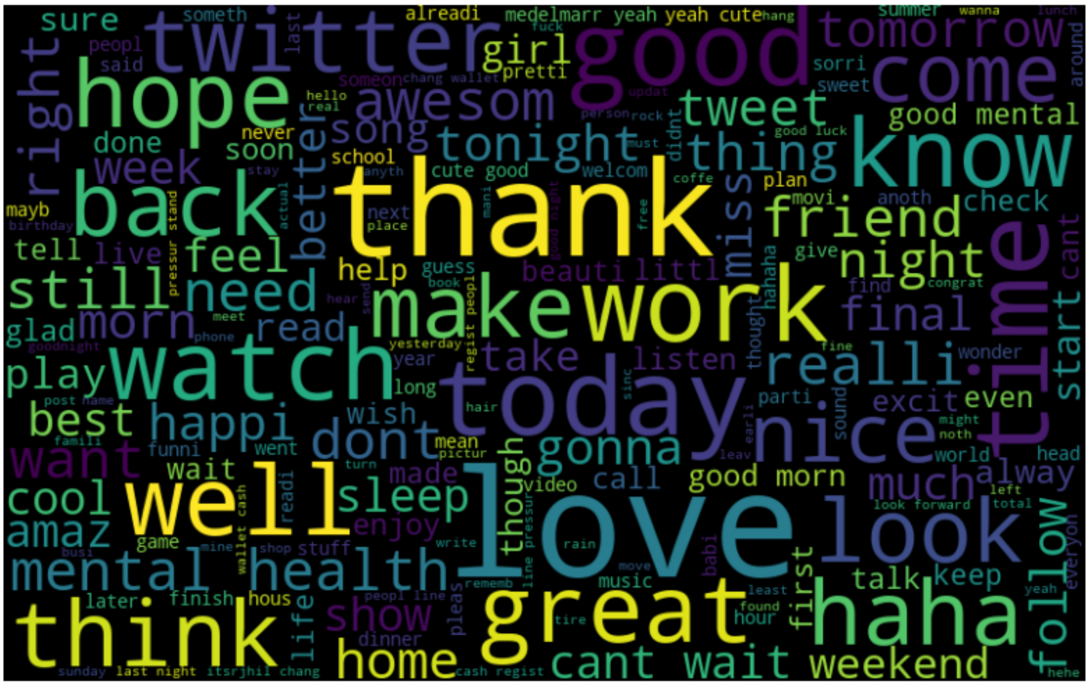

Figure 3.11: Negative word cloud from analysis to identify depression

3.3.8 Training and Testing data

In machine learning, one of the most popular hobbies is researching and creating methods this are discover to information also create forecasts founded this information. That models utilize these incoming information that create the quantitative simulation, that someone use to draw conclusions and evaluate the data, in order to achieve their objectives. Before being used to create a model, these input types are frequently divided into numerous data sets. Three different data sets are usually used in the model-building process: training, validation, and testing.

Total Data	Unique Word	Train Data	Unique Word	Test Data	Unique Word
45218	53405	36174	53405	9044	53405

Table 3.4: Training testing and unique word table of our dataset

3.4 Research Subject and Instrument

It was necessary for us to employ a variety on applications and instruments, the majority on that were freely available tools-to carry out our research on depression detection. We generally used open-source tools, thus there were restrictions. We used the Python programming language to complete this work mostly on our Windows 10 home PC.
Here we can use different types of package, Below we discuss it part by part:

3.4.1 Language Used

Python is a fairly sophisticated programming language, to put it simply. It's a kind of programming language with a focus on object-orientation used to create apps. Its integrated data structure, when combined alongside energetic composing also attach to, it's fantastic also appealing to quickly developing applications. It's also a fantastic language for ML programming. This kind of ML project likewise makes extensive use of this language.

3.4.2 Face Pager

Face pager is used to retrieve publicly accessible data from websites that rely on web scraping and APIs, such as: Facebook, YouTube, Twitter, and other social networking networks. Every piece of information that was scraped from various websites is kept in a SQLite database and may be exported to CSV.

3.4.3 JupyterLab

Jupyter is an interactive on the internet growth surroundings to code, notepads, additionally information manipulation. Users may easily reorganize and customize analysis of information, writing for scientific purposes, artificial intelligence, additionally academic calculating activities because of its flexible user interface. Because of that's modular design, you may easily add new features and enhance the ones that are already there.

3.4.4 Libraries Used

- **Matplotlib:** Pyplot is a set of functions in Matplotlib that are used as one of the tools for plotting data. It enables you to conduct many tasks, such as defining a plot's boundaries and identifying lines within it when creating forms.
- **NumPy:** One common option for manipulating arrays in Python is the NumPy library. It covers the Fourier transform, matrix operations, and linear algebra. An assortment of tools and methods for handling arrays of different sizes is provided by NumPy, a Python package. Mathematical and logical array operations are made possible by NumPy. NumPy is, towards placed this Just that, the numerical computing Py Module. The phrase Python's in numbers is also used with this one.
- **Pandas:** Panda's primary use case is data analysis. JSQN, SQL, and even Microsoft Excel are just a few of the many data formats that Pandas can handle. Panda, for instance, enables users to select, arrange, reset, merge, and modify data.
- **Sklearn:** An intuitive and powerful tool for analyzing predictive data is Sklearn. Everyone can use it for free and customize it to suit their needs, created with Numerical Python, Scientific Python, and matplotlib.
- **Seaborn:** This seaborn is a Python's information representation module that is compatible with matplotlib. It's a user-friendly platform for making informative and aesthetically pleasing data visualizations.

3.5 Conclusion

There are numerous approaches for utilizing text data from social media platforms to identify depression in people. Using natural language processing techniques to examine the sentiment and linguistic characteristics of social media messages is one suggested approach. Patterns indicating depression symptoms can be found by applying machine learning methods, such as neural networks or support vector machines. Contextual data can also improve the accuracy of depression detection models. Examples of this data include the user's engagement and posting tendencies.

Chapter 4

Experimental Result Analysis

4.1 Introduction

We've conducted tests using a wide variety of algorithms and information sources to forecast potential results. This chapter presents additionally discusses the results of our experiments. Chapter 4.2 discusses the detailed experiment results; Chapter 4.3 covers the descriptive analysis; and Chapter 4.4 concludes with a summary of the entire experiment.

4.2 Experimental Setup

We gathered a dataset from social media platforms such as Facebook, Twitter, and Instagram that included messages from both sad and non-depressed people in order to conduct an experiment on the detection of human depression using text data. We preprocessed the data using stemming, stop word removal, and tokenization. After that, we retrieved features like sentiment scores and TF-IDF vectors. For classification, we used machine learning algorithms such as Random Forest, Logistic Regression, and Support Vector Machines (SVM), and we evaluated the models using k-fold cross-validation. To evaluate the efficacy of various methods in depression identification, we lastly evaluated performance parameters such accuracy, precision, recall, and F1-score.

4.3 Experimental Result

It was common to run across difficulties when implementing various algorithms for depression identification. For this reason, we approached the procedure using various techniques. We experimented and researched many approaches to see which would work best for the experiment. We experimented with many approaches to enhance our work and its results.

We made use of dictionaries, content categorization strategies, and readily available Python packages. In the process, we discovered a connection between discouragement and the use of dialects. The sad posts had more frequent phrases. According to our earlier research, simple to understand, outrage, a antagonistic attitude, uneasiness, additionally abusive conversations, ideas, and emotions are frequently observed as dialect markers of melancholy.

We used two distinct datasets, one of which we personally acquired and the other which we gathered from the internet.

4.4 Descriptive Analysis

Our results varied according to the classification methods we employed. We applied the Random Forest Classifier, SVC, AdaBoost Classifier, Logistic Regression, and XGB Classifier to the following problems: BOW from NLP and Instances per Term Counterfactual Document Frequency. We labeled our own data in order to facilitate our algorithm's work. Each algorithm operated on the same dataset, which included our own dataset that we downloaded from the internet in addition to the pre-existing data. Following the completion of the dataset process, we verified the algorithms correctness using Python and its built-in modules.

The performance of various classifiers is displayed in this section. Open-source

Classifier	Accuracy (in percentage)
Extreme Gradient Boosting (XGB)	93.00
Random Forest (RF) classifier	92.50
Logistic Regression	95.30
Support Vector Classifier (SVC)	94.80
AdaBoost Classification	93.60
Term frequency-inverse document frequency (TF-IDF)	93.23
Bag of Words (BOW)	93.47

Table 4.1: Accuracy comparison of multiple classifier models

programmes Jupyter and CoLab were used throughout the entire process. The following seven classifiers were utilized in total: TF-IDF, BOW, SVC, AdaBoost, Random Forest, Logistic Regression, and XGB.

4.5 Comparison The Model Performance With Existing Datasets

In the given table 4.2 we've used two datasets from github and Kaggle that are relevant to our proposed system. We obtained some accuracy, precision, recall, and f-1 score that are compatible with our system as we applied the same model to these datasets that we used for our system. Similarly we can see the comparison between our existing system using Logistic regression, Support vector machine and Random forest classifier.

Dataset	Algorithm	Accuracy	Precision	Recall	F1- Score
Own_Dataset.csv	Logistic Regression	95.43%	96.67%	92.24%	94.4%
mental_health.csv	Logistic Regression	90.12%	92.85%	87.86%	90.21%
Own_Dataset.csv	SVC	95.33%	96.5%	92.3%	94.5%
mental_health.csv	SVC	89.55%	92.12%	86.34%	89.12%
Own_Dataset.csv	Random Forest	94.96%	95.12%	91.3%	93.2%
mental_health.csv	Random Forest	89.33%	90.10%	89.56%	88.32%
Own_Dataset.csv	Logistic Regression	95.43%	96.67%	92.24%	94.4%
Depressed.csv	Logistic Regression	84.40%	81.55%	89.45%	85.21%
Own_Dataset.csv	SVC	95.33%	96.5%	92.3%	94.5%
Depressed.csv	SVC	84.52%	82.43%	89.81%	85.75%
Own_Dataset.csv	Random Forest	94.96%	95.12%	91.3%	93.2%
Depressed.csv	Random Forest	83.34%	89.21%	90.94&	82.84%

Table 4.2: Compare between two dataset of our model performance

4.6 Summary

One of the most crucial components of any experiment is its data. The data provided can have a significant impact on the results of the same sort of experiment. We were certain that the outcomes the fact peoples obtained applying about us experiment records that proved previously accessible on the internet would differ from ours because we used a combination of the two datasets. We may have higher or worse accuracy since we used more data. From the beginning of this study, our aim has been to identify depression using the information we have collected from social media.

We used a variety of machine learning techniques to accomplish our goal. We have employed seven different algorithms in total for this research. Before beginning our work, we had to search for a few various things. We did begin working on our algorithm as soon as we decided on it. Next, we obtained each algorithm's accuracy. As previously stated, we have independently labeled our data. We refer to the word use frequency within the same data class.

We have also discovered that the algorithm is unable to provide an accurate prediction in the event of a little inaccuracy in the data. Depending on the data pattern, it will provide false positive or false negative results. issues such as omitting a letter or putting a sentence in the wrong place. With 95.3 percent accuracy, we achieved the highest accuracy in logistic regression.

4.7 Conclusion

Using text data from social media sites to identify human depression, it was found that the approaches produced encouraging results. By utilizing machine learning algorithms and natural language processing techniques, the models were able to identify depression symptoms from social media posts with a remarkable degree of accuracy. The results of the investigation showed that contextual data, and language aspects were important in improving the detection process efficacy. The trials also shown the possibility of using social media data for depression patients to be identified early and treated.

Chapter 5

Conclusion

5.1 Introduction

We've learned a lot about this subject thanks to this investigation. Even today, mental health remains a touchy subject. This explains why well-informed news and information are not as widely disseminated. We employed machine learning that identify the data concealed information for this purpose. ML makes predictions based on these patterns.

5.2 Summary of The Study

As we have already indicated, Facebook data was used for our study since we wanted a social media platform where users felt at ease or frequently shared their emotions. Facebook lived up to our expectations. Our algorithms were trained and learned phrase patterns with the aid of this data, which was then utilized to diagnose depression. A small number of issues were initially resolved earlier.

We succeeded in achieving the objective we had been pursuing. We received varied results from different algorithms. In the following section, we went over it in more detail.

5.3 Main Contribution of Our Existing System

In the given figure 5.1 we can see the unique concept of our proposed system. We can use ensemble methods for better accuracy. These techniques aim at improving the accuracy of results in models by combining multiple models using a single model. We use different types of models like XGB classifier, Logistic regression, Random forest, Support vector machine, Adaboost, Naive bayes. The ensemble approach was used to determine the average accuracy of all the models, as each model produced varying levels of accuracy.

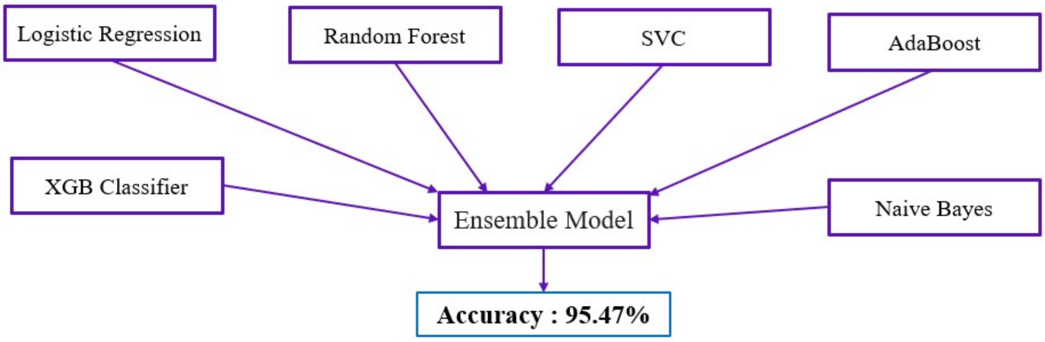

Figure 5.1: Unique contribution of our proposed system

5.4 Possible Impacts

We think that the results of our work vary depending on the location. We made sure our work was done correctly because of this. We wanted to ensure the authenticity of our work. We employed a unique kind of dataset for this reason. Our research has applications in computer science and physiology, among other domains. Because it can aid in the regulation of depression within the general population, it would be beneficial to society as a whole. It can also save lives and make our lives easier. If this effort complies with its future work plan, it can also aid in the general public's knowledge of the circumstances. It can also aid in disseminating knowledge about the dangers of depression that are often overlooked.

5.5 Implication of Further Study

In our own work, there are numerous opportunities for additional research in this area. Many strategies for improving our work were discovered. We have discovered several errors, as we have already indicated, and these errors will help us improve our research. We have had false negatives in our projections, as we have previously said. Such an event was not anticipated. Should this arise during real-world implementation, it can result in a fatal mistake. Fixing false negatives is part of our plan. In order to teach our system to handle minor errors like misspelled or incomplete words, this is necessary.

We have additional objectives in mind. We want to improve upon our prediction. In this manner, we can employ algorithms to interact with an application that allows users to forecast the type of phrases they will input. Additionally, we would like to be able to anticipate what type of depression an individual is experiencing based on data from social media.

We believe that this type of activity can help people by offering advice on how to improve their circumstances. We may also be able to assist individuals without disclosing their personal information to third parties.

5.6 Conclusion

Our work indicates that the approaches and research results we used are excellent. We believe and hope that this study's conclusion will further research in this area. We will have a lot of options for growing our work thanks to this study. Throughout our work, we have discovered a few errors. We observed fresh avenues for which this research might be developed. It will enable us to address any mistakes or other issues that arose while working on this research. Additionally, we are considering how to combine algorithms and create more effective solutions to address the issues raised by that investigation to later years. That research would enable us regarding the learn additional regarding the entire area during investigations.

Everyone anticipate which would further advance the development of technology that supports human mental health and open up new avenues for assistance. For a better study outcome, we have combined psychological and practical information. Based on the results of this investigation, we intend to provide a novel method for detecting depression.

References

[1] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601, 2018.

[2] Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of depression-related posts in reddit social media forum. *Ieee Access*, 7:44883–44893, 2019.

[3] Nafiz Al Asad, Md Appel Mahmud Pranto, Sadia Afreen, and Md Maynul Islam. Depression detection by analyzing social media posts of user. In *2019 IEEE international conference on signal processing, information, communication & systems (SPICSCON)*, pages 13–17. IEEE, 2019.

[4] Anu Priya, Shruti Garg, and Neha Prerna Tigga. Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science*, 167:1258–1267, 2020.

[5] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *2014 science and information conference*, pages 372–378. IEEE, 2014.

[6] Mandar Deshpande and Vignesh Rao. Depression detection using emotion artificial intelligence. In *2017 international conference on intelligent sustainable systems (iciss)*, pages 858–862. IEEE, 2017.

[7] Cassendra Frederick. Multiple vdt prolonged exposures: On fatigue depression, ocular strains of sleep management pre-liminary smaller gerontology

overviews. *Asian Journal of Behavioural Sciences*, 3(2):61–79, 2021.

[8] Raymond Chiong, Gregorius Satia Budhi, Sandeep Dhakal, and Fabian Chiong. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, 135:104499, 2021.

[9] Kuhaneswaran AL Govindasamy and Naveen Palanichamy. Depression detection using machine learning techniques on twitter data. In *2021 5th international conference on intelligent computing and control systems (ICICCS)*, pages 960–966. IEEE, 2021.

[10] Md Zia Uddin, Kim Kristoffer Dysthe, Asbjørn Følstad, and Petter Bae Brandtzaeg. Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*, 34(1):721–744, 2022.

[11] Riza Bob Subhan. On the review of image and video-based depression detection using machine learning. 2020.

[12] Irene Li, Yixin Li, Tianxiao Li, Sergio Alvarez-Napagao, Dario Garcia-Gasulla, and Toyotaro Suzumura. What are we depressed about when we talk about covid-19: Mental health analysis on tweets using natural language processing. In *Artificial Intelligence XXXVII: 40th SGAI International Conference on Artificial Intelligence, AI 2020, Cambridge, UK, December 15–17, 2020, Proceedings 40*, pages 358–370. Springer, 2020.

[13] . What are we talking about when we talk about covid-19? biopolitics, the current pandemic and a sociological approach to global disorder. , pages 41–61, 2020.

[14] David William and Derwin Suhartono. Text-based depression detection on social media posts: A systematic literature review. *Procedia Computer Science*, 179:582–589, 2021.

[15] Roman Klinger. Where are we in event-centric emotion analysis? bridging emotion role labeling and appraisal-based approaches. In *Proceedings of the Big Picture Workshop*, pages 1–17, 2023.

[16] Hatoon S AlSagri and Mourad Ykhlef. Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Transactions on Information and Systems*, 103(8):1825–1832, 2020.