

**MACHINE LEARNING TECHNIQUES FOR DETECTING HUMAN DEPRESSION
USING SOCIAL MEDIA DATA**

BY

**NAFIS IQBAL
ID: 191-15-2675**

AND

**ESRAT JAHAN
ID: 191-15-2556**

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

Supervised By

Dr. S.M. AMINUL HAQUE
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

MD. SABAB ZULFIKER
Senior Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

30 JANUARY 2023

APPROVAL

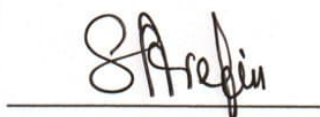
This Project titled “MACHINE LEARNING TECHNIQUES FOR DETECTING HUMAN DEPRESSION USING SOCIAL MEDIA DATA”, submitted by Nafis Iqbal 191-15-2675 and Esrat Jahan 191-15-2556 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 30 January, 2023.

BOARD OF EXAMINERS



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

Chairman



Dr. Mohammad Shamsul Arefin
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Ms. Sharmin Akter
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



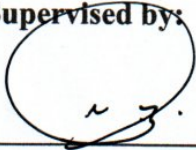
Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

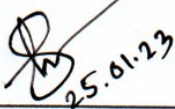
We hereby declare that, this project has been done by us under the supervision of **Dr. S.M. Aminul Haque, Associate Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree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

Supervised by:



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

Co-Supervised by:



Md. Sabab Zulfiker
Senior Lecturer
Department of CSE
Daffodil International University

Submitted by:



Nafis Iqbal
ID: -191-15-2675
Department of CSE
Daffodil International University



Esrat Jahan
ID: -191-15-2556
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully.

We really grateful and wish our profound our indebtedness to **Dr. S.M. Aminul Haque, Associate Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of machine learning to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

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

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work. Specially our friends who helped us during our work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Depression is one of the most serious issue among human civilization in modern time. It is one of the well-known mental health issues in this era. In this paper we worked on this mental health issue known as depression. This has affected countless people of different gender, age and race. It is also taboo for some people to talk about which makes it more so serious. Now people also share their depression thoughts through the social media. It is also to be mentioned many people does not even realize they are depressed but their posts on social media shows they are. One of the social media where people share their thoughts are twitter and this is what we chose to gather our data from twitter. This method required both depressed and not depressed social media data so our algorithm can distinguish between depressed and not depressed post. NLP and Machine Learning is used for the process.

This research aims to explore the use of machine learning techniques for detecting human depression using social media data. The study will focus on the use of various algorithms including TF-IDF, BOW, XGB Classifier, Random Forest Classifier, Logistic Regression, SVC, ADA Boost Classifier and Naive Bayes. The objective of this research is to develop a model that can accurately identify individuals who may be at risk for developing depression based on their social media data. The research will utilize a dataset of social media posts and interactions, which will be preprocessed and used to train and test the machine learning algorithms. The performance of each algorithm will be evaluated using metrics such as accuracy, precision, recall, and f1-score. The final model will be chosen based on the highest performance. The research will also consider the ethical aspects of using social media data for detecting depression, such as privacy concerns, accuracy and reliability, bias, access, and responsibility. The results of this research could have significant implications for the identification and treatment of depression, as well as for the overall well-being and quality of life of those affected.

The results of this research will provide valuable insights into the use of machine learning techniques for detecting human depression using social media data. The findings will be useful for mental health professionals, researchers, and policymakers to understand the potential of social media data for the early identification of individuals at risk for depression. The research will also provide valuable information for the development of more accurate and efficient models for detecting depression in the future, which could ultimately lead to better outcomes for those affected and a reduction in the overall burden of depression on society.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v

CHAPTER 1: INTRODUCTION	PAGE
1.1 Introduction	1
1.2 Motivation	3
1.3 Rationale of the Study	4
1.4 Research Question	5
1.5 Report Layout	6

CHAPTER 2: BACKGROUND	PAGE
2.1 Introduction	8
2.2 Related Work	10
2.3 Research Summery	12
2.4 Scope of the problem	13
2.5 Challenges	13

CHAPTER 3: RESEARCH METHODOLOGY	PAGE
3.1 Introduction	15
3.2 Research Subject and Instrument	16
3.3 Data Collection Procedure	18
3.4 Statistical	19
3.5 Implementation	34

CHAPTER 4: EXPERIMENTAL RESULT AND DISCUSSION	PAGE
4.1 Introduction	36
4.2 Experimental Result	36
4.3 Descriptive Analysis	36
4.4 Summery	37

CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	PAGE
5.1 Impact on Society	39
5.2 Impact on Environment	40
5.3 Ethical Aspects	41
5.4 Sustainability Plan	43

CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	PAGE
6.1 Summery of the study	45
6.2 Conclusion	45
6.3 Possible Impacts	46
6.4 Implication of further study	46
References	47

LIST OF FIGURES

FIGURES	PAGE
Figure 2.1.1 Social Media user count growth from 2017 to 2027	08
Figure 2.1.2 Top 10 Countries with the Highest Rates of Depression	09
Figure 2.1.3 Top 10 Countries with the Highest number of depressions	10
Figure 3.1.1: Proposed model for research work	15
Figure 3.4.1 Twitter user count	20
Figure 3.4.2 Word cloud from consideration	24
Figure 3.4.3 Random Forest work flow	26
Figure 3.4.4 Logistic Regression example	27
Figure 3.4.5 AdaBoost workflow	29
Figure 3.4.6 TF-IDF workflow	31
Figure 3.4.7 Bag of Words workflow	33
Figure 3.5.1 Data training method	35

LIST OF TABLES

TABLE	PAGE
Table 3.2.1: Tools used for the project	16
Table 3.3.1: Data count	19
Table 3.4.1: Unlabeled non processed data examples	22
Table 3.4.2: Labeled processed data examples	23
Table 4.3.1 Algorithm accuracy	37

LIST OF ABBREVIATIONS

ML	Machine Learning
SM	Social Media
TP	True positive
FP	False Positive
FN	False Negative
NLP	Natural Language Process
BOW	Bag of Word
TF-IDF	Term Frequency Inverse Document Frequency

CHAPTER 1

INTRODUCTION

1.1 Introduction

One of the most known word in this era of human history is “Depression”. It would be hard to find any person who uses the modern communication methods and have not heard the word depression. Specially during the recent COVID-19 pandemic this topic got more attention. But there is question that might occur which is what is this depression. From the World Health Organization’s description of depression, we can say the general definition of depression is it is a mental disorder which can be characterized by less pleasure from the previously enjoyable activities or overall lack of interest to anything and the main one which is persistent sadness. Other symptoms can also be felt like losing appetite or lack of sleep. Other symptoms can be poor concentration and tiredness.

Because of recent corona virus we have seen all kinds of problem arise in our society. Because of waves and waves of corona pandemic lockdown we have faced so many disruptions in our life. Also, for a long time we had to stay behind the closed doors. This created problems which include rise of depression among the general population. Corona pandemic have influenced everyone regardless their age, gender or situation. Many people lost their job and the had to face isolation, income loss, fear of death etc. Level of insomnia, anxiety was increased significantly. One of the reports suggest 18% alcohol and 36% drug use went up during the pandemic. For all the reasons mentioned depression level also went up. The numbers suggest it went up almost 25%.

Data suggests that around 5.07 billion people are using internet that means around 72 percent people use or have access to the internet. From this around 4.74 billion or 59.3% people use social media around the world. The numbers are only increasing day by day which shows that more and more people use the internet to access the social media. Social media is a great tool to communicate with each other which allows them to share their personal thoughts either using personal message or through the use of public posts. They share data like opinions, photo, voice record, video and many more. This data also reflects the user’s emotions such as mood, sentiment or feeling.

Detection of depression is a method that can be done from text, photo or video. For our particular task we have used the depression detection from text of social media. Our study has the aim to analyze data from social media platform which have public access to their data in order to detect depression from the gathered data. We purpose to use different machine learning method as and stable and efficient method for detecting depression.

As mentioned before according to WHO depression is a mental disorder where a person has symptoms like losing interest in previously interested activities also felling increasingly sadder than before. It creates other problems like other mental health issue with physical problems such as reduce of work efficiency and laziness. Symptoms can vary from mild to severe. Patients can also feel guilty without reason or worthless.

According to the Debra Fulghum Bruce, PhD there are five types of common depression.

- **Bipolar Disorder:** It is also known as manic depression. In this type of depression there are different episodes of mood which ranges from the high energy to the low depressive periods. When a person in his or her low phase he or she can or might have the symptoms of major depression which we will mention later on.
- **Major Depression:** Major depression is kind of self-explanatory because it is one of the most known depression. People with severe depression calls it “global gloom”. People suffering from major depression shows signs like loss of interest, thoughts of suicide, trouble in making decision, trouble in concentration, feeling worthless also maybe guilty, lack of energy to do anything, insomnia or trouble of sleeping, feeling agitated maybe feeling sluggish or slowed down mentally and physically.
- **Persistent Depression Disorder:** This type of depression has an old name which is dysthymia. As the name suggests it is usually characterized by persistent or even with continues symptoms that can or also might last at least two years. People who are diagnostic with the chronic depressive illness are able to easily cope with everyday tasks but rarely display signals of any excitement. Change of energy level, food consumption, sleep condition and self-esteem might get changes.
- **Postpartum Depression:** This type of depression can be seen among one in seven new mothers. It creates worry, exhaustions among the new mothers which can make it difficult for them. Which makes accomplishing their daily task harder.

- **Seasonal Affective Disorder:** This term is also self-explanatory because this type of disorder occurs during mostly in the winter time. Days became shorter and less sunlight. Its effect goes away during the summer or spring time.

Covid Depression Spike: During the covid time rapid spread of corona virus has created considerable level of fear, anxiety around the globe. To cope with the virus and death many decisions were taken among which we can see lockdown where people had to stay home for months without leaving their home for a moment. For this reason, many people have developed depression without even realizing.

Symptoms of Depression:

- Feeling of frustration
- Frequently interrupted sleep
- Change of behaviors like appetite decrease
- Aches in the body, pain, discomfort in the stomach and other type of pains also can be the symptoms of depression
- Having difficulty to sit still
- Trouble concentrating on work or other things
- Weight might also increase or decrease because of depression

Machine Learning: We know machine learning is a special automation process for data analytics. It is also known as a subset of artificial intelligence. It uses provided data to learn patterns and create its own pattern and make judgment using this without human touch.

1.2 Motivation

Worldwide almost 280 million people suffer from depression according to the information shown on WHO site. That means almost 3.8% population of the world is suffering from depression. Also 700000 people suicides because of depression. Also, this are the numbers that have been reported but there are many cases that are not reported.

Twenty first century has added many useful things to our list one of which is social media. More than half of the human population use social media. Social media is a great way to connect with each other and share their thoughts. People like to share their emotion on the internet through the social media. They also share their sadness and stories. Where we can find clues about their depression.

Our goal is to use this data from the social media to make a model where using ML we can automatically detect depression from a text data. Because we know depression is still a taboo for the people also lot of people don't know they suffer from the depression. So, they can't even find help for their depression. But if the social media can detect their users mental state, they might help them without the user being awkward.

From all the social media we choose the twitter to collect our data from. Because it has a wide range of users and people share their feeling via small passage mostly to the public. It is also easy to collect huge amount of data related to our work form there. Even though we know that human emotion specially depression is a problem form most people to detect from. We kept out data collection according to our ease of data sorting. So that it can work with basic emotion detection.

1.3 Rationale of the study

It is a really common practice among the recent researches to use AI or ML in different fields in order to automate their process. Also, it is being used in the emotion detection and text type detection researches a lot. Our work follows this path where we are using the social media data and detection depression from them using ML. There have been many works done in this field previously. Same type of themes was practiced but in different ways. Different data gave different outcomes to different works. Our work focuses use of recent acquired data from the social media data and mixing with previous data and creating our own dataset.

A person might feel that is difficult for them to concentrate and assess when he or she starts suffering from depression, also it affects the agility. Depression also affects one's brain and might generate difficulty when recollecting their memory. Depression can affect a person's psychology, but also it can affect a person's physical structures. It is also possible that depression might be able to affect one's central nervous system of a person. It can or might also cause permanent damage in the brain, that is why many who suffer from depression might

feel difficulty when remembering things. It is also noteworthy that around 20 percent people who suffer from this might never recover.

Around 16% or 1.2 billion people of world population are young people who are around 15 to 24. Also, most people who suffer from depression is around 18 to 29 years. Also, most significantly impacted group is the age group of people are age group from 45 to 65. It is important to note that 90 percent of the people who are adult in between age group of 18 to 29 uses the social media.

That is why if we use the current social media, it will be more likely that we are working with data from the young group of people. That means it will allow us to get more data on depression detection of young people mostly if we are successful in our work the reason being most people who uses social media are young people and it is more likely that young people will share their emotion on the internet rather than the older group of people.

1.4 Research Questions:

Before we even start with our research work there are question that are essential to our work. These questions will create an outline for our work. The following questions are the ones we encountered

- Why do we need depression detection?
- Why do we need to use the social media for data?
- How will we collect our required data?
- What kind of methods to use for data collection?
- Why we will use machine learning?
- Which kind of ML we will use?
- What will be our criteria of data collection?

We will discuss about a few quantitative results from studies that has been done before moving on to the results. In order to find keywords for our work we looked into the social media posts related to our work. It enhanced the performance in our depression detection, and as a result we started to notice more indications of frustration. We tried to make sure that word representing despondency includes feelings like frustration, anger, happiness, rage.

So, as a result to this research it will offer insight on how to identify depressive symptoms in written texts as effectively and quickly as possible. Depression is known to impact human language use in the past. We have also studied different papers in search for great techniques which can accurately detect depression better than the others.

We have two main sources of data for our project. One source is an online available dataset which contains data from twitter. It has around 36000 data related to depression. It has 3 type of data one depressed one not depressed and the last one is neutral data. We have collected data from twitter using tools. This data dates from January 2022 to June 2022. We manually labeled the data what we collected from the twitter. Depending on different criteria we also labeled our data depending on our online dataset

1.5 Report Layout:

In the **Chapter 1** is where the introduction part for our research project is and we shed light on the basics of our thesis. We discuss how social media is related to the users and how it has a great effect on our thesis data collection. Our motivation is also discussed on this chapter. We have also mentioned our efforts and basic steps to finish this work. We have also talked out potential outcomes in short.

Chapter 2 is the background study of our work. In this chapter we talked about our idea behind our work, why we felt to work on this idea. We have also tried to provide basic overview of global social network analysis which we have gathered from the internet. We have also discussed what challenges we have to work on.

Chapter 3 Experimental analysis part has the main purpose of establishing our research strategy and describing it. This is our main point which we have discussed. Because our main importance is the type of data we are using and it bears the most importance on our work. Because our result and outcome depend on the type of data we are choosing. It has the most lime light in this chapter. We have talked about our two datasets and discussed about its aspects. We explained and examined the dataset.

In the **Chapter 4** experimental result and discussion is this particular chapter where we have also talked about the studies we have done during our experiment. We discussed the

mathematical techniques we used on this research and the algorithms that we used. We have talked about the concepts, experiment finding and description of them.

Chapter 5 has the summary of our total research, conclusion about what we have done so far, recommendations, and where we might be able to work with it in the future with the idea of where this might be implementable and what effects it might have.

CHAPTER 2

BACKGROUND

2.1 Introduction:

Because of the most recent pandemic there were many changes in our life. These changes were felt all over the world for most of people. It has increased the tension and anxiety for most of the people around the world. One of the most important things during this pandemic is that people had to stay home for so long many of them started to develop depression. So, for this reason it is necessary to monitor health of the general public.

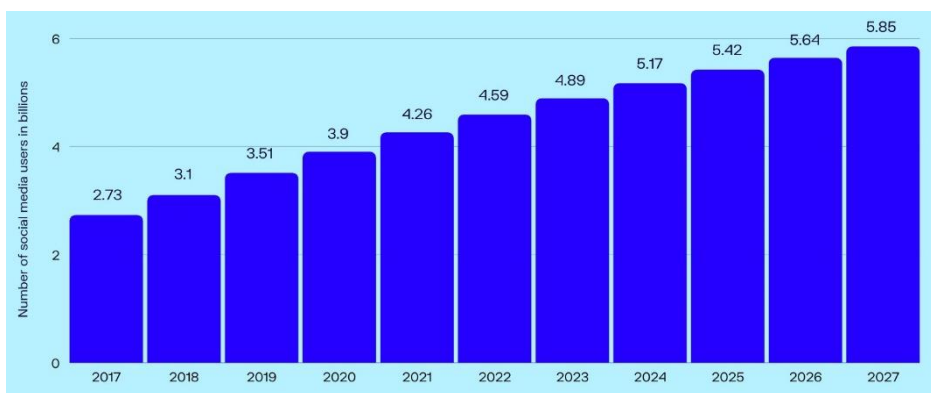


Fig 2.1.1 Social Media user count growth from 2017 to 2027

In **figure 2.2.1** it shows that social media user count from 2017 to 2027. Here they showed from actual count of users how many users were using social media from 2017 to 2021. After the 2021 means from 2022 to 2027 they have projected social media user count using previous few years social media user count. All the user numbers in billion, which means in 2021 there were 4.26 billion. This information was taken from “statista.com” website.

Many changes have occurred because of the lockdown. Many people started to use the social media as a result of not having anything to do during the time of lockdown. Changes in number of social media users have raised significantly during last two years. Social medias like Reddit, Twitter, Tiktok and Facebook etc. saw a huge user boost during this time.

This study focused on the people who uses the social media to show their emotions because in between all the emotion there are emotions like sadness, loneliness and other symptoms which shows the depression might have affected the person.

The goal of this study is to make use of the available data on the internet which are input by the normal users. Every input is a data for our study depending on the information stored on the post of the user. As we know there are many types of data on the internet most are on text. Data like age, gender, preference and many more. Which gives us a lot about a person. But a person’s emotional situation can be understood by the posts or their text data. Our goal is to collect this data from the social media and make use of it.

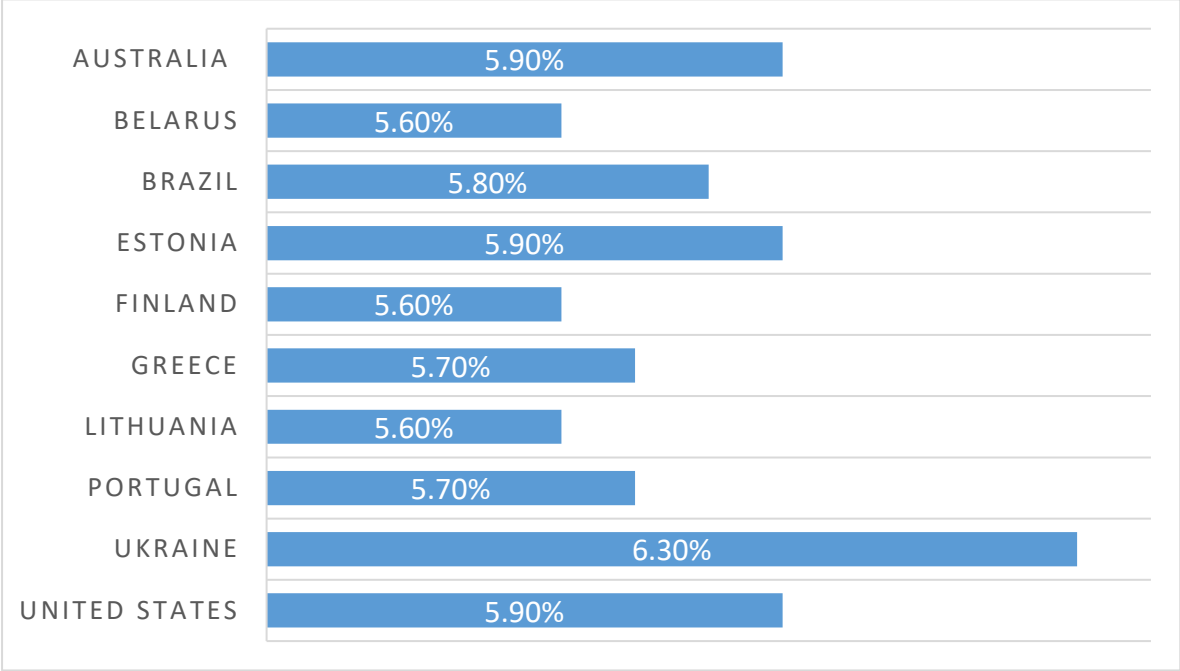


Fig 2.1.2 Top 10 Countries with the Highest Rates of Depression

In the **2.1.2 figure** it shows top ten countries that have the greatest number of depressed people. It means Ukraine has the greatest number of people in percentage from total number of populations. According to the ranking Ukraine has most percentage of depressed person among their population. This information was taken from “worldpopulationreview” site

If the amount of data had to put in numbers, then every minute 527,700 photos are shared on Snapchat. More than 4 million videos are watched, 456,00 posts are twitted on the Twitter. Around 16 million texts are sent and around half million comments are posted and 293000 statues are updated on the Facebook alone.

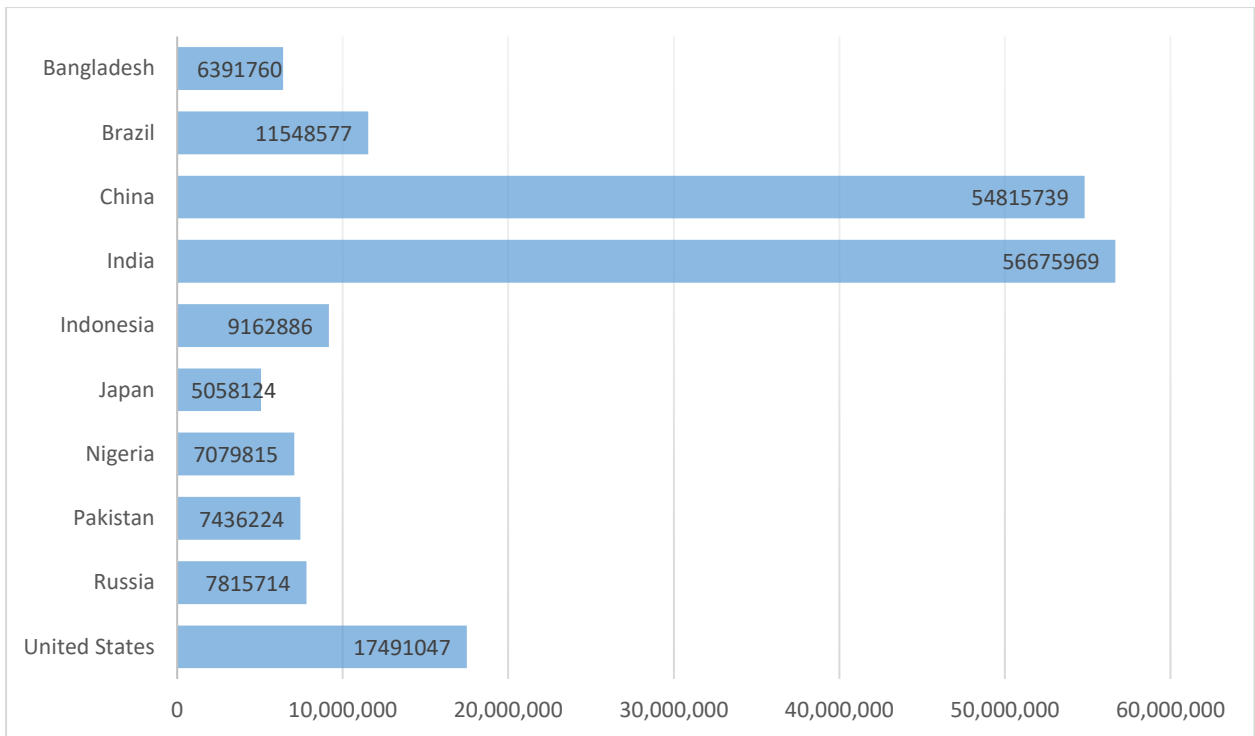


Fig 2.1.3 The top 10 Countries with the highest number of depressions

In the **2.1.2 figure** get information about total number of depressed people among all the countries. Here depression cases are shown in normal numerical number. So, regard of a nation's total population this number shows top countries total cases of depression. Depression case numbers was taken from "worldpopulationreview" website.

2.2 Related Work:

The authors claim that individuals whose personalities or demographics fit a certain profile have more possibly to share information about their own mental health diagnoses on to the social media. According to the results of our research, the MLP classifier had the best performance when it came to recognizing and understanding the presence of sadness in the reddit the social media network. Its accuracy was 91%. A score of 91% was achieved by the MLP classifier, which demonstrates the power and effectiveness of the combined features. [1]

This article presents a methodology used for obtaining usernames from users who post on to the social media in order to determine the level of depression risk. An invitation was sent out to fifty people on Facebook, asking them to contribute their posts from the most recent years. These posts from users were then incorporated into a machine learning model. It has been shown that depression can or might lead to major mental illness or even suicide, as well as

how to detect depression using machine learning approaches. Additionally, it has been established that machine learning approaches can be used for detecting depression. [2]

Using methods of machine learning, the proposed system is able to determine whether or not the user is depressed. Before determining whether or not there is any indication of depression in the text, the algorithm performs a reading of emotion based on the text that was provided as input by the user. The fact that the user can access the system in the comfort and privacy of one's own home shields him or her from the social stigma that is prevalent in his environment. [3]

The diversity also the richness of the features set of the method that was proposed has allowed it to achieve a higher accuracy than the method that was proposed previously. The frequency words are chosen in accordance with the higher frequency that the user feels they should have. The sentiments of each tweet are determined by applying the "feeling of the sentences" method, and also then the overall average sentiment from all tweets is determined for the user-mixed sentences. [4]

Not only because of the emergency itself, but also because of the social problems that followed, like unemployment, a lack of resources, and a financial crisis, people's mental health problems would be especially sharply increased by a major catastrophe, such as diseases caused by the coronavirus 2019 (Covid-19). [5]

Their analyses shed light on societal sentiment through time and bad issue topics in Weibo posts. Therefore, conducting research on posts on social media that have a negative tone in order to obtain a deeper understanding of the experiences of the Chinese general people during the outbreak of COVID-19 and to provide examples for other nations is something that could be helpful. The findings from the posts on Weibo offered useful public health advice., and it is possible that transparency and scientific advice will reduce public worry. [6]

The emotion of sadness can be detected in people through the use of sound or video recordings by machine learning, which employs a variety of different methods. It has been put to use in medical diagnostics as well as in classifying and diagnosing many diseases, including neurodegenerative diseases. Images and videos can be parsed for facial feature data, and put through the wringer of artificial intelligence analysis tools in order to establish a diagnosis of depression. [7]

In this piece, Twitter is considered to be the primary data source for analytical purposes, specifically in the form of tweets from the users. When compared to the sizes of storage space required to hold the same amount of text, audio, and video information is considerably smaller. Twitter has proven to be the best platform for applying emotional artificial intelligence to detect depression. This is due to the fact that the maximum amount of characters that may be included in a tweet is set by Twitter. [8]

The authors constructed the EmoCT dataset with the intention of sorting tweets about Covid-19 into distinct feelings in order to research the issue of mental health, and they used the BERT (ft) model, which was derived from predicting the emotion label from just one label, in a single-label classification job on the one million randomly selected English tweets data on April 7, 2010. [9]

A methodical approach to determining how frustrated people is with the messages they receive from social media platforms. It is impossible for the algorithm to determine a person's level of frustration because it gathers information from tweets using keywords rather than from posts on Facebook. The machine learning model provides for a six-point scale to be applied to the depression criteria. [10]

2.3 Research Summary:

Our work on this project focuses on different available method out in the community. We have used total of seven algorithm and it also have different algorithm with our own dataset. Here we have used twitter as our primary data source. As we mentioned earlier our dataset has both previously used data and our own data that we acquired recently. It will allow us to look for things like impact of new data that was added by us from the same source and it will allow us to get accuracy of seven algorithm we used. We have kept the new data as the previous dataset we combined with. That means there are same class and same type of labels. Python was our main language of choice, and ML techniques were used in our feature extraction procedures. There are many algorithms that has been used in countless works on the same topic. So, there are many existing information already on the internet about this topic. There are many ways that others have done their wok on this topic. This means we saw many opportunities which we previously thought available as new but after few searches we saw works already have been don on the topic we wanted to work with.

2.4 Scope of the Problem:

It is essential for influence analytics systems that can be applied in a broad variety of contexts to have the capacity to recognize sadness in written content. It's possible that those who are confined inside for the duration of the lockdown may experience some degree of mental health distress as a result. For example, students may feel upset or depressed within the context of schoolwork, young people may be anxious about their jobs or careers, and businessmen may be concerned about their companies.

This article works on by focusing to the part where input will be a text then there will be a proper outcome which will be able to determine proper state of emotion. This proposed technique will allow a person to find if a person whose data was input is depressed or not. System has been taught by the so many uses input because of this reason, it is able to tailoring recommendations and give appropriate response that addresses main problem of the topic.

Our thesis has one particular objective which is to build a concept. A concept that will be able to meet the requirement of all the person who are or might be suffering from depression. A concept where a model based on text which is computationally efficient. A great goal where this project is able to help all the people out there.

We focus on detection of depression before it comes to a incurable stage or to a stage where a person might take bad decision. Bad decision like suicide can be prevented for many people if the concept is a success.

- **Data availability:** One of the biggest challenges in using machine learning techniques for detecting depression is the availability of appropriate data. Social media data is vast and diverse, and it can be difficult to obtain a representative and reliable dataset for training and testing machine learning algorithms. This can impact the accuracy and generalizability of the models developed.
- **Algorithm performance:** Another important aspect of the problem is the performance of the machine learning algorithms themselves. There is a need to develop algorithms that can accurately and reliably identify individuals who may be at risk for depression based on their social media data. This can be challenging, as depression is a complex condition that is influenced by a wide range of factors.

2.5 Challenges:

Data analysis to detect depression is greatly aided by textual information, which is the most widely used form of communication.

During our work on this project, we faced challenges like

1. **Data availability and quality:** One of the biggest challenges is obtaining a representative and reliable dataset for training and testing machine learning algorithms. Social media data can be difficult to obtain, and it may not always be of high quality. Additionally, social media data can be highly unstructured and may require extensive preprocessing before it can be used for training machine learning models.
2. **Algorithm performance:** Another challenge is the performance of the machine learning algorithms themselves. Depression is a complex condition that is influenced by a wide range of factors, and it can be difficult to develop algorithms that can accurately and reliably identify individuals who may be at risk for depression based on their social media data.
3. **Generalizability:** Another challenge is to ensure that the model generalizes well across different demographics and populations. This includes avoiding bias towards specific groups, and ensuring that the model can accurately identify individuals from diverse backgrounds.
4. **Model's interpretability:** Machine learning models can be complex and difficult to interpret, which can be a challenge when trying to understand how the model is making its predictions and identifying the features that are most important for depression detection.

In this research, we attempt to provide a framework that can determine extent of emotional distress experienced by users based on the content of their social media posts. Like text summarization and machine translation, depression detection is a job that has a reputation for being difficult because to the complexity of language. Identifying signs of depression in communication written in a variety of languages is a very challenging and difficult endeavor. Since it is challenging to collect enough annotated training data, the supervised machine learning technique has not been extensively used or adopted. This is the complexity of determining a depressive disorder which was also quite high during the COVID-19 situation using several types of online social media. There are a number of challenges that may be faced when working on the use of machine learning techniques for detecting human depression using social media data. Some of the key challenges include:

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction:

ML methodological frameworks for identifying depressive disorder or dd checking out the online posts or data from the internet from users. For our particular work we have work considered the social media known as Twitter. Our study covers a wide range of people because we are using the English language that makes our data collections range wider. This will allow us to cover people from different age group. Then there are people of different groups like normal people, student or other type of occupation. Our work model contains two class one depressed another non-depressed.

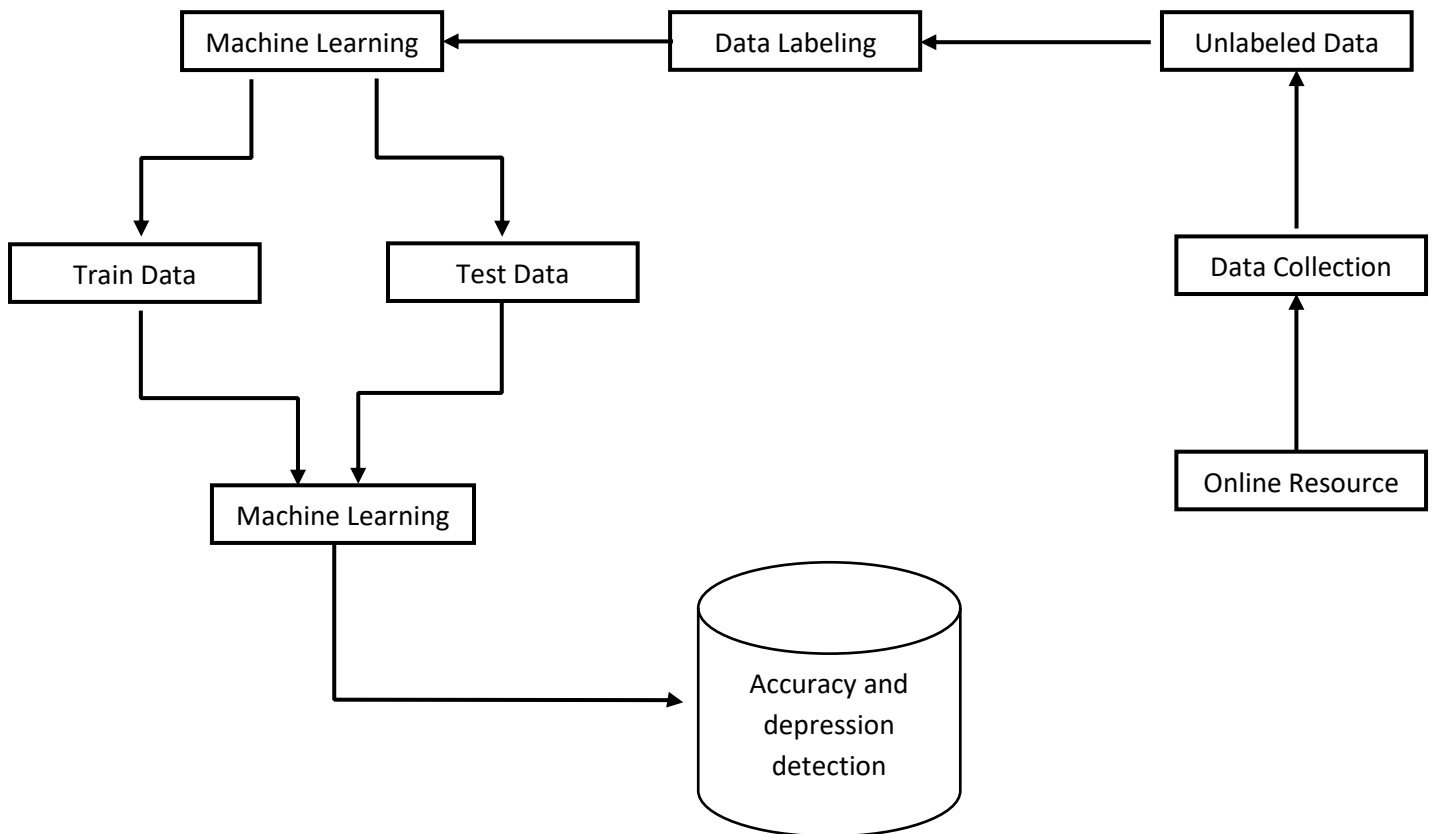


Fig 3.1.1: Proposed model for research work

Figure 3.1.1 shows the basic model that we have used on this research project. First step is to look for data available on the internet and then extraction of data from the source. Then we

will get the unlabeled data. But we need labeled data for our work. So, we have to label data that we have collected. Then this data will be useable for the machine to be fed. Then we can use this data to train and test data using our desired ML methods. Then this algorithm will give us accuracy and also allow us to detect depression which is our goal here.

We have to process the raw data so that they can be used for ML and this process is also called data preparation. It is an important process for the ML model and also crucial stage. It is also necessary to get rid of data that are unwanted in the process.

3.2 Research Subject and Instrument:

For the purpose of this research on detecting depression, we had to use different type of software, tools and most of which of open-source tool. There was restriction because we have used open-source tools mostly. This work was done on our personal computer which has windows 11 and mostly using python programming language

Table 3.2.1: Tools used for the project

Package Name	Description	URL Link
Python	To put it simply, Python is a very advanced programming language. It's a type of language used for developing applications with an emphasis on object-orientation. It has built in data structure which combines with dynamic typing and binding which is great and attractive for fast application development. Also, it is a great programming language for ML. It is also a widely used language in this type of ML projects.	www.python.org

Face Pager	Face pager is used for fetching publicly available data from social media platforms like YouTube, Twitter and other websites on which work on the basis of APIs and web scraping. All the data scrapped from different sites stored in an SQLite database and also available for exporting to csv.	github.com/strohne/Facepager
Jupyter Lab	Jupyter is a web-based development environment which is interactive for codes, notebooks, and data tinkering. Its adaptable user interface makes it simple for users to customize and rearrange data science, scientific reporting, machine learning, and scientific computing operations. Its modular structure makes it open to customization by adding new features and improving upon those already present.	jupyter.org

Libraries used:

- **Matplotlib:** Among the tools for plotting data in Matplotlib is a series of functions known as Pyplot. It allows you to do things like identify lines in a plot, set the bounds of a plot during form generation, etc.
- **NumPy:** The Python package NumPy is a popular choice for array manipulation. Linear algebra, the Fourier transform, and matrix operations are all under its purview. Python's NumPy is an array of objects and techniques for working with arrays of varying dimensions. NumPy allows for the mathematical and logical execution of arrays. To put it simply, NumPy is a Python library for numerical computation. Also, this term used for "numerical Python".

- **Pandas:** Panda's main function is data analysis. Pandas is compatible with a wide variety of data formats, including JSQN, SQL, and even Microsoft Excel. As an example, Panda allows users to merge, reset, choose, organize, and alter data.
- **Sklearn:** Sklearn is an easy-to-use and effective program for analyzing predictive data. free for everybody to use and adapt to their own purposes Made using matplotlib, SciPy, and NumPy.
- **Seaborn:** This seaborn is known as a matplotlib-compatible Python package for visualizing data. It's an accessible platform for creating visually appealing and insightful data visualizations.

3.3 Data Collection Procedure:

Collecting data is the process which is mostly self-explanatory. Though there are many ways to gather data. There are many aspects that had to be met when data is being collected. Aspects like data type, when was data produced, if it meets our work expectation and so on.

There are so many types of data available for use and because of internet it is easily accessible. There are available tools which allows the easy collection of data. One of the tools is Face pager tool. We used this tool for our data collection process.

Face Pager: It's an automatic data retrieval tool developed by Jakob Junger and Till Keyling in 2019.

Its ca retrieve data from many social media sources like Twitter and others. It can be set to retrieve data in different presents. It is possible to change settings according to ones need.

There are different parameters that can be used to collect data for our work we used parameters like

- id
- message

While our data collection depending on the parameters, we also collected data like where we had both depressed and non-depressed data.

Table 3.3.1: Data count

Type of data	Quantity
Total number of data	40000
Positive or not depressed data	28000
Negative or depressed data	12000

Label Encoder:

In the field of machine learning, we often work with datasets that include numerous labels in either one column or many columns. Words or numbers may be used to represent these designations. The training data are often labeled with words in order to make the data human-readable or to make the data more intelligible.

Label encoding refers to the process of transforming labels into a format that a computer can interpret. A necessary step in this process is to convert the labels to a numerical format. Ultimately, ml algorithms may make an informed decision as to how those labels should be used. In supervised learning, this phase of pre-processing the structured dataset is very significant.

3.4 Statistical Analysis:

i. Data Analysis:

In order to extract meaningful insights, draw valid conclusions, and provide useful guidance for making decisions, data analysis is performed. Data analysis is multifaceted and multimethod, including a wide range of procedures that go by a wide variety of names

and finding use in many fields of business, science, and the social sciences. The modern business world could benefit from data analysis if it leads to more data-driven decisions and improved operational efficiency.

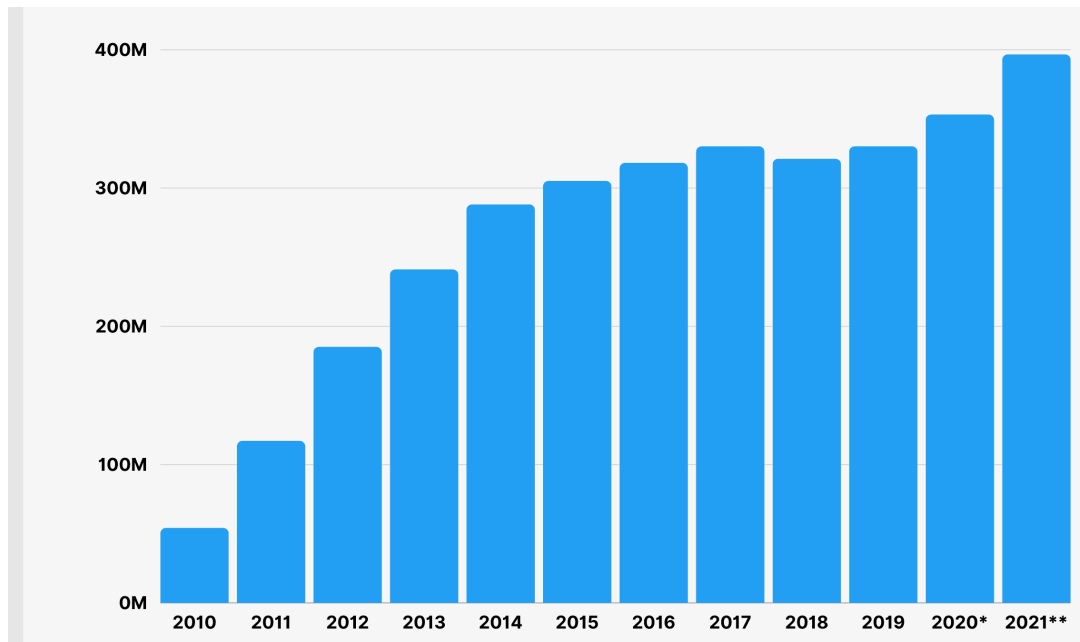


Figure 3.4.1 Twitter user count

Our primary data source is twitter and according to backlinko.com **Figure 3.4.1** shows number of twitter users from 2010 to 2021.

ii. **Data processing:**

Discovering usable information, informing conclusions, and providing assistance for decision-making are all goals of the data analysis process, which consists of examining, cleaning, converting, and data modeling to accomplish these objectives. Data analysis encompasses a broad range of techniques known by many different names and may be tackled from numerous angles. It finds application in a range of fields, including business, science, and the social sciences. Data analysis plays a crucial role in the current professional world, allowing for more objective decision-making and facilitating the smoother running of businesses.

We began by collecting unlabeled data and raw data from Facebook, both of which lacked any kind of classification. After that, we assigned labels to the data by hand. The following approach was used for the purposes of cleaning and pre-processing.

- Removed all the stop word.
- Remove all the links were present in the comment.
- Finding duplicate data and removed it.
- Sometimes, after the top has been eliminated, the article is empty.

iii. **Building dataset:**

Building a dataset for machine learning techniques for detecting human depression using social media data can be a complex and challenging task. The first step in building a dataset is to collect appropriate data from social media platforms. This may include text data, such as posts and comments, as well as other types of data such as images, videos, and metadata. It is important to ensure that the data is representative of the population being studied, and that it is obtained in a legal and ethical manner, such as obtaining informed consent from individuals.

When collecting data for machine learning techniques for detecting human depression using social media data, there are a few key factors to consider in order to ensure that the dataset is suitable for training and testing machine learning algorithms. These include:

1. **Representativeness:** It is important to ensure that the dataset is representative of the population being studied. This may involve collecting data from a diverse range of individuals, including those from different age groups, genders, ethnicities, and socioeconomic backgrounds.
2. **Quality:** The data should be of high quality and should be free of errors and inconsistencies. Social media data can be unstructured and may require a significant amount of preprocessing to make it usable for machine learning algorithms.
3. **Quantity:** A sufficient quantity of data is needed to train machine learning algorithms. The more data that is available, the more accurate and reliable the algorithms will be.
4. **Timeliness:** The data should be relevant and up-to-date, as depression can be a dynamic condition that can change over time.

Table 3.4.1: Unlabeled non processed data examples

No	Unlabeled data
1	RT @samthingsoweto: I?m sorry for the silences. I thought I was wack but today I only discovered how dope I am. #depression
2	RT @vigorous_man: Modern men suffer from depression because of: 1. Porn 2. Sugar 3. High body fat 4. Poor sleep 5. No purpose 6. Low Testo?
3	@michtosincere Heureusement car si j'étais élu j'oscillerais entre la dépression profonde et les élans d'enthousias? https://t.co/3pBZ6M5mi2 (Trash data)
4	RT @LaQueenJ: Its high time we accept that unemployment is the main cause of depression among the youths
5	It wasn't that long ago that I was in need of such support for clinical depression. Even then these services were u? https://t.co/oOpou3kQsb
6	RT @ThisIsMduh: @LindokuhleMboma Twitter: You?re HIV positive. Person: I?m not. Twitter: Prove it Person: *shows results* Twitter: You?
7	RT @heresyfinancial: During the Great Depression the US government burned crops and slaughtered cattle Why? To? help? with the? problem?
8	I might add that I had just told them that the cause of this episode was gender dysphoria and depression over trans? https://t.co/qiUrauTcn7
9	@fcoconsrv @Suzy_NotSuzy I fell for the craziness too as I am immunocompressed and wore a mask everywhere, mostly s? https://t.co/1y0VEojcu4
10	RT @starsmitten_: @sleepy I?ve come to realize depression is like a multiplier debuff, so whatever other debuff we?re hit with during the t?

As we have shown in **Table 3.4.1** there are only two data columns for our dataset one of which is actual data where users have showed their own emotion and in the other column, we have labeled in numerical data. As it can be seen that this data is not properly pre-processed. In this data there are unwanted things in the dataset. We can see links, punctuation marks and name. Also, there are unwanted stop words which can be problem while working on algorithms.

Table 3.4.2: Labeled processed data examples

No	Example	Label
1	did i miss my chance does anybody else feel like they will never be able to experience the joy of certain big things in life like the path you took is too far astray i do about a lot of things like college marriage having a son a real relationship comfort lasting friendship trust sanity accomplishment	0
2	goodmorning i hate mondayzzz tt	1
3	i hate being steps back from anything i want to do in life oh you want to get a job sorry nobodys going to hire you unless you put your life on hold and go to therapyoh you want a relationship suck it loser not today get your ass to therapyoh you want to maintain your friendships hope theyre still around in a year to the doctors with youits just like i get it but i also have a life to get to and its starting right now today it doesnt seem worth all of the lost time and effort to eventually get to the same square one regular people are starting frommaybe theres something great waiting for me on the other side but its getting to the point where im not very interested in finding out	0
4	rt humansofny i dont think im going to miss eighth grade its been a tough year a lot of my friends are struggling with depr	0
5	flowers check cologne check candles check date crap bottle of lotion check its going to be a good night	1
6	mijavera yes excited ive never had one before and everyone else does i feel so left out lol	1
7	my thoughts just listen my cause of death will always be suicide unless someone kills me i want to achieve small goals and kill myself after as punishment for trying	0
8	was so sad that i felt nauseous	0
9	rt noturmcm me wow im actually starting to feel happy again my depression	1
10	therapy via skype skype therapy for overcoming anxiety	1

As we can see in **Table 3.4.2**, we can see the dataset has been preprocessed and labeled. From the previous table we can see this table has many changes. Changes like no punctuation marks, no links. Also, we can see that there is no uppercase letter neither. We have also labeled our data. There are there columns now one of them is new. The new column is labels and the other two are data and number column. We have pre-processed data and this will allow the algorithms to work without any problems.

iv. **Feature Extraction:**

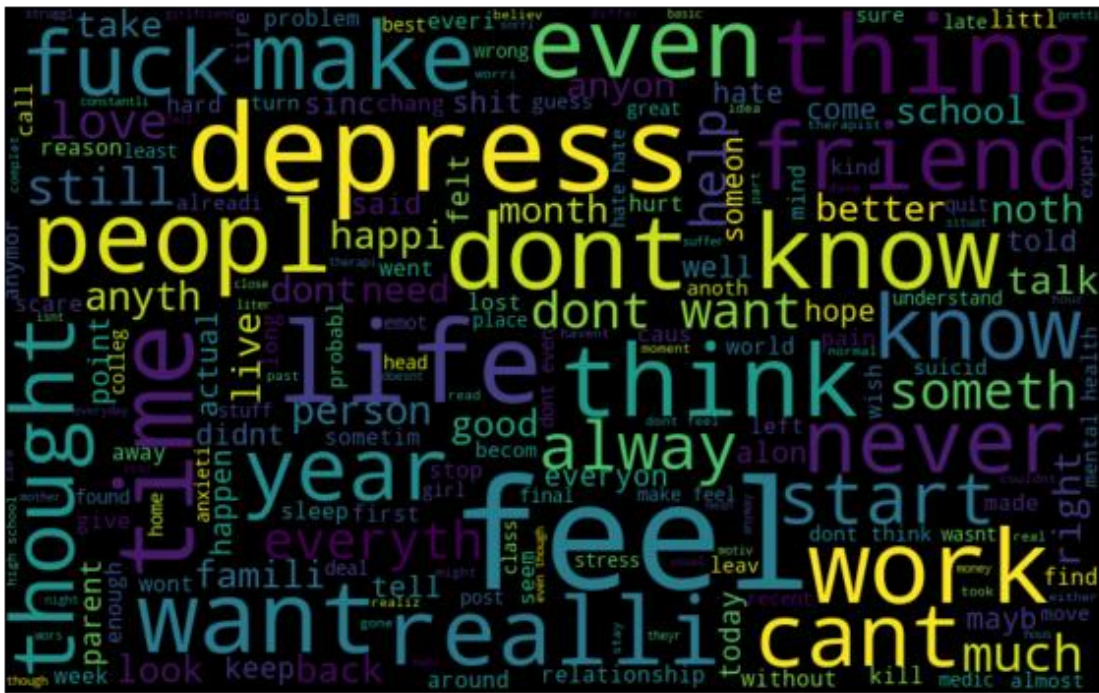


Figure 3.4.2 Word cloud from consideration

In **Figure 3.4.2** we have world cloud. This is the world cloud of our own dataset. In our dataset we have total of forty thousand individual data. In every data we have different amount of data which allows to show emotion of different people. When we made world cloud from our dataset, we got the words which are most used in the dataset. Which means words like depress, feel, can't etc. words should be in the dataset.

With the goal of facilitating the following learning and generalization steps and, in some cases, resulting in better human interpretations, feature extraction in ML, pattern classification, and image recognition begins with an initial set of measured values and constructs effectiveness of (features) designed to be insightful and non-redundant. Dimensionality reduction and feature extraction go hand in hand.

Whenever the size of the data fed into an algorithm exceeds its processing capacity in its entirety and it is suspected that some of the data is redundant (examples include identical units of measurement written in both feet and meters and the monotony of pixelated image displays.), if the data is turned into a more manageable collection of characteristics, then much more may be accomplished. Feature selection refers to the process of picking a few key characteristics from a larger pool. Information of interest in the input data is expected to be present in the chosen features. As a consequence, the intended job will be able to be carried out by employing this reduced representation rather than the whole source data.

v. **Training and Testing data:**

Researching and developing algorithms that can learn from data and provide predictions based on that data is one of the most common activities in the area of machine learning. As a means of accomplishing their goals, these algorithms first use the input data to build a mathematical model, which they then employ to make inferences or assessments. These kinds of inputs are often partitioned into many data sets before they are utilized to build a model. The model-building process typically employs three distinct data sets: training, validation, and test.

vi. **Classification Model:**

- **XGB Classifier:** Machine learning is the practice of creating software that can learn without being taught. It's a branch of artificial intelligence that uses mathematical and statistical methods to design and build computer programs. Neural networks are one sort of data categorization system that form the basis of many machine learning algorithms. In the field of machine learning, this XGB neural network classification is the gold standard. It was meant to identify spoken words in audio recordings. The XGB classifier uses a two-step process to identify speech: training and recognition. Training involves feeding the audio source to the classifier and giving it a set of target words to create a model. The model is then used to identify unknown words in the audio recordings. After training, the classifier can be used for speech recognition- it identifies words with high accuracy and precision in noisy environments. XGB is used for automated speech recognition in many applications such as dictation services and voice-enabled vehicles. Some gaming platforms have also adopted it for automated text-to-speech functions. The XGB speech recognition classifier is also used in medical applications for analyzing breath sounds, heartbeats and more. In addition,

military organizations use XGB for automatic speech translation; this allows human translators only for high-priority translations.

- **Random Forest:** For both classification and regression, use the RF Classifier's tree-based approach. In the context of machine learning, it is an algorithm for creating a tree-based hierarchy. It's an AI technique that constructs a hierarchy of "decision trees." As an ensemble technique, Random Forest Classifier generates numerous decision trees and then takes an average. The issue of overfitting is alleviated in this manner. Because it can be applied to any situation where there is a large amount of data, machine learning is now one of the most discussed subjects in business. RF Classifier is an extremely well-liked ML method., which has a lot of advantages over other algorithms.

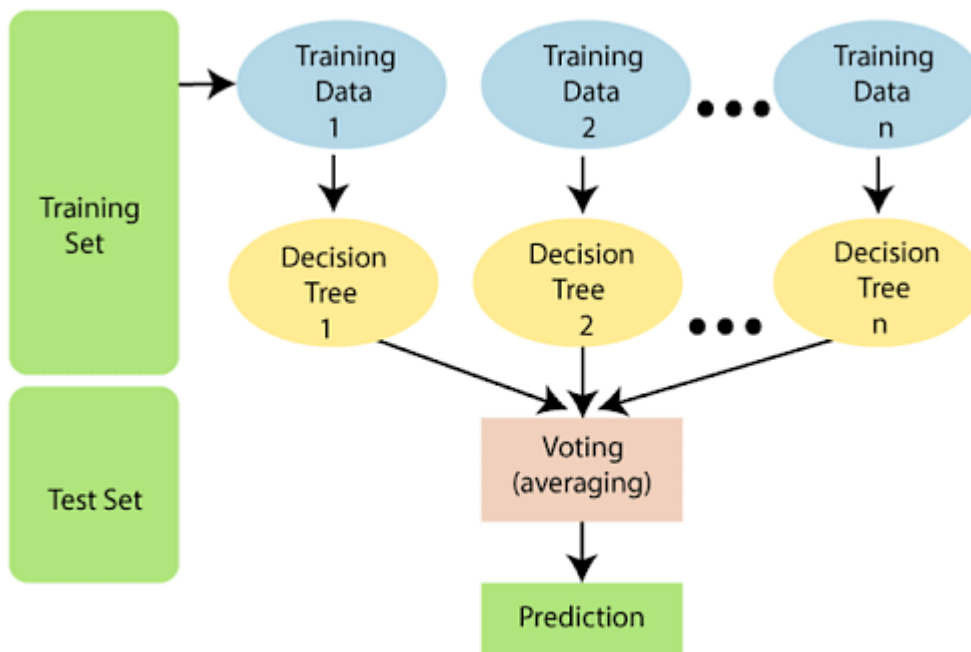


Figure 3.4.3 Random Forest work flow

Figure 3.4.3 is a visual representation of “random forest” algorithm. Figure 3.4.2 was taken from “simplilearn.com”. As we know random forest is a supervised machine learning algorithm. It selects random data from sample from train dataset. For every training data it creates a new decision tree. It is determined by voting after averaging the decision trees. Most voted prediction is selected as the final result.

The algorithm was first published in 1997 and was created to process large datasets. It can be applied to any problem that involves classification as long as there are some training data sets with labels for each possible category, and each set has at least 2 samples. Random Forest Classifier belongs to supervised classifiers because it needs

training data sets given by a supervisor. Once there are enough datasets, this algorithm makes predictions based on them through succeeding stages: selection, splitting and out-of-bag estimation.

- **Logistic Regression:** Logistic regression is a sort of classification method. It makes a decision between two outcomes based on a variety of parameters taken separately. So, in that case, what does this imply? One definition of a binary result is one in which there are only two distinct outcomes that might take place: It's either going to happen (1), or it won't happen at all (0). To put it simply, independent variables are anything other the dependent variable that may affect the outcome of a study on the final result (or dependent variable). Therefore, if you are working with binary data, the appropriate method of analysis is called a logistic regression. When the outcome or determining factor is of a binary or categorical type, you are dealing with binary data; to put it another way, you will know you are dealing with binary data if it falls into one of two categories.

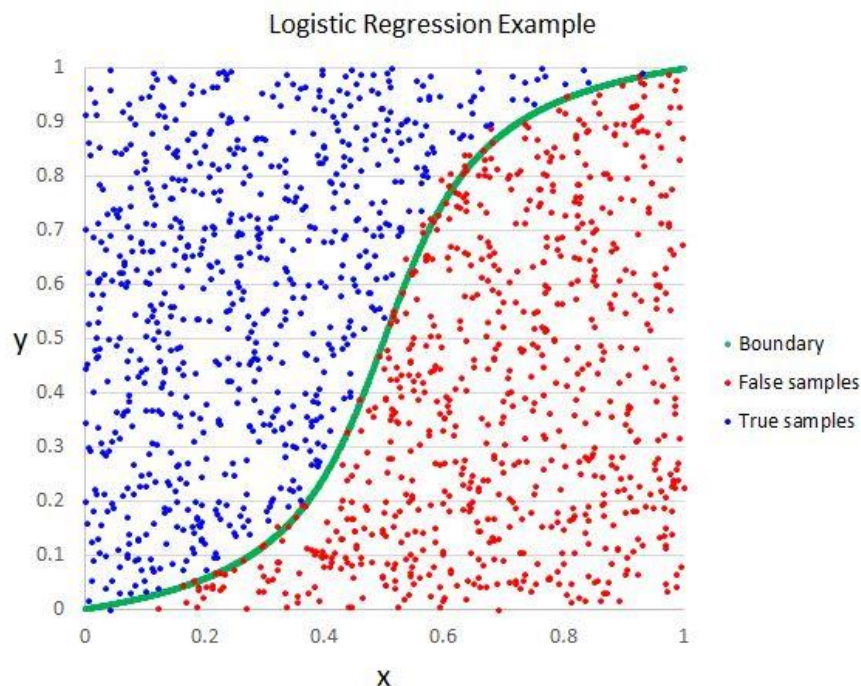


Figure 3.4.4 Logistic Regression example

The **figure 3.4.4** is Logistic regression which is a supervised learning algorithm used for classification tasks. It works by using a logistic function to model the probability of an event occurring. The logistic function is defined as: $f(x) = 1 / (1 + e^{(-x)})$ where x is the input to the

function. The output of the function is always between 0 and 1, which makes it suitable for modeling binary outcomes (e.g., 0 or 1, yes or no, true or false). To train a logistic regression model, you need to provide a training dataset that includes input features and a binary label indicating the class of each example. The model will learn the relationship between the input features and the output label by adjusting the parameters of the logistic function. During the training process, the model will make predictions on the training examples and compare them to the true labels. It will then use an optimization algorithm to adjust the parameters of the logistic function to minimize the error between the predicted and true labels. Once the model is trained, you can use it to make predictions on new examples by inputting the features into the trained logistic function and using the output to predict the probability of the example belonging to each class. You can then choose a threshold (e.g., 0.5) to decide whether the example should be classified as one class or the other.

- **SVC:** By definition, SVC is a known as nonparametric clustering method that does not presuppose anything about the data's underlying structure, including the number of clusters or their relative sizes. Our prior research has shown that it performs best with low-dimensional data; so, a preprocessing step, such as performing the principal component analysis is usually necessary if somehow the dimensionality of your data is large. The original approach has been improved in a few different ways, and these improvements provide specialized algorithms for computing the clusters by only computing a fraction of the edges in the adjacency matrix. Several of these modifications have been proposed.
- **AdaBoost:** With AdaBoost, you can boost the efficiency of your existing machine learning technique. When employed with less capable students, it is most effective. These are models that perform just slightly better than random chance when applied to a classification challenge. One-level decision trees are by far the most popular type of algorithm used in conjunction with AdaBoost since they are the most suitable. In an attempt to build a robust classifier, "boosting" combines the outputs of many weaker classifiers. To do this, a model is constructed by chaining together several simpler models. First, by feeding in the training data, a model can be built.

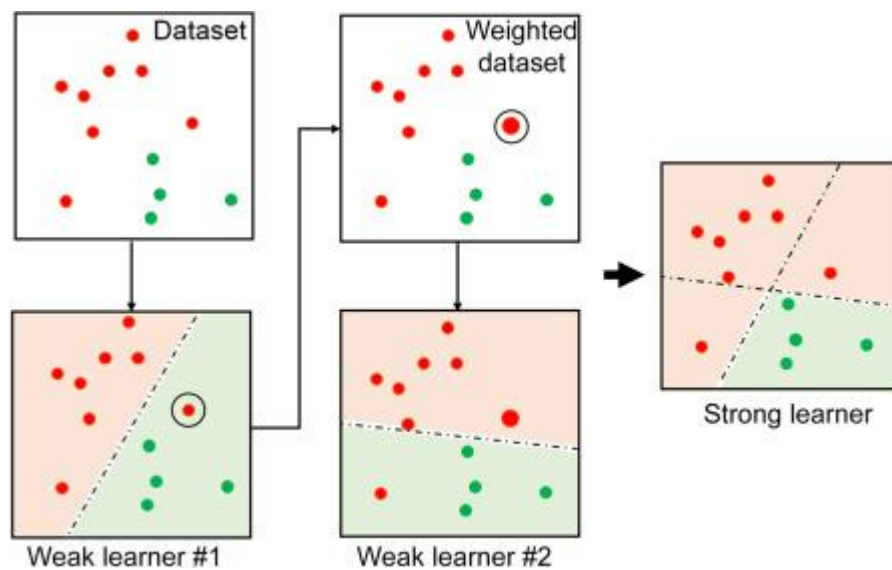


Figure 3.4.5 AdaBoost workflow

Figure 3.4.5 shows AdaBoost (short for Adaptive Boosting) is an ensemble learning algorithm that can be used for classification or regression. It works by combining multiple "weak" learners to form a strong learner. A weak learner is a model that performs better than chance, but not significantly better.

The basic workflow of AdaBoost is as follows:

1. Select a base learning algorithm (e.g., decision tree).
2. Train the base learner on the training data and make predictions.
3. Calculate the error rate of the predictions.
4. Increase the weight of the misclassified examples so that the base learner pays more attention to them on the next iteration.
5. Train the base learner again using the updated weights.
6. Repeat steps 2-5 for a fixed number of iterations or until the error rate reaches a certain threshold.
7. Combine the weak learners into a single strong learner by weighting each learner according to its error rate.

During the training process, AdaBoost pays more attention to examples that are misclassified by the previous weak learners. This helps the algorithm to focus on the hard examples and improve the overall accuracy of the model.

Once the model is trained, you can use it to make predictions on new examples by inputting them into the strong learner and combining the predictions of the weak learners according to their weights.

Then, a secondary model is built to try to fix the problems with the first one. This process is repeated, and further models are added, until either the entirety of the training data set is accurately predicted or the maximum number of models has been added. AdaBoost was the name of the very first boosting algorithm that was developed with the objective of binary classification in mind particularly, and it proved to be rather effective. The phrase "Adaptive Boosting," which is shortened as "AdaBoost," refers to an incredibly common type of the "boosting" methodology. This method combines a large number of "weak classifiers" into a unique "classification algorithm."

- **TF-IDF:** The Inverse Frequency Document, or IFFD frequency (also known as TF-IDF) is indeed a metric used in statistics to measure how important a word is in relation to a set of texts. Two measurements are multiplied together to get this result, the frequency with which a certain term occurs in a manuscript is the first., and the second is TI-DF of the phrase over a set of documents. It can be used for many different things, most notably in the field of computerized text analysis, and It's really useful for evaluating words in NLP-related machine learning algorithms. In order to search for documents and retrieve information, they created TF-IDF to do this. This is achieved by increasing inversely as the frequency with which a term occurs in the source manuscript, in the meanwhile diminishing inversely proportionate to the number of papers containing the word. hence, common phrases found in all texts, such as this, that, if and what, low scores despite their

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

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

$$TF - IDF(t, d) = TF(t, d) * IDF(t)$$

prevalence in the paper because they lack significance there. Contrarily, if the word "Bug" occurs multiple times in one paper but not for others, it's likely because the information contained inside is very important. For instance, if our goal is to determine the categories to which certain NPS answers belong, we may conclude that the word

"Bug" is most closely associated with the category "Reliability," given that the vast majority of replies that contain the word will be about that category.

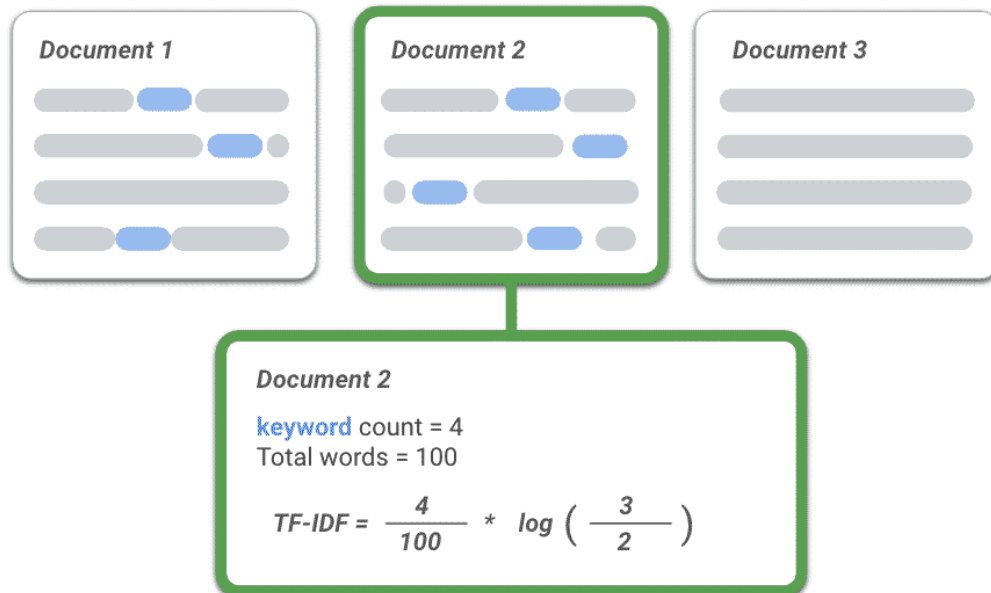


Figure 3.4.6 TF-IDF workflow

In **figure 3.4.6** it is Term frequency–inverse document frequency (TF-IDF) which is a numerical statistic that is used to reflect the importance of a word in a document or a collection of documents. It is commonly used in information retrieval and natural language processing tasks.

TF-IDF (term frequency-inverse document frequency) is a mathematical term used to represent the importance of a word in a document. The TF-IDF algorithm is used to weigh a word's importance within a document, in relation to an entire corpus of documents.

The algorithm calculates two values for each word in a document:

1. **Term Frequency (TF):** This is the number of times a word appears in a document, normalized by the total number of words in the document. It is calculated as $TF = (\text{number of occurrences of a word in a document}) / (\text{total number of words in the document})$
2. **Inverse Document Frequency (IDF):** This is a measure of how rare a word is across all the documents in the corpus. It is calculated as $IDF = \log\left(\frac{\text{total number of documents in corpus}}{\text{number of documents containing the word}}\right)$.

The final TF-IDF score for a word in a document is the product of its TF and IDF values. This results in words that are common across all documents in the corpus receiving a lower score, and words that are specific to a particular document receiving a higher score.

The basic workflow of calculating TF-IDF is as follows:

1. Tokenize the documents into individual words (also called "tokens").
2. Calculate the term frequency (TF) of each token in each document. Term frequency is the number of times a token appears in a document, normalized by the total number of tokens in the document.
3. Calculate the inverse document frequency (IDF) of each token. Inverse document frequency is a measure of how rare a token is across all documents in the collection. It is calculated as the logarithm of the total number of documents in the collection divided by the number of documents that contain the token.
4. Calculate the TF-IDF score of each token in each document by multiplying the term frequency and inverse document frequency of the token.

The resulting TF-IDF scores can be used to identify the most important tokens in each document or to compare the importance of tokens across different documents. For example, a token that occurs frequently in a single document but rarely in the collection as a whole will have a high TF-IDF score, indicating that it is important to that document.

- **BOW:** Natural Language Processing makes use of a text modeling technique called bag of words (NLP). We may refer to it as a method of feature extraction using text data if we speak in words that are more technical. Using this approach, feature extraction from papers may well be carried out in a manner which is both easy and flexible. The term "bag of words" refers to a visualization of texts that shows how often certain terms appear inside a given document. [Case in point:] We only count the number of words and pay no attention to the specifics of the grammar or the sequence of the words. The term "bag" is used to refer to a collection of words since no information about the structure or order of the contents of the text is retained. The model cares mostly about the presence or absence of certain phrases in the text, it ignores the order within which the words occur in the text. In that case, why resort to such a grab bag of words? Exactly what is wrong with a piece of writing that is simple and easy to comprehend? Due to the fact that machine learning algorithms work best when presented with structured, fixed-length inputs that are well-

defined, and the Bag-of-Words technique, which allows us to convert texts of varying lengths into a vector of constant size, the messiness and lack of structure that characterizes text is one of the most significant challenges faced when working with it. In addition, ML models focus on numeric data, as it can be broken down into finer details than textual data. To be more precise, we use the bag-of-words, or BOW, method to transform a phrase into a vector of integers.

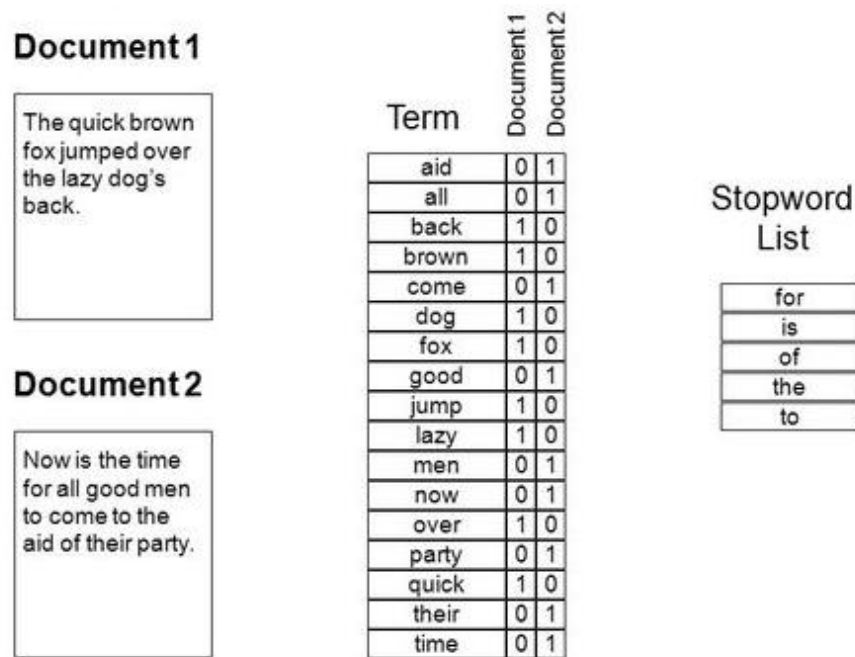


Figure 3.4.7 Bag of Words workflow

In **Figure 3.4.7** the bag of words model is a simple representation of text data used in natural language processing tasks. It represents a text document as a bag (unordered collection) of words, ignoring grammar and the order of the words.

The bag of words (BOW) algorithm is a method used to represent text data in a numerical format that can be used as input for machine learning algorithms. The workflow of the BOW algorithm includes the following steps:

1. **Tokenization:** The first step is to split the text data into individual words or tokens. This process is known as tokenization and typically involves removing punctuation and special characters, and converting the text to lowercase.
2. **Removing stop words:** The next step is to remove stop words, which are common words that do not contain any useful information, such as "a", "an", "the", "is", etc.

These words are removed to reduce the dimensionality of the data and to focus on the more meaningful words.

3. **Stemming or Lemmatization:** The next step is to reduce the words to their base form. This can be done by stemming which is the process of reducing words to their root form. Or by Lemmatization which is the process of reducing words to their base form by considering the context and the part of speech.
4. **Creating a vocabulary:** The next step is to create a vocabulary of all the unique words that appear in the text. This vocabulary will be used to create a numerical representation of the text data.
5. **Vectorization:** Once the vocabulary is created, the next step is to convert each document into a numerical vector representation. This is done by counting the number of occurrences of each word in the vocabulary for each document and creating a vector with the resulting counts. This vector will be used as input for the machine learning algorithm.
6. **Weighting:** To assign a weight to each word in the vector, the words can be weighted using a technique like TF-IDF (term frequency-inverse document frequency), which assigns a weight to each word based on its importance to the document.
7. **Dimensionality reduction:** Once the vectors are created, it may be necessary to reduce the dimensionality of the data. This can be done using techniques such as principal component analysis (PCA) or singular value decomposition (SVD).

The resulting feature vectors can be used as input to machine learning algorithms for tasks such as classification or clustering. The bag of words model is simple but effective, and it has been widely used in natural language processing tasks.

3.5 Implementation:

For implementation we have divided our work in main five part. These parts are steps we need for our project to be successful

- Data Collection
- Data pre process
- Data process
- Algorithm implementation
- Result discussion

For data collection we used free tools for data scraping but, in the end, we used Face pager tool. Then we started to work on the data preprocess. Here we removed all the unnecessary part of our data like removing stop words, removing the links, removing unnecessary symbols and etc. We extracted the frequency of stop words when the tokenization was done.

We started our algorithm implementation where we had to work on our code for the desired algorithm. We used total seven algorithm and looked for the accuracy.

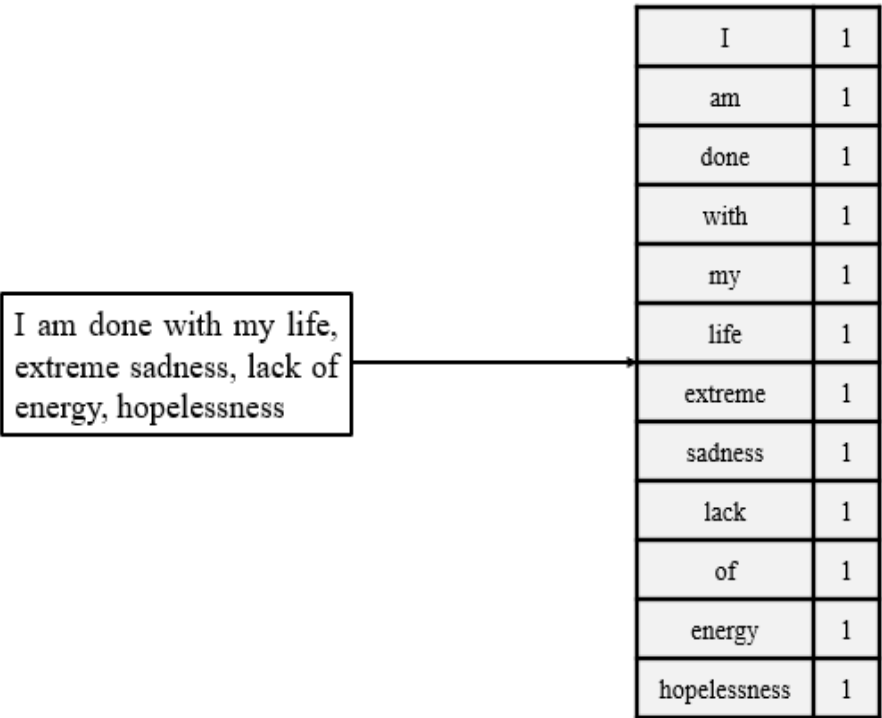


Figure 3.5.1 Data training method

As shown in the **Figure 3.5.1** it shows the method how the Bag of Words algorithm works. This is how BOW takes a sentence and splits it as individual words. After that it is assigned as a token. After this the algorithm does stemming, lemmatization, create vocabulary, weighting and dimensionality reduction.

When we were done with the algorithm, we looked at the accuracy. Comparing the accuracy, we saw which will be better for our work. But there are other parts which we wanted to work on like implementation of NLP. For this reason, we have picked algorithm like TF-IDF and BOW which we used for our detection of machine learning.

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 Introduction

In order to predict conceivable consequences, we have experimented with a broad range of information sources and algorithms. Our experimental outcomes are presented and discussed throughout this chapter. Detailed experiment result is discussed on chapter 4.2, descriptive analysis is covered in chapter 4.3 and in the end total summary of the whole experiment is talked in the 4.4 chapter.

4.2 Experimental Result

Challenges while implementing different algorithm for depression detection was normal. We used different methods for the process for this reason. We tried different methods and studied different methods for the best method for the experiment. We tried different ways to improve our work and its outcome

We used available python libraries, content categorizing techniques and dictionary. While doing so we found out link between dialect usage and discouragement. There were more frequent terms in the depressing posts. As our previous findings we saw that self-explanation, indignation, hostile mood, unease and self-destructive talks, thoughts or expressions were common as a dialect indication of melancholy.

We used two different datasets where one of the datasets was collected from the internet another one, we collected ourself.

4.3 Descriptive Analysis

Depending on our classification algorithms that we used we got different outcomes. We used XGB Classifier, Random Forest Classifier, Logistic Regression, SVC, Ada Boost Classifier and We Applied It To: Instances per Term Counterfactual Document Frequency and BOW from NLP. For the purpose of making things easy for our algorithm we labeled our own data. Every algorithm worked on the same dataset which consisted of both the pre available data and our own dataset that we have acquired form the internet. After we were done with the dataset process, we used python and its prebuilt libraries to check accuracy of the algorithms

Table 4.3.1 Algorithm accuracy

Classifier	Accuracy
XGB Classifier	94.12 %
Random Forest Classifier	93.55 %
Logistic Regression	94.40 %
SVC	93.98 %
Ada Boost Classifier	94.00 %
TF-IDF	83.79 %
BOW	87.22 %
Naive Bayes	90.00 %

In **Table 4.3.1** we have shown our algorithms accuracy. As shown, there are algorithms from machine learning also from natural language process. This algorithm shows that only two NLP algorithm have lowest accuracy where they are less than 90%. But normal most of the machine learning algorithm are more than 90%. Also, all the NLP are the lowest among all the algorithm. Even though only two of them are here.

This part shows the performance of different classifiers. For this whole process CoLab and Jupyter was used which are open-source software. Total of seven classifiers were used which are XGB Classifier, Random Forest Classifier, Logistic Regression, SVC, Ada Boost Classifier, TF-IDF and BOW.

4.4 Summery:

Data is one of the most important aspect of one experiment. Same type of experiment can have very different results depending on the data provided. As we used mixture of two dataset, we were sure there would be difference between results that others have acquired using one of our experiments datasets which was available online before. Because we used more data it is possible to have better or worse accuracy. Our goal for this research from the start is to detect depression from the data we gathered from social media.

We achieved our goal by working with different methods of ML. There is total 7 algorithm we have used for this total work. There is different thing we had to look for a while before starting our work. When we choose our algorithm, we did start our work on it. Then we got the accuracy of all these algorithms. As mentioned, before we have labeled our data on our own. We say frequency of word use in same class of data.

We have also found out there if there was slight error on the data the algorithm won't be able to give proper prediction. It will give false positive or false negative depending on the data pattern. The problems like missing a letter or miss placing a sentence than usual. Our most accuracy was in Logistic Regression where we got 94.4% of accuracy.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society:

The use of machine learning techniques for detecting human depression using social media data, as it has the potential to have a significant impact on society. Here are a few potential ways in which these techniques could impact society:

1. **Improved mental health:** One of the most significant potential impacts of machine learning techniques for detecting depression is the improvement of mental health outcomes for individuals who are identified as being at risk for depression. By catching the condition early, it may be possible to intervene and provide support and resources that could help to prevent the condition from worsening. This could ultimately lead to better outcomes for those affected and a reduction in the overall burden of depression on society.
2. **Reduced stigma:** The use of machine learning techniques for detecting depression could also help to reduce the stigma associated with mental health conditions. By providing a more objective and unbiased way of identifying individuals who may be struggling with mental health issues, machine learning techniques could help to reduce the perception that mental health conditions are something to be ashamed of or hidden. This could encourage more people to seek help and support, which could ultimately lead to better outcomes for those affected.
3. **Economic benefits:** The use of machine learning techniques for detecting depression could also have economic benefits for society. For example, it could help to reduce the overall burden of depression on the healthcare system, as early intervention could prevent the condition from becoming more severe and requiring more intensive and costly treatment. It could also lead to improved productivity and economic outcomes,

as individuals who are experiencing depression may be less able to work or contribute to society.

4. **Social benefits:** In addition to the potential economic benefits, the use of machine learning techniques for detecting depression could also have social benefits for society. For example, it could help to improve relationships and social connections, as individuals who are experiencing depression may be more isolated and have fewer social connections. By identifying and addressing depression, it may be possible to improve social connections and support networks, which could have a positive impact on the overall well-being of individuals and society.

Overall, the use of machine learning techniques for detecting human depression using social media data has the potential to have a significant impact on society. By improving mental health outcomes, reducing stigma, and providing economic and social benefits, these techniques could have a positive impact on the well-being and quality of life of those affected and on society as a whole.

5.2 Impact on Environment:

The use of machine learning techniques for detecting human depression using social media data is not likely to have a direct impact on the environment. However, it is possible that the development and use of these techniques could have indirect environmental impacts through their impact on resource consumption and emissions. Here are a few potential ways in which the use of these techniques could impact the environment:

1. **Energy consumption:** The development and use of machine learning techniques can be resource-intensive, particularly if they require a lot of computational power. This could result in increased energy consumption, which could have environmental impacts through the production of greenhouse gas emissions. To minimize these

impacts, it is important to consider the energy efficiency of the algorithms and hardware being used, and to take steps to minimize energy consumption where possible.

2. **Data storage:** Machine learning algorithms require large amounts of data in order to be trained and tested. This data needs to be stored, which can also have an impact on the environment through the use of resources such as energy and materials. To minimize these impacts, it is important to consider the environmental impacts of data storage and to take steps to minimize resource consumption where possible.
3. **Transportation:** If the machine learning techniques are being used in a distributed manner, with data and algorithms being accessed from multiple locations around the world, it is possible that the use of these techniques could result in increased transportation emissions. To minimize these impacts, it is important to consider the environmental impacts of transportation and to take steps to minimize emissions where possible.

Overall, the use of machine learning techniques for detecting human depression using social media data is not likely to have a direct impact on the environment. However, it is important to consider the potential indirect environmental impacts of the development and use of these techniques, and to take steps to minimize these impacts where possible. This could include reducing resource consumption and emissions through energy-efficient algorithms and hardware, minimizing resource consumption in data storage, and minimizing transportation emissions where possible.

5.3 Ethical Aspects:

The use of machine learning techniques for detecting human depression using social media data raises a number of ethical concerns. Here are a few of the key ethical aspects to consider:

1. **Privacy:** One major ethical concern is the issue of privacy. Social media data is personal and sensitive, and it is important to ensure that it is handled responsibly and in accordance with relevant laws and regulations. This includes protecting the data from unauthorized access or misuse, and obtaining appropriate consent from individuals before using their data. It is also important to consider the potential impacts on an individual's privacy if they are identified as being at risk for depression based on their social media data.
2. **Accuracy:** Another important ethical consideration is the accuracy of the machine learning algorithms being used to detect depression. If the algorithms are not accurately identifying individuals who are at risk for depression, it could have serious consequences, including wrongly labeling individuals as being at risk and potentially causing them unnecessary distress. On the other hand, if the algorithms are too conservative and do not identify enough individuals as being at risk, it could result in missed opportunities for early intervention and support. It is important to ensure that the algorithms are thoroughly tested and validated to ensure their accuracy.
3. **Bias:** Another ethical concern is the potential for bias in the machine learning algorithms. For example, if the algorithms are trained on data that is not representative of the population, it could result in biased predictions. It is important to ensure that the data used to train the algorithms is diverse and representative of the population, in order to reduce the potential for bias.
4. **Access:** Another ethical issue to consider is access to the machine learning techniques for detecting depression. If the techniques are only available to certain groups or

individuals, it could result in inequities in the ability to identify and address depression. It is important to ensure that the techniques are widely available and accessible to all who may benefit from them.

5. **Responsibility:** Finally, there is the question of who is responsible for the decisions made based on the output of the machine learning algorithms. It is important to consider who is responsible for ensuring that the algorithms are being used ethically and appropriately, and for ensuring that individuals who are identified as being at risk for depression are connected with the appropriate resources and support.

Overall, the use of machine learning techniques for detecting human depression using social media data raises a number of ethical concerns that must be carefully considered and addressed. It is important to ensure that the techniques are developed and used in an ethical and responsible manner, in order to maximize their potential benefits while minimizing any negative impacts.

5.4 Sustainability Plan:

A sustainability plan for the use of machine learning techniques for detecting human depression using social media data should consider a number of factors in order to ensure that the use of these techniques is sustainable over the long term. Here are a few key elements that should be included in a sustainability plan:

1. **Data privacy and security:** Ensuring the privacy and security of social media data is critical to the sustainability of the machine learning techniques for detecting depression. This includes implementing appropriate safeguards to protect the data from unauthorized access or misuse, and obtaining appropriate consent from individuals before using their data. It is also important to consider the potential impacts on an individual's privacy if they are identified as being at risk for depression based on their social media data.

2. **Accuracy and reliability:** Ensuring the accuracy and reliability of the machine learning algorithms is also essential for the sustainability of the techniques. This includes thoroughly testing and validating the algorithms to ensure their accuracy, and regularly reviewing and updating the algorithms as needed.

3. **Bias:** Reducing the potential for bias in the machine learning algorithms is also important for the sustainability of the techniques. This includes ensuring that the data used to train the algorithms is diverse and representative of the population, and regularly reviewing and adjusting the algorithms as needed to reduce the potential for bias.

4. **Access:** Ensuring that the machine learning techniques are widely available and accessible to all who may benefit from them is also important for their sustainability. This could include developing strategies to make the techniques more affordable or accessible in areas with limited resources, or partnering with organizations or individuals who can help to distribute the techniques more widely.

5. **Responsibility:** Clearly defining the roles and responsibilities of those involved in the development and use of the machine learning techniques is also critical for their sustainability. This could include establishing clear guidelines for the ethical use of the techniques, as well as defining the roles and responsibilities of those who are responsible for ensuring that the techniques are being used ethically and appropriately.

Overall, a sustainability plan for the use of machine learning techniques for detecting human depression using social media data should consider a range of factors in order to ensure that the use of these techniques is sustainable over the long term. By addressing issues such as data privacy, accuracy and reliability, bias, access, and responsibility, it is possible to ensure that the techniques are developed and used in a responsible and ethical manner that maximizes their potential benefits while minimizing any negative impacts.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study:

This study has given us many insights in this topic. Still mental health is a sensitive topic in this time. This is the reason it is not so wide spread well aware news or information. That is the reason we used ML to recognize the hidden information in the data. These patterns are what ML uses to do predictions.

Already we have mentioned that while doing our study we have used data from Twitter and the reason behind this is we needed a social media where people are comfortable or frequently share their emotion. Twitter met our expectation. This data helped our algorithms to train and learn sentence pattern and which was used for depression detection. There are few problems which were addressed before in the beginning.

We were able to achieve our goal which we were working for. Different algorithm gave us different kind of outcome. We discussed about it in details in next part.

6.2 Conclusion:

Our research result and used methods are great according to our work. After concluding this study, we think and hope it will enhance the study of this field. This study will give us many ways for expanding our work. We have found out few errors during our work. We saw new ways which this study might be possible to expand further. It will allow us to fix the errors or other problems that we faced during our work on this project. We are also thinking of ways to combining algorithms and making more efficient ways to fix our problems of this study in the future. This study will allow to add more knowledge about our field of study. We hope that it will add more to progress of human mental health development and add new ways to help technology that helps human mental health. We have used both practical and psychological information for better result of our study. We hope to propose a new system for depression detection depending on this study.

6.3 Possible impacts:

We believe that there are different effects of our work in different places. For this reason, we made sure our work was properly done. We wanted to make sure our work be genuine. That is why we used our own kind of dataset. Our work can be used in different fields like physiological field or computer science field. It would be a good asset for the society because it can help to control depression among the general society. It also can make our life easy and save lives. It can both help regular people understanding the situation if this work meets its's future work plan. Also, it can help spread out information about the hidden dangers of depression.

6.4 Implication of further study:

There are many possibilities for further study in this field also in our own work. We found out different ways to make our work better. As we mentioned we have also found out some errors and these errors are way for making our study better. As we said before we have got false negatives during our predictions. It was not supposed to happened. It possibly can create fatal error if we face this in real life implementation. We have plans for fixing false negatives. That means training our algorithm where it can work with small errors like words missing letters or misspelling.

We also other goals in our mind. We would like to make our prediction is better. That way we can use algorithm to work with an application where people can use our work to predict their input sentences type. We also would like to make prediction from social media data where we can predict what kind of depression one person is suffering.

We think we can use this kind of work to suggest people about their situation and what they can do to get better. We might also be able to help people without sharing their data to the public.

Reference

- [1] Trotzek, M., Koitka, S., & Friedrich, C. M. (2018). Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences. arXiv. doi.org/10.1109/TKDE.2018.2885515
- [2] M. M. Tadesse, H. Lin, B. Xu and L. Yang, "Detection of Depression-Related Posts in Reddit Social Media Forum," in IEEE Access, vol. 7, pp. 44883-44893, 2019, doi: 10.1109/ACCESS.2019.2909180.
- [3] N. A. Asad, M. A. Mahmud Pranto, S. Afreen and M. M. Islam, "Depression Detection by Analyzing Social Media Posts of User," 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), 2019, pp. 13-17, doi: 10.1109/SPICSCON48833.2019.9065101.
- [4] Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science*, 167, 1258-1267.
- [5] Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." 2014 science and information conference. IEEE, 2014.
- [6] Deshpande, M., & Rao, V. (2017, December). Depression detection using emotion artificial intelligence. In 2017 international conference on intelligent sustainable systems (iciss) (pp. 858-862). IEEE.
- [7] Soares Passos, L. M., Murphy, C., Zhen Chen, R., Gonçalves de Santana, M., & Soares Passos, G. (2020, February). The prevalence of anxiety and depression symptoms among Brazilian computer science students. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 316-322).
- [8] Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, 135, 104499.
- [9] Govindasamy, K. A., & Palanichamy, N. (2021, May). Depression detection using machine learning techniques on twitter data. In 2021 5th international conference on intelligent computing and control systems (ICICCS) (pp. 960-966). IEEE.

- [10] Uddin, M. Z., Dysthe, K. K., Følstad, A., & Brandtzaeg, P. B. (2022). Deep learning for prediction of depressive symptoms in a large textual dataset. *Neural Computing and Applications*, 34(1), 721-744.
- [11] Arelan Asraf, Teddy Surya Gunawan, Bob Subhan Riza, Edy Victor Harynto, Zuriati Janin. 2020. "On the review of image and video-based depression detection using machine learning." 2502-4752.
- [12] Marcel Trotszek, Sven Kotika, Christoph M Friedrich, Member, IEE. March 2020. "Utilizing Neural Networks and Linguistic Metadata For Early Detection Of Depression Indications in Text Sequences." *IEEE* 32: 588-601.
- [13] k. Sudhan, S. Sreemathi, B. Nathiya, D. Rahinipriya. 2020. "Depression Detection Using Machine Learning." *IJRAD* 2581-4451.
- [14] Irene LI, Yixin Li, Sergio Alvarez-Napagao, Dario Garcia-Gasulla, Toyotaro Suzumura. 2020. "What are we depressed About When We Talk About Covid-19 Mental Health Analysis on Tweets Using Natural Language Processing." *Springer* 12498: 358- 370.
- [15] Michael M Tadesse, HONGFEI LIN, BO XU, LIANG YANG. 2019. "Detection of Depression-Related Posts in Reddit Social Media Forum ." *IEEE* 7: 44883-44893.
- [16] Mandar Deshpande, Vignesh rao. 2017. "Depression Detection Using Emotion Artificial Intelligence." *IEEE*.
- [17] David William, Derwin Suhartono. 2021. "Text-based Depression Detection on social Media posts: A literature review." *Science direct* 179: 582-589.
- [18] Md Rafiqul Islam, Mahmud Ashad Kabir, ashir Ahmed, abu Raihanm. kamal, Hua Wang. 2018. *Springer Nature Switzerland* 1-12.
- [19] Hatoon AlSagri, Mourad Yakhlef. n.d. "Machine learning-based Approach For Depression Detection on Twitter Using Content and Activity Features."
- [20] David William, Derwin Suhartono. 2021. "Text-based Depression Detection on social Media posts: A literature review." *Science direct* 179: 582-589.

MACHINE LEARNING TECHNIQUES FOR DETECTING HUMAN DEPRESSION USING SOCIAL MEDIA DATA

ORIGINALITY REPORT

12%

SIMILARITY INDEX

9%

INTERNET SOURCES

4%

PUBLICATIONS

8%

STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	Submitted to Daffodil International University Student Paper	2%
3	"Early Detection of Mental Health Disorders by Social Media Monitoring", Springer Science and Business Media LLC, 2022 Publication	1%
4	Submitted to Jacksonville University Student Paper	<1%
5	Submitted to University of Sunderland Student Paper	<1%
6	newscatcherapi.com Internet Source	<1%
7	library.samdu.uz Internet Source	<1%
8	www.researchgate.net Internet Source	<1%

9	github.com Internet Source	<1 %
10	Submitted to Xiamen University Student Paper	<1 %
11	Submitted to Manchester Metropolitan University Student Paper	<1 %
12	Submitted to Kingston University Student Paper	<1 %
13	www.irjmets.com Internet Source	<1 %
14	Mohammed Kasri, Marouane Birjali, Abderrahim Beni-Hssane. "A comparison of features extraction methods for Arabic sentiment analysis", Proceedings of the 4th International Conference on Big Data and Internet of Things, 2019 Publication	<1 %
15	"17th International Conference on Information Technology–New Generations (ITNG 2020)", Springer Science and Business Media LLC, 2020 Publication	<1 %
16	"Artificial Intelligence in Healthcare", Springer Science and Business Media LLC, 2022 Publication	<1 %

17	Submitted to University of Birmingham Student Paper	<1 %
18	www.annalindsey.com Internet Source	<1 %
19	Submitted to The Robert Gordon University Student Paper	<1 %
20	Submitted to University of Southampton Student Paper	<1 %
21	Submitted to University of Greenwich Student Paper	<1 %
22	Submitted to Virginia Polytechnic Institute and State University Student Paper	<1 %
23	wap.iol.co.za Internet Source	<1 %
24	"Computational Vision and Bio-Inspired Computing", Springer Science and Business Media LLC, 2022 Publication	<1 %
25	Blaber, Amanda, Harris, Graham. "Ebook: Assessment Skills for Paramedics, 3e", Ebook: Assessment Skills for Paramedics, 3e, 2021 Publication	<1 %