Masters of Science in Computer Science and Engineering

Human Depression Detection from Social Network Data

by

Md Rafiqul Islam

Student No. 154603

Supervisor

Dr. Abu Raihan Mostofa Kamal

Associate Professor, Department of CSE, IUT

Department of Computer Science and Engineering (CSE)

Islamic University of Technology (IUT)

Organisation of Islamic Cooperation (OIC)

Gazipur- 1704, Bangladesh

June, 2018

Human Depression Detection from Social Network Data

by

Md Rafiqul Islam

Student No. 154603

Supervised by

Dr. Abu Raihan Mostofa Kamal

Associate Professor, Department of CSE, IUT

A thesis submitted to the Department of Computer Science and Engineering (CSE) is partial fulfilment of the requirements for the award of the Degree of Masters of Science in Computer Science and Engineering (M.Sc. Engg. CSE)

Department of Computer Science and Engineering (CSE)

Islamic University of Technology (IUT)

Organisation of Islamic Cooperation (OIC)

Gazipur- 1704, Bangladesh

June, 2018

Recommendations of the board of examiners

The thesis titled "**Human Depression Detection from Social Network Data"** Submitted by Md Rafiqul Islam, Student No. 154603 of Academic Year 2015-2016, has been found satisfactory and accepted as partial fulfilment of the requirement for the Degree of Masters of Science in Computer Science and Engineering (M.Sc. Engg. CSE) on June 06, 2018.

1.	....	
	Dr. Abu Raihan Mostofa Kamal Associate Professor, Department of Computer Science and Engineering, IUT, Board Bazar, Gazipur-1704, Dhaka, Bangladesh.	Chairman (Supervisor)
2.	....	
	Prof. Dr. Muhammad Mahbub Alam Head, Department of Computer Science and Engineering, IUT, Board Bazar, Gazipur-1704, Dhaka, Bangladesh	Member
3.	....	
	Dr. Md. Kamrul Hasan Associate Professor, Department of Computer Science and Engineering, IUT, Board Bazar, Gazipur-1704, Dhaka, Bangladesh	Member
4.	....	
	Prof. Dr. Abu Sayed Md. Mostafizur Rahaman Professor, Department of Computer Science and Engineering Jahangirnagar University, Dhaka, Bangladesh	Member (External)

Declaration of the Candidate:

This is to certify that the work presented in this thesis titled "**Human Depression Detection from Social Network Data**" is done by the candidate under the Supervision of Dr. Abu Raihan Mostofa Kamal, Associate Professor, in the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh.

It is also declared that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

....

(Signature of the Candidate)

Md Rafiqul Islam

Network and Data Analysis Group (NDAG)

Student No: 154603

Session: 2015-2016

June, 2018

....

(Signature of the Supervisor)

Dr. Abu Raihan Mostofa Kamal

Associate Professor

Department of Computer Science and Engineering

Islamic University of Technology (IUT)

Gazipur- 1704

June, 2018

Dedication

"This dissertation is dedicated to my parents and teachers for all their continuous support, *love and inspiration"*

Abstract

Social networks have been developed as a great point for its users to communicate with their interested friends and share their opinions, photos, and videos reflecting their mood, feelings and sentiments. This creates an opportunity to analyze social network data for user's feelings and sentiments to investigate their mood and attitude when they are communicating via these online tools. Existing literature reflects on the use of various Application Programming Interfaces (API) such as Graph API, REST API and Streaming API to collect data from social networks and used for depression detection. Although diagnosis of depression using social networks data has picked an established position globally, there are several dimensions that are yet to be detected. In this study, we aim to perform depression analysis on Facebook data collected from an online public source by using machine learning techniques. We have evaluated the efficiency of machine learning techniques using a set of various psycholinguistic and textual features. The result shows that in different experiments Decision Tree (DT) gives the highest accuracy than other ML approaches to find the depression. It is anticipated that the analysis reported in this study would contribute in developing any electronic disease management system both for communities and healthcare professionals groups.

Keyword: Social network, Emotions, depression, sentiment analysis.

Acknowledgements

First of all, I would like to thank to our God for keeping me in sound mind and health during the work and giving me the ability to work hard successfully.

Secondly, I would like to thank and express my deepest appreciation and sincere gratitude to my supervisor Dr. Abu Raihan Mostofa Kamal, who inspired me to take pride in my research. His enthusiasm for research efforts surely will have a significant effect on my future research. His advice throughout this process kept me focused on the right direction. I am really thankful to him for sharing his ideas and necessary insights.

It was my pleasure to get the cooperation from the Head of the Department, Professor Dr. Muhammad Mahbub Alam during various phases of the work. I am grateful to him for his encouragements and valuable advice. In addition, I would like to especially thank to Dr. Md. Kamrul Hasan, Dr. Hasanul Kabir and Prof. Dr. Abu Sayed Md. Mostafizur Rahaman for their kind appreciation.

I would like to thank the jury members of my thesis committee for the many interesting comments and criticism that helped improve this manuscript. Lastly, I would acknowledge and recall all of the teachers, friends and well-wishers in the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT). They all supported me really a lot with their reviews, comments and suggestions during the research work.

Md Rafiqul Islam

June 2018

Preface

This master's thesis is outlined based on enhancing knowledge discovery (KD) about the depression from social network data using data mining integrated environment. The research was carried out in the Network and Data Analysis Group (NDAG) under the Department of Computer Science and Engineering (CSE) of Islamic University of Technology (IUT), OIC, Dhaka. It includes six chapters which are briefed as follows:

Chapter-1

Chapter 1 provides a discussion on the importance of the work that has been done and why the current topic is selected for the Master's Thesis along with an informative introduction.

Chapter-2

Chapter 2 discusses the background and related works to understand the depression concepts from social network data.

Chapter-3

Chapter 3 discusses the proposed system. It describes the proposed depression detection system with the respective of four heuristics.

Chapter-4

Chapter 4 describes about the experimental analysis. It shows the precision recall, Quantitative and Empirical evaluation.

Chapter-5

Chapter 5 provides the overall discussions of the work.

Chapter-6

Chapter 6 provides conclusion based on the KD concept using data mining environment and gives ideas for the future scope.

Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interfaces
KD	Knowledge Discovery
SNS	Social Networking Site
HSE	Health and Safety Executive
GMHAT	Global Mental Health Assessment Tool
DSR	Design Science Research
ML	Machine Learning
LIWC	Linguistic Inquiry and Word Count
IT	Information Technology
MD	Mental Disorder
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
PD	Personality Disorder

Chapter 1

Introduction

In this chapter, we first present an overview of our thesis that includes the significance of the problem and the problem statement in details. Besides, we also discuss about the different research challenges what we are going to face in the whole scenario. After that, we present our thesis objectives and contributions. The chapter ends with a short description of the organization of this thesis.

1.1 Motivation

Social networks have been an upcoming trend over the last few years and they have gained an established position in the World. The proliferations of internet and communication technologies, especially the online social networks have rejuvenated how people interact and communicate with each other electronically [1]. The applications such as Facebook, Twitter, Instagram and alike not only host the written and multimedia contents but also offer their users to express their feelings, emotions and sentiments about a topic, subject or an issue online. Such data are embedded with useful information about persons' well-being and life-style [2]. Most of the social media or organization's media sites provide publicly available large dataset that contain relevant attributes of professionals [3]. This is great for users of social networking site to openly and freely contribute and respond to any topic online; on the other hand, it creates opportunities for people working in the health sector to get insight of what might be happening at mental state of someone who reacted to a topic in a specific manner. In order to provide such insight, machine learning techniques could potentially offer some unique features that can assist in examining the unique patterns hidden in online communication and process them to reveal the mental state (such as *'happiness'*, *'sadness'*, *'anger'*, *'anxiety'*, depression) among social networks' users [4, 5]. Moreover, there is growing body of literature addressing the role of social networks on the structure of social relationships such as breakup relationship, mental illness (*'depression'*, *'anxiety'*, *'bipolar'* etc.), smoking and drinking relapse, sexual harassment and for suicide ideation.

Mental health has been an important determinant of communities' well-being. This may influence not only by individual attributes, but also from different social circumstances and organisational environment in which people work and live. Besides, mental health condition is one of the main causes of occupational dis-satisfaction and sickness in the far-reaching consequences of the workplace [6, 7]. So, the main idea of this master's thesis is to analyse social network data to detect any factors that may reflect the mental health issues (depression, anxiety, bipolar etc) of relevant social network users. It is interesting to discover depressive post/comments and to observe these data from social network users [8]. Especially, Facebook offers an exciting potential for the study of mental health issues because Facebook users consists of their individual followers, while, on the other hand, their posts/comments offer a powerful content network.

In the light of the brief discussion of depression detection, the following six connected components addressed in this thesis are:

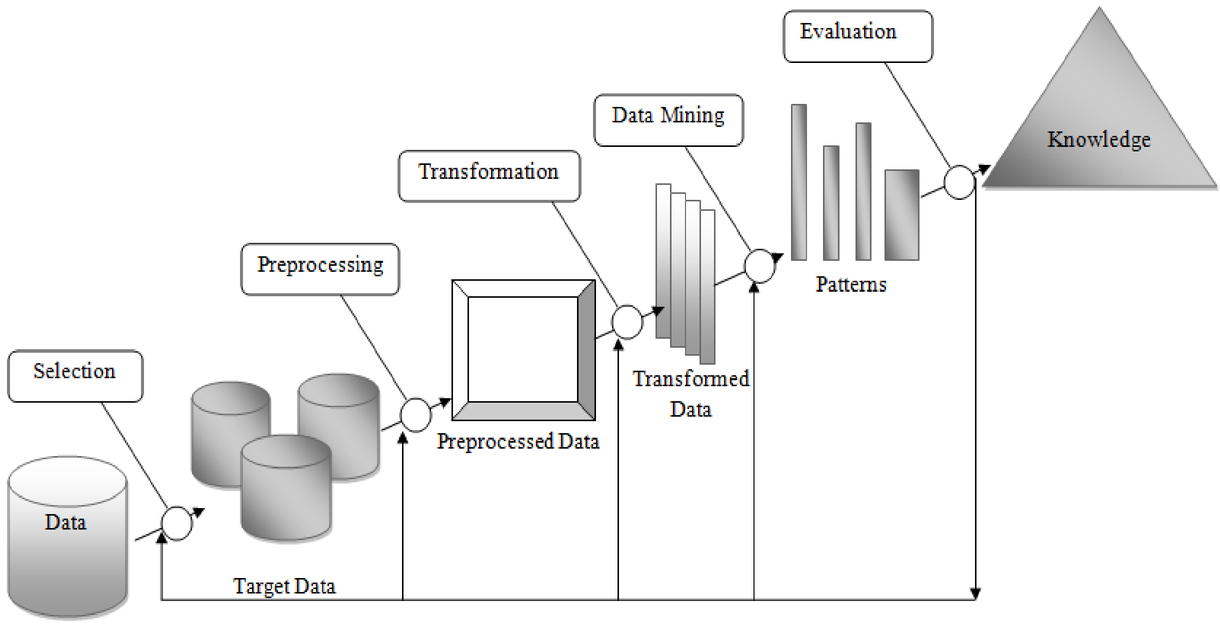

Figure 1.1: The complete system overview of our project

1.2 Objective and Research Challenges

One of the secondary objectives of this thesis is to present an overview of current mental health issues and social network analysis in general. Furthermore, our aim is to provide the understanding around problems and decision support approach development through the use of machine learning (ML) techniques for mental health problems and possible data-driven solutions. The main objectives of this thesis are as follows:

- Knowledge discovery (KD) about the depression in around the world from Facebook data.
- Demonstration of Facebook data suitability for depression detection based on its key symptoms.
- Supervised machine learning algorithm implement for depression related Facebook data classification.
- Interactive visualization development for exploratory analysis.

Considering the key objective of this study, the following are subsequent research challenges are addressed.

- Define what depression is and what are the common factors contributing toward depression.
- What are the features to look for depression detection in Facebook post and comments and how to extract these features?
- What are the most influential machine learning techniques for detecting depression from Facebook post and comments?

1.3 Contribution to Knowledge and Statement of Significance

In this thesis work, we have proposed a Depression Detection (DD) system using the Design Science Research (DSR) sequences. The main contributions of this thesis are summarized as follows:

- We synthesised the literature on various emotion detection techniques to detect depression.
- We have created a new dataset and the data is captured in real-time and exact location.
- We extracted the psycholinguistic and textual features for our specific research problem and elaborate on the lesson learned from using each type.
- Emotional process, temporal process and linguistic style are utilized to extract the psycholinguistic information's.
- To train the feature, we used Matlab 2016b.
- Proper training method for the classification of each feature is done using Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT) and Ensemble.

- The detection performance is evaluated with different Machine Learning (ML) sub classifier with two types of Datasets.
- Our experimental outcomes served as an alternative source of knowledge and attempt to fill the gaps of the traditional health reports and surveys.

1.4 Thesis Outline

The remainder of the thesis is organized as follows: Section 2 presents background of the proposed study while the section 3 describes details of the proposed system. The experimental analysis is presented in Section 4 and its discussion in section 5. Finally, Section 6 provides a conclusion and future directions of the study.

Chapter 2

Background Study

In this chapter, we first present a discussion on Social Network Analysis and Mental Health Illness of Facebook users which is followed by a review on different depression appearance-based methods. Finally, we end the literature review with a description of the depression analysis and the relationship between Facebook and Depression in this study.

2.1 Social Network Analysis

Over the past several years, the number of social networking site (SNS) users has escalated significantly, rising from 8% of adult Internet users in 2005 to 37% in 2008 [9]. These SNSs allow users to create profiles showcasing demographic information, personal thoughts, and a wide range of interests [10]. Additionally, these sites offer users the opportunity to maintain online connections with online friends, as well as to share photographs, videos, and stories. SixDegrees.com, launched in 1997, was the first SNS, and allowed users to create personalized profiles, maintain lists of friends, and browse these friend lists. Interest in this type of Internet interaction led to the creation of various sites such as LiveJournal, LinkedIn, MySpace, and Facebook. In 2010, 72% of online 18-29 year olds used SNSs, and the rate of use among teens was similar. Motivations for SNS use differ between individuals and are related to factors such as frequency of site visitation [11]. For example, those who engage in more frequent site visitation report being more social in their online communication, as opposed to those users who visit SNS sites infrequently. Additionally, SNS preference is associated with the user's level of attachment to that particular site [12].

In 2004, a Harvard undergraduate named Mark Zuckerberg founded a SNS he called "The facebook" as a way for his Harvard University classmates to communicate on the web [13]. In the past seven years, Facebook has evolved into a globally-utilized site with over 800,000,000 users that post photographs, share feelings, and update relationship statuses [14]. The average user of the SNS creates and shares 90 pieces of personal information each month [15]. Features of Facebook commonly used are Friend Requests (the ability to add a connection with another Facebook user), the Wall (a public posting site for each individual user where friends may share public messages and exchange comments), and Photos (publicly sharing a single photo or a complete online album with the option to “tag” users as a means of identification). Individual users differ in regards to the manner in which they use Facebook [16, 17]. For instance, there is still ambiguity over what type of conduct and communication is appropriate with this new form of social media, in regards to the degree to which a person chooses to share personal details on a public forum [18].

Nowadays, in the world, social media technologies allow people to express their views freely online in which they leave valuable content that can enable opportunities of creating new insights or support information [19]. As a result, growing number of researchers are using social media data to recruit for a range of online health, medical, and psychosocial studies. Choudhury et al. [20] investigated the possibility to utilize online networking to identify and analyze any sign of significant depression issue in people. They took data from twitter. Through their twitter postings, they quantified behavioral credits identifying with social engagement, feeling, dialect and semantic styles, sense of the self-system, and notices of antidepressant medications. Thus, while there have been some studies conducted on the relationship between Facebook and depression, there is still a large gap. Research needs to be done that examines the direct relationship between Facebook use and depression. Additionally, studies need to be conducted to investigate the factors that moderate this relationship. For instance, researchers attempted to explore the impact of different Facebook advertisement content for the same study on recruitment rate, engagement, and participant characteristics using the five Facebook advertisement sets (“resilience”, “happiness”, “strength”, “mental fitness”, and “mental health”). However, it is important to carefully consider the online content of target population for the assessment of generalisability [21]. Joyce et al. [22] identified that depression is a most common factor of mental disorders in most developed countries while carrying out a systematic meta-review examining the effectiveness of the interventions of workplace mental health. The study by Joyce et al. [22] concluded that their findings demonstrate there are empirically supported interventions that workplaces can utilize systematic solutions to aid in the prevention of common mental illness.

2.2 Depression Analysis

A factor that is thought to influence individual differences in normal cognitive aging is the presence of depressive symptoms. Recently there has been a growing awareness that mood disorders may be associated with a distinct pattern of cognitive impairment [23]. For example, an inverse association between depression and cognitive function has been reported in clinical studies in both younger and elderly samples, with cognitive deficits found on tests of attention [24], memory functions [25], psychomotor functions and executive functions [26].

Recently, Chodosh, Kado, Seeman and Karlamangla [27] have investigated the hypothesis that depression may even lead to a faster rate of cognitive decline. They investigated the association between depressive symptoms and longitudinal cognitive changes in older adults who were high-functioning at baseline and found that a higher number of baseline depressive symptoms were strongly associated with greater seven-year decline in cognitive performance.
They therefore concluded that depressive symptoms independently predict cognitive decline in previously high-functioning older persons. Sachs-Ericsson, Joiner, Plant and Blazer [28], in a longitudinal study of community-dwelling elderly persons, found that depressive symptoms were associated with subsequent cognitive decline, even after controlling for baseline cognitive status and demographic and physical functioning variables. This was also found by Chi en Chou [29] in a study with Hong Kong Chinese older adults.

In order to study the relationship between depression and dementia, Wee et al. [30] and Haberler et al. [31] investigated whether depressed elderly individuals with normal baseline cognition were at increased risk of cognitive decline and Alzheimer's disease. They found that depression increased the risk of Alzheimer's disease and cognitive decline, but only among people with higher levels of education. Aldarwish et al. [32] stated that depression could be a serious risk factor for dementia and cognitive decline, and offers a few hypotheses. One hypothesis is that depression could be a possible prodrome of dementia, which is supported by studies of patients who are initially diagnosed with a depression and progress to dementia. Choudhury et al. [33] stated that social media as a measurement tool for depression detection. They examined that Facebook is progressively researched as methods for recognizing psychological well-being status. Their investigation revealed that to recognize the level of depression is 49%. A possible biological explanation is that depression as a prodrome of dementia could arise from subcortical cerebrovascular disease. A second hypothesis is that depression is an early reaction to cognitive decline, which may occur if people in the earliest stage of dementia have an awareness of their declining cognitive abilities. According to this hypothesis, although it would seem that depression precedes the diagnosis of dementia, it would actually follow early cognitive decline. A third possibility is explained by the threshold hypothesis. Diagnosis of dementia occurs when a threshold is reached where it begins to significantly impair daily life. Depression involves cognitive deficits which may cumulate with those in early dementia, leading to an earlier stage of reaching the threshold.

The final possibility is that depression could play a causal role in dementia and is explained by the ‘glucocorticoid cascade’ hypothesis of Sapolsky. This hypothesis states that stressors trigger a release of adrenocorticotropic hormone by the pituitary gland which in turn triggers secretion of glucocorticoids from the adrenal glands. The role of the glucocorticoid receptors in the hippocampus is to inhibit further glucocorticoid secretion. Although short term secretion of glucocorticoids is useful, prolonged secretion can have harmful consequences by damaging the hippocampus leading to impairment of feedback inhibition and hippocampal damage. It is indeed found that depression often involves disregulation of the hypothalamic-pituitary-adrenal (HPA) axis and that the hippocampus is atrophied in individuals with Alzheimer’s disease.

2.3 Exploring the Relationship between Facebook and Depression

Depression, defined as when preteens and teens spend a great deal of time on social media sites, such as Facebook, and then begin to exhibit classic symptoms of depression” [33]. Nowadays,

A huge number of researchers are conducting their research on the relationship between Internet and mental health that utilizes empirical evidence to make its claims [34]. They found that some concerns exist regarding the widespread prevalence of the Internet, as some of the uses of the Internet have been associated with negative impacts on psychological well-being [35, 36]. For example, “Internet addiction,” a self-explanatory condition associated with excessive use, withdrawal, tolerance, and negative repercussions of the Internet, has been suggested as a disorder that should be included in the DSM-V [37]. Additionally, in adult users, web browsing has been positively correlated with loneliness and negatively correlated with overall life satisfaction [38]. Finally, misrepresentation of the self, a common practice on SNSs, has been linked to decrease social skills, decreased self-esteem, increased social anxiety, and increased aggression in a young population [39]. While there are several positive uses that have been identified with Internet use and Facebook participation, it is apparent that downsides also exist.

Moreno et al. [40] investigated the depression condition of college student. They found a small amount of research exists that examines the relationship between styles of Facebook use and depressive symptoms. Approximately one quarter of college students disclose symptoms of depression on Facebook via status updates and Wall posts. It is suggested that students that are highly involved in SNS communities of friends are more likely to share personal information than students who are low-frequency SNS users.

Nonetheless some of the above mentioned studies have addressed mental health issue– specifically, none of these studies have considered depression detection from Facebook data – in particular for develop a solution design understanding. Moreover, the studies reported above developed and applied various approaches to examine mental health issue among Facebook users, but none of them have developed technological solutions or approaches for instance, by applying latest data-driven for the purpose of depression detection. This thesis attempts to collect social media/open sourced data from online and examine various attributes of depressive and non-depressive users and diagnose their mental health conditions by applying ML technique to demonstrate a solution design aspect for better management of the health issues.

Chapter 3

Proposed System

In this chapter, we explain the proposed Design Science Research (DSR) procedure for designing the Depression Detection (DD) system introducing the data collection and features extraction. In addition, we present how we can build ground truth table from our datasets. After that, we present how we can classify and incorporate the data for the depression detection.

3.1 Proposed Depression Detection System

The study adopted a DSR framework for the design construct as artefacts, consisting problem definitions and solution strategies in principal informing through to ML technique intended to address the decision making requirements of mental health conditions for the appropriate management of depression issues.

Baskerville et al. [41] described that design science as representing (a) a design-science research project, (b) an artifact "build and evaluate" project such that a research project may entail, (c) the production of new knowledge from design and development, and (d) the creation of reports or articles describing this design-science research project. Baskerville et al. [41] also suggested that DSR is an approach that enables structural guides for researchers to create modernized or innovative system artifacts. Hevner et al. [42] provided four useful criteria for defining a DSR study problem space for specifying a design based solution artefact, implementing the design solution, evaluating the design artefact and communicating study details and results. Our study follows this model to produce new construct in form of an artifact design and its evaluation. In addition, our description is guided by Hevner et al. [42] explicating the level of contribution to artefact abstraction and knowledge sharing. To build up a depression detection system, we utilise four design phases such as for data collection and processing, data classification, problem statement and diagnosis of mental health issues (illustrated in figure 3.1).

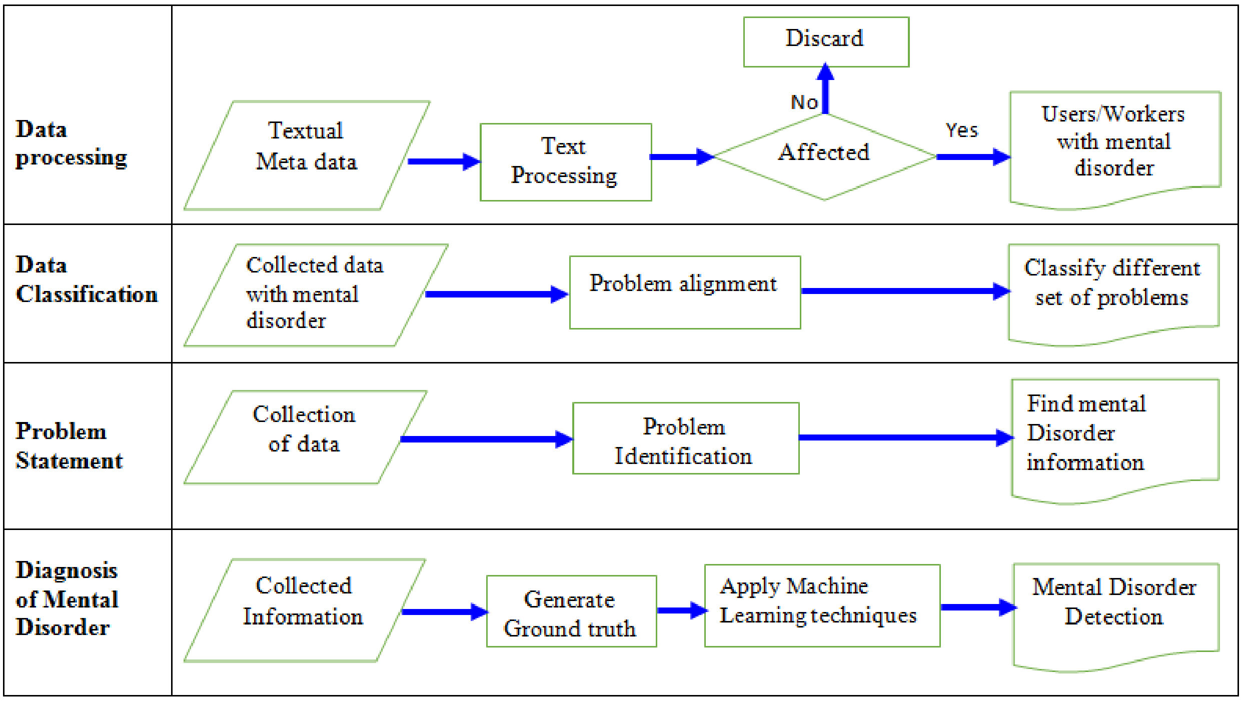

Figure 3.1: Methodological steps for identifying mental health issues of our proposed DD system

3.2 Data Sets and Data Collection

3.2.1 Description of the First Data Collection

We use publicly available dataset [43] containing attributes of workers working in IT/tech workplace. The dataset contains the number of attributes of IT/Tech workplace workers. In this dataset, the total number of records in each part is approximately, 10000. We used part 1 to analyze the situation in IT workplace and part 2 to detect the percentage of IT workplace workers who are suffering from mental stress.

3.2.2 Description of the First Data Set

This dataset contains 100 columns where each column represents a unique piece of information of a tech worker. Although each of the individual columns provides the valuable information for diagnosis of tech workplace workers mental health conditions, in this paper, we took 23 columns.

We didn’t take the remaining 77 columns because these columns are not necessary for detecting mental health condition of tech workplace workers. So, we examined these 23 columns and divided our dataset into two parts to detect and discuss the mental stress of IT workplace workers. Overall, these columns contain both mentally sick and mentally sound workers information. Part 1 contains 13 columns and part 2 contains 10 columns based on the following questions which are shown in below.

Overview of part1

- Have you heard of or observed negative consequences for co-workers who have been open about mental health issues in your workplace?
- Do you believe your productivity is ever affected by a mental health issue?
- If yes, what percentage of your work time (time performing primary or secondary job functions) is affected by a mental health issue?
- Do you think that discussing a mental health disorder with previous employers would have negative consequences?
- Would you be willing to bring up a physical health issue with a potential employer in an interview?
- Would you bring up a mental health issue with a potential employer in an interview?
- Do you think that team members/co-workers would view you more negatively if they knew you suffered from a mental health issue?
- Which of the following best describes your work position?
- How willing would you be to share with friends and family that you have a mental illness?
- What is your age?
- What is your gender?
- What country do you live in?
- What US state or territory do you live in?

Overview of part2

- Have you observed or experienced an unsupportive or badly handled response to a mental health issue in your current or previous workplace?
- Have your observations of how another individual who discussed a mental health disorder made you less likely to reveal a mental health issue yourself in your current workplace?
- Do you have a family history of mental illness?
- Have you had a mental health disorder in the past?

- Do you currently have a mental health disorder?
- If yes, what condition(s) have you been diagnosed with?
- If maybe, what condition(s) do you believe you have?
- Have you been diagnosed with a mental health condition by a medical professional?
- Have you ever sought treatment for a mental health issue from a mental health professional?
- If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?

3.2.3 Description of the Second Data Collection

We worked on Facebook users' comments for depressive behavioral exploration and detection. We collected data from the Facebook [44]. Preparing of social network data, in particular Facebook user's comments is one of the primary challenges which bear information on whether or not they could contain depression bearing content. To tackle this issue we use NCapture for collecting data from Facebook [45, 46]. For better understanding the explosion of unstructured data in the world today, NCapture is powerful software for qualitative data analysis. It's designed to help us organize, analyze and find insights in unstructured or qualitative data like interviews, open-ended survey responses, articles, social media and web content. In addition it gives us a place to organize and manage our material so that we can start to find insights in our data in a more efficient way [47, 48].

3.2.4 Description of the Second Data Set

Our primary dataset contains total 21 columns where 13 columns represent the linguistic style (articles, prepositions, auxiliary verbs, conjunctions, personal pronoun, impersonal pronouns, verbs, negation etc.) information, 5 columns represent the emotional (positive, negative, sad, anger and anxiety) information, 3 columns represent the temporal process (past, present and future) information and each column gives the individual information's about depressive behavior (refer to table 3.1).

Table 3.1: Characteristics of raw data

Characteristic	Quantity
Total number of cells	150045
Total number of depressive indicative cells based on our manual response (with Zero values)	87129

Total number of non-depressive indicative cells based on our manual response	62916
Total number of depressive indicative cells based on our manual response for linguistic style (without Zero value)	43551
Articles	3200
Prepositions	3842
Auxiliary Verbs	3813
Conjunctions	3619
Personal Pronoun	3875
Impersonal Pronoun	3444
Verbs	4004
Negations	2637
Pronoun	3989
Adverb	3407
Adjective	3342
Comparative	2436
Interrogative	1943
Total number of depressive indicative cells based on our manual response for emotional process(without Zero value)	13884
Positive	2676
Negative	4149
Sad	1733
Anger	1177
Anxiety	4149
Total number of depressive indicative cells based on our manual response for temporal process(without Zero value)	8237
Past	2527
Present	3909
Future	1801

3.2.5 Building Ground Truth Dataset

This section discusses the process employed to construct our dataset with ground truth label information (on whether the comments is depression indicative). It is divided into two sets (a) for the positive (YES) class (depression indicative comments) and (b) for the negative (NO) class (non-depression indicative comments).

From First Dataset

We now discuss how we constructed our dataset with ground truth information (on whether the IT/Tech workers conditions are mentally sick). We use our dataset and divided into two sets (1) for the positive (YES) class (mentally sick) and (2) for the negative (NO) class (mentally sound). We justified the information of each set by two experts manually. We analyzed approximately 10000 records of IT workplace workers where 58% obtained YES and 42% obtained NO. Table 3.2 illustrated the overview of mental condition of IT/Tech workers.

Table 3.2: Overview of mental condition of IT/Tech workers

Total number of IT/Tech workers	1433
Mentally Sick workers	933
Mentally Sound	500

From Second Dataset

Out of the total 7145 comments, 58% obtained *'YES'* for depression indicative comments and 42% obtained *'NO'* for non-depressive indicative comments. Table 3.3 illustrated the dataset information and a few examples of depression indicative comments are given in Table 3.4.

Table 3.3: Depression indicative distribution of dataset

Data set information	Quantity
Total number of Facebook Users comments	7145
Depression Indicative comments	4149
Non-depression Indicative comments	2996

Table 3.4: Examples of Depression Indicative Comments

Examples	Response
I am currently having the problem of restlessness and needing to move but I also don't feel like moving.	YES
I feel sad and con not concentrate in my studies	YES
I find faults in all the people around me and I feel lonely and alone.	YES
My daughter started on fepakote at age 16. She did ok but, when she started lithium things changed for the better.	YES
I hate the fact that I know some of my triggers but can't avoid them...l have to just keep up the exposure as I've been told this is better than isolating myself in fear.	YES
I'm having a terrible day.	YES
Put an alarm on your phone I need to again it works	NO
I use to use rubbing alcohol and worked whl younger but dint give a rats ass now.still get teased by it by insecure men. But they can go fuck themselves.	NO
Story of my life., I struggle with these things daily	NO
I take Latuda at night because it makes me sleepy and xanax throughout the day for anxiety.	NO

3.3 Measuring Depressive properties

3.3.1 Measuring Depressive Behaviour from First Dataset

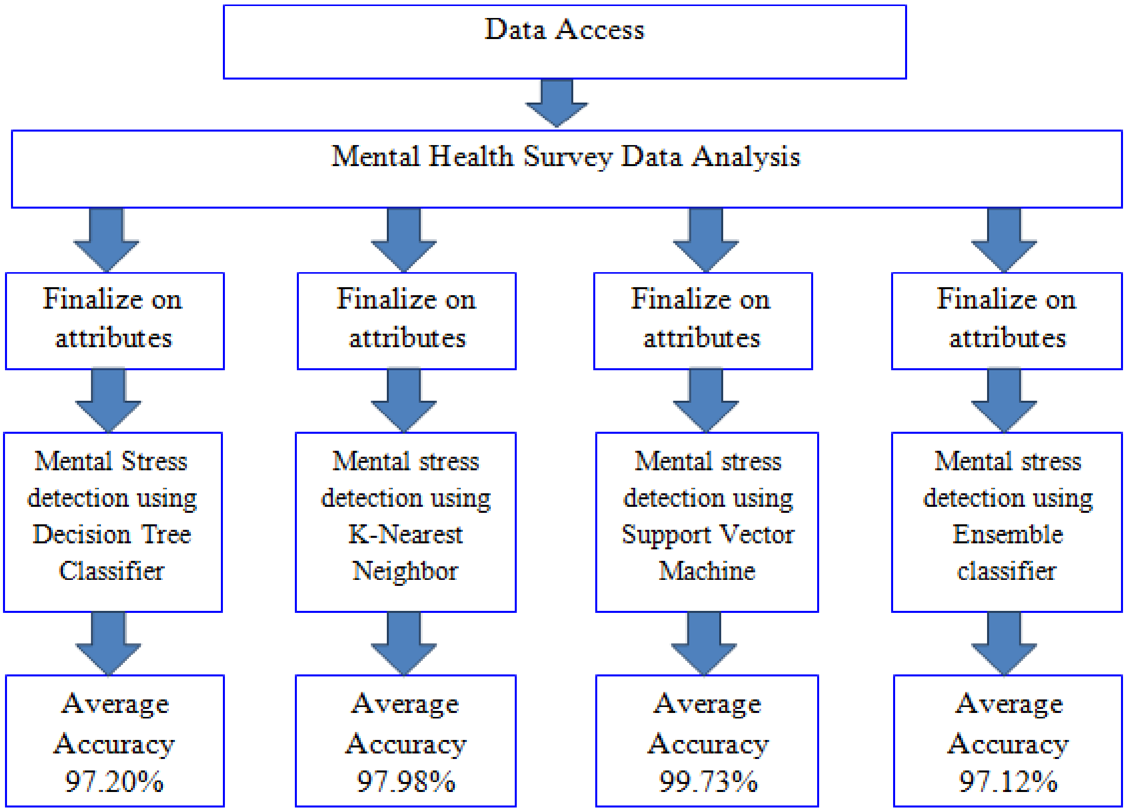

Figure 3.2: Detecting procedures that were used in first dataset.

Figure 3.2 represents a measuring procedure for detecting mental health conditions using ML technique. In this procedure, we collect data for finalizing the features of the dataset and then applied the ML technique for detecting the mental health issues that are centered to anxiety, depressions, bipolar, personality and mood that we focused initially for our investigation.

We first analyze a set of attributes such as age group, sex, occupation, and country that have used to represent the range of mentally sick people in IT/Tech workplace (Table 3.5). Attributes such as family problem, past condition, present condition, and workplace environment condition are used for detecting the mental disorder in the workplace (Table 3.6).

Table 3.5: Categorization of IT/Tech workers in various mental disorders

Variable	Total	Anxiety Disorder	Mood Disorder	Others Disorder
Age Group				
20-30	222	134	53	35
31-40	239	150	69	20
41-50	76	38	29	9
>50	27	13	10	4

Sex				
Male	363	207	102	54
Female	177	114	52	11
Occupation				
Back-end Developer	139	72	44	23
Supervisor/Team Lead	75	49	18	8
Executive Leadership	28	13	11	4
DevOps/ SysAdmin	55	34	16	5
Dev Evangelist/ Advocate	27	16	11	0
Support	26	17	6	3
One-person shop	47	20	15	12
Front-end Developer	76	55	16	5
Designer	24	16	5	3
Other	66	40	19	7
HR	3	3	0	0
Country				
United States of America	384	228	115	41
United Kingdom	62	43	17	2
Australia	16	9	6	1
Canada	23	15	3	5
Germany	10	4	5	1
Netherland	22	8	5	9
France	3	2	0	1
Denmark	2	1	0	1
Other Country	45	26	11	8

Table 3.6: Number and proportion of IT/Tech workplace workers mental disorder into various conditions

		Sex	Anxiety	Mood	Other	Total
Mental health disorder in the past?	YES	M	137	22	13	172
		F	82	15	3	100
	NO	M	70	14	6	90
		F	33	2	1	36
Family history of mental illness?	YES	M	86	13	8	107
		F	57	9	1	67
	NO	M	121	23	11	155
		F	58	8	3	69
Currently have a mental health disorder?	YES	M	124	22	15	161
		F	72	14	2	88
	NO	M	83	14	4	101
		F	43	3	2	48
Mental health issue in current or previous workplace?	YES	M	107	15	9	131
		F	58	11	2	71
	NO	M	100	21	10	131
		F	57	6	2	65
Diagnosed with a mental health condition by a medical	YES	M	93	16	9	118
		F	61	12	3	76

professional?	NO	M	114	20	10	144
		F	54	5	1	60
If you have a mental health issue, do you feel that it interferes with your work when being treated effectively?	NA	M	103	20	10	133
		F	58	6	1	65
	Rarely	M	46	5	3	54
		F	21	4	1	26
	Some Times	M	47	7	5	59
		F	30	6	2	38
	Often	M	11	4	1	16
		F	6	1	0	7

3.3.2 Measuring Depressive Behaviors from Second Dataset

To characterize and differentiate between depressive and non-depressive posts, we analyze and construct the various feature sets based on psycholinguistic dimensions, and textual features from the user posts. They are explained briefly as follows:

A. Psycholinguistic Features: LIWC package [49] is a psycholinguistic lexicon created by psychologists to recognize the various emotional, cognitive, and linguistic components lies on user's verbal or written communication. For each input in the corpus, it returns more than 70 output variables with higher level hierarchy of psycholinguistic features such as

--- Linguistic dimensions, other grammar, informal language

--- Temporal, affective, social processes

--- Cognitive, biological, perceptual processes

--- Personal concerns, drives, relativity.

These higher-level categories are further specialized in subcategories such as

--- Biological processes - body, sexual, health and ingestion.

--- Affective processes - positive emotion, negative emotion

--- And negative emotion further sub-classified as anger, anxiety, and sadness.

--- Drives - affiliation, achievement, power, regard, and risk.

For our task, we selected 38 features among 54 variables, and converted each clinical and non-clinical post into mathematical vectors based on psycholinguistic features. The word in the posts could fit some categories and not fit into some categories. Hence, we could consider the classification of posts based on psycholinguistic dimensions, to which category it belongs. Figure 3.3 shows the detecting procedures of depression and Table 3.7 shows the various categories of LIWC psycholinguistic processes.

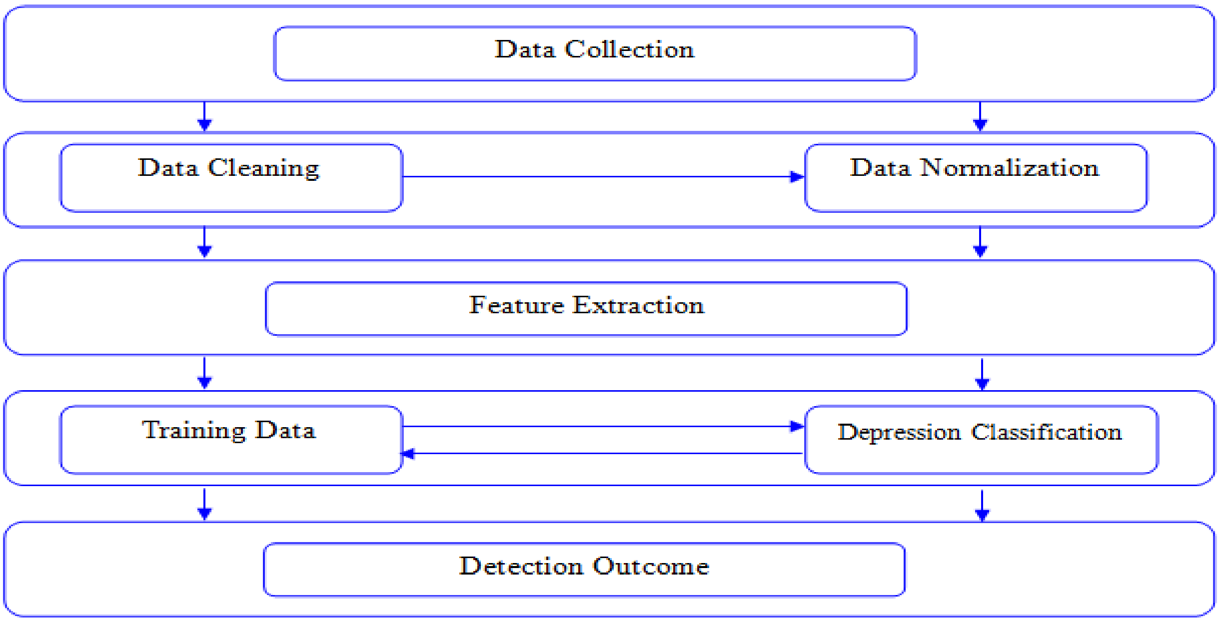

Figure 3.3: Detecting procedures that were in first dataset.

Table 3.7: Break down of target terms into various categories.

LIWC derived cues	Example word
Emotional process	
Positive emotion words	happy, love, nice, sweet
Negative emotion words	worthless, loser, hurt, ugly, 'nasty'
Sadness words	worry, crying, grief, sad
Anger words	stop, shit, hate, kill, annoyed
Anxiety words	worried, fearful
Temporal process	
Present focus	today, is, now
Past focus	Ago, did, talked
Future focus	Shall, may, will, soon
Linguistic Style	
Articles	A, an, the
Prepositions	For, in, of, to, with, above
Auxiliary verbs	Do, have, am, will
Adverbs	Quickly, Slowly, very, really
Conjunctions	And, But, whereas
Total pronouns	I, them, itself
Personal pronoun	I, them, her
1st Person singular pronoun	I, me, mine
1st Person plural pronoun	we, us, our
2nd Person	you, your
3rd Person singular pronoun	He, she, her, him
3rd Person plural pronoun	they, their, they'd
Impersonal Pronouns	it, it's, those
Verbs	Go, Good
Negation	Deny, Dishonest, no, not, never

In this table, positive, negative, sad, anger, anxiety refers to terms chosen for emotional process. Then, present, past, future refers to terms chosen for temporal process and finally, articles, prepositions, auxiliary verbs, adverbs, conjunctions, pronouns, verbs refers to terms chosen for linguistic style.

Emotional processes: Emotional processes, a complex experience of consiousness, bodily sensation, and behaviour that reflects the personal significance of a thing, an event, or a state of affairs. The analysis of the emotional comments of social network data can be leveraged to produce reliable predicts in a variety of circumstances. We use psycholinguistic resource LIWC (http://www.liwc.net/) for considering five features of the emotion state manifested in the comments: positive affect (PA), negative affect (NA), sadness affect (SA), anger affect (AA), and anxiety affect (AnA).

Temporal Process: Generally, temporal process word provides information about past focus category, present focus category and future focus category of how people are referencing each other and their degree of emotionality.

Linguistic Style: We introduce features to characterize comments based on the use of linguistic style. Facebook user comments processing is built on the idea that language conveys information beyond the literal meaning of the words used. Some studies have shown that the way in which people use language can reveal information about their thoughts and emotions. We also use here LIWC for determining 9 specific linguistic styles: articles, prepositions, auxiliary verbs, adverbs, conjunctions, personal pronoun, impersonal pronouns, verbs, and negations. It was designed to measure word use in psychologically meaningful categories. In addition it has been successfully used to identify relationships between individuals in social interactions, including relative status, deception, and the quality of close relationships.

B. Textual Features

We extracted the textual features to find the most frequent words in the underlying posts between two classes. The reason for predicting the textual features from the user posts is to analyze the most frequent words, that underlies in the textual content and how it varies depends on the context. For instance, when predicting the depressive posts, we can find the terms such as sad, unhappy, unlucky, weak, and depressed as the most frequent.

The Bag of Words (BoW) model is the general way of extracting textual features, when modeling with Machine Learning classifiers. This is very simple and flexible model, and this algorithm counts the number of occurrence of words in the corpus. In our context, it can be represented as a binary word presence representation that indicates, if a word is present in each post. Each distinct word in the corpus corresponds to a feature in the representation. This provides list of words with their word counts per post. Table 3.8 illustrates the top uni-grams belongs to each class.

Table 3.8: Top Textual features associated with depressive and non-depressive posts

Class	Example Word
Depressive Posts	loser, depress, lonely, sad, alone, weak, useless, life, imbalance, blame, problems, unsuccessful, suicidal, torture, safe, escape, worry, intimidate, uncomfortable, therapy, medication, shit, pressure, conversation, hurts, myself, worth, break, nobody, mine, painful, hate, suck
Non-Depressive Posts	lol, work, weekend, say, friends, brilliant, follow, tips, bieber, love, amazing, hello, now, bored, awesome, beautiful, romantic, fuck* , perfect, excited, smile, meet, tonight, life, movie, football, favorite, sleepy, great, night, team, good, anyone, you, your, tomorrow, money

3.3.3 Classification Model

This stage constructs prediction model for depression post/comments recognition, by considering the psycholinguistic and textual features as input. Consider our training corpus B = p_1; p_2;…..p_n of n posts/comments, such that each post/comments p_i is labelled with the class either as depressive or non-depressive, where L = $l_1|l_2$ The task of a classifier f is to find the corresponding label for each post.

$$f : B\epsilon L \quad f(p) = l$$

Here, we employ four popular classifiers: Support Vector Machine, Decision tree, Ensemble, and k-Nearest Neighbor (kNN).

Support Vector Machines (SVM): a non-probabilistic linear binary classifier, that maps an input feature vector into a higher dimensional space and find a hyperplane that separates the data into two classes with the maximal margin between the closest samples in each class.

Decision Tree (DT): an interpretable classifier creates the hierarchical tree of the training instances, in which a condition on the feature value is used to divide the data hierarchically. For the classification of text documents, the conditions on decision tree nodes are commonly defined in terms and a node may be subdivided to its children based on the presence or absence of a term in the document.

Ensemble: Ensemble methods use multiple learning algorithms of decision tree for better predictive performance.

k-Nearest Neighbor (kNN): a proximity-based classifier use distance- based measures i.e., the documents which belong to the same class are more likely similar or close to each other based on the similarity measures. The classification of the test document is reported from the class labels of the k nearest similar documents in the training set.

Chapter 4

Experimental Analysis

In this chapter the experiments for the two datasets are presented and discussed. At the beginning section 4.1, it covers details about the experimental setup. The evaluation of precision, recall, F-measure and accuracy are described and furthermore also the experimental analysis is stated. In the following sections each individual experiment is explained. Finally, results are discussed and potential conclusions are drawn.

4.1. Experimental Setup

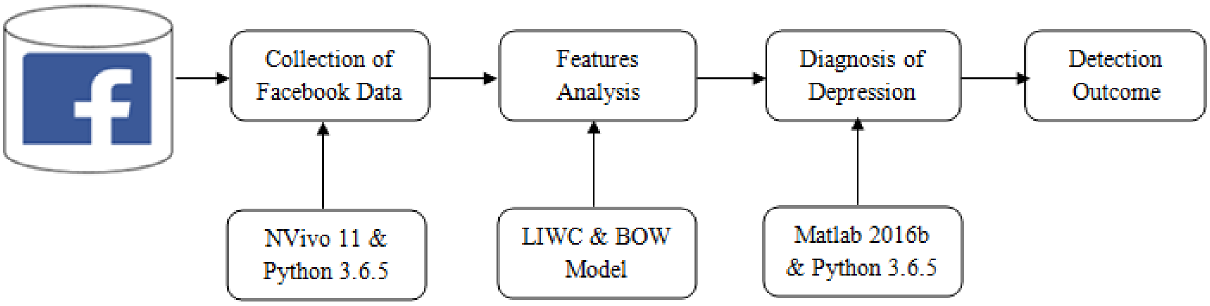

Figure 4.1 an experimental overview of Facebook data analysis for depression detection

To evaluate performance and effectiveness of our experiment, we applied several software and open source tools as shown in the figure 4.1. The experiments were carried out in a computer with Windows 7 Operating System, Intel Core I 5 Processor, RAM of 4GB with 3G internet connection of local operator.

Here, the table 4.1 shows the name of the Package, Descriptions and URL link used in that project. The column file says that total number files are considered from the project sources.

Table 4.1: Open Source Software/Projects found from different Enterprise Repository online

Package name	Descriptions	URL link
NVivo 11	**NCapture** is a free web browser extension,	http://www.qsrinternatio

	developed by QSR that enables you to gather material from the web to import into **NVivo**. It can use **NCapture** to collect a range of content—for example, articles or blog posts. It can also collect social media content from Facebook, Twitter and YouTube.	nal.com/nvivo/nvivoproducts/nvivo-11-for-windows
LIWC	LIWC (Linguistic Inquiry and Word Count) is a text analysis program. It calculates the degree to which various categories of words are used in a text, and can process texts ranging from e-mails to speeches, poems and transcribed natural language in either plain text or Word formats.	http://liwc.wpengine.com/
Matlab 2016b	MATLAB is a multi-paradigm numerical computing environment. A proprietary programming language developed by Math Works, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, C#, Java, Fortran and Python.	https://www.mathworks.com/products/new_products/release2016b.html
Python 3.6.5	Python is a widely used high-level, general-purpose, interpreted, dynamic programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than would be possible in languages such as C++ or Java. The language provides constructs intended to enable clear programs on both a small and large scale.	https://anaconda.org/anaconda/python

4.2 Evaluation of Precision, Recall, F-Measure and Accuracy

In this section, we investigated the performance of different classifiers in detecting depression in a shorter time. The experiment is conducted using MATLAB 2016b. We applied four major classifiers: Decision trees, KNN, SVM and Ensemble. Each classifier has sub-classifiers such as Decision trees- Simple DT, Medium DT, and Complex DT; SVM- Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian, and Coarse Gaussian; KNN- Fine, Medium, Coarse, Cosine, Cubic and Weighted, Ensemble- Boosted tree, Bagged tree, Subspace discriminant, Subspace KNN, RUSBoosted tree.

Using the above classification techniques, we examined detection performance of Facebook user comments and IT workers information. To comprehend the significance of different feature types, we applied above mentioned four classifier techniques. The results of the analysis are reported in table 4.2, 4.3, 4.4 and 4.5. Performance of these classifiers has been computed by using the evaluation matrices parameters such as precision, recall, F-measure and accuracy. It is based on four possible classification outcomes. True Positive (TP) = The depression cases that are positive and predicted positive. True Negative (TN) = The depression cases that are negative and predicted negative. False Negative (FN) = The depression cases that are positive but predicted to be negative. False Positive (FP) = The depression cases that are actually negative but predicted to be positive.

All the evaluation metrics are defined as follows.

Precision is the ratio of true positives to the cases that are predicted as positive. It is the percentage of selected cases that are correct.

$$Precision\ (P) = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall is the ratio of true positives to the cases that actually positive. It is the percentage of corrected cases that are selected.

$$Recall\ (R) = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

F-measure is the mean of Precision and Recall. It takes both false positives and false negatives into an account. F-measure is calculated as:

$$F - Measure = 2\frac{PR}{P + R}$$

The experiment is carried out by using 10-fold cross-validation on all testing data set. For every classifier, we show the value of its sub-classifier which persists to high F-measure (refer to Table 4.2, 4.3, 4.4 and 4.5).

4.3 First Experimental Analysis

This experiment was performed on the first Facebook dataset of this thesis, which is described in section 3.3.1. The dataset was crawled over a time period of three months. The main aspect of this dataset is that it measures attitudes towards mental health and frequency of mental health disorders in the tech workplace based on 1200 Workers. To recap the specifications of this dataset the main characteristics are:

- Family history of mental illness workers.
- Workplace environment.
- Age of the mental illness workers.
- Mental health condition of IT workers.
- Mental health issues with direct supervisor.

The following table 4.2 shows the performance evaluations of first dataset that has Precision, Recall F-measure and Accuracy.

Table 4.2: Precision and recall corresponding to best F-measure of the different classifier

Algorithm	***Precision***	***Recall***	***F-measure***	***Accuracy***
Complex Tree	0.994623	0.991425	0.993022	99
Medium Tree	0.994577	0.982851	0.988679	98.5
Simple Tree	0.957974	0.952840	0.955400	94.1
Fine KNN	0.997851	0.995712	0.996781	99.5
Medium KNN	0.988159	0.983922	0.986036	98.1
Coarse KNN	0.933537	0.978563	0.955520	94
Cosine KNN	0.997826	0.983922	0.990825	98.7
Cubic KNN	0.988159	0.983922	0.986036	98.1
Weighted KNN	0.997851	0.995712	0.996781	99.5
Linear SVM	0.997860	1	0.998929	99.9
Quadratic SVM	0.996794	1	0.998394	99.8
Cubic SVM	0.996794	1	0.998394	99.8
Fine Gaussian SVM	0.997851	1	0.998924	99.5
Medium Gaussian SVM	0.996788	0.998927	0.997856	99.7
Coarse Gaussian SVM	0.996791	0.998928	0.997858	99.7
Ensemble Boosted Tree	0.996784	0.996784	0.996784	99.5
Ensemble Bagged Tree	0.996791	0.998928	0.997858	99.7
Ensemble Subspace Discriminant	0.95010	0.979635	0.964643	95.3
Ensemble Subspace KNN	0.904244	0.981779	0.941418	92

Ensemble RUSBoosted Tree	0.995694	0.991425	0.993555	99.1

Figure 4.2 represented the result of precision and recall value of different classifier where the x-axis represents the different sub-classifier of Decision Tree, KNN, SVM and Ensemble techniques. we mentioned (1-20) as a symbol of Decision trees- Simple Tree, Medium Tree, and Complex Tree; SVM- Linear, Quadratic, Cubic, Fine Gaussian, Medium Gaussian, and Coarse Gaussian; KNN- Fine, Medium, Coarse, Cosine, Cubic and Weighted, Ensemble- Boosted tree, Bagged tree, Subspace discriminant, Subspace KNN, RUSBoosted tree respectively and y-axis represents the range of value.

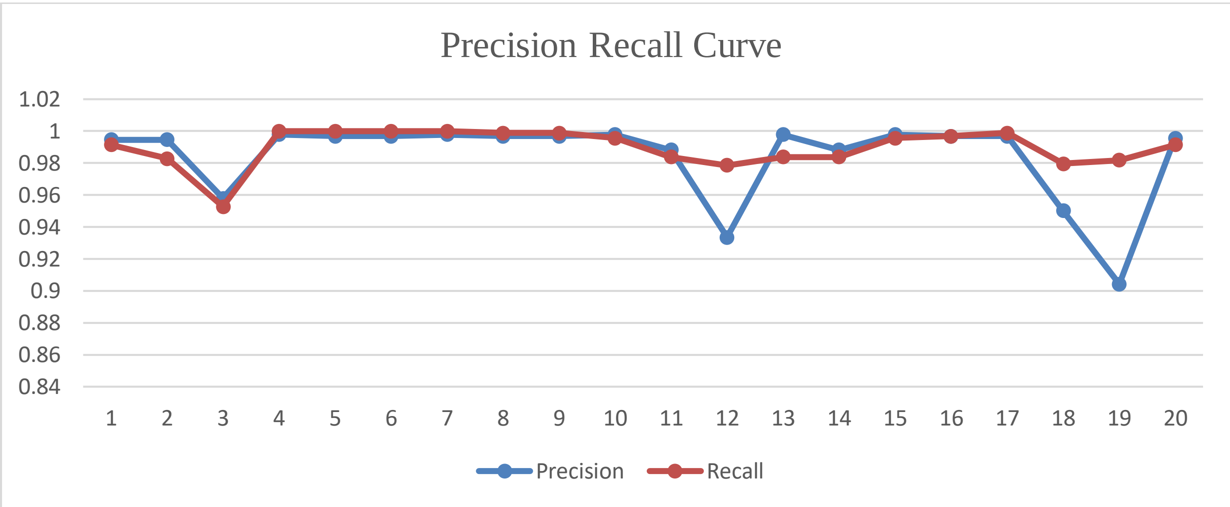

Figure 4.2: Precision-recall curve for the 20 sub-classifier of machine learning techniques

Similarly, Figure 4.3 shows the result of F-Measure where the x-axis represents the name of different sub-classifier and y-axis represents the percentage of the sub-classifier result.

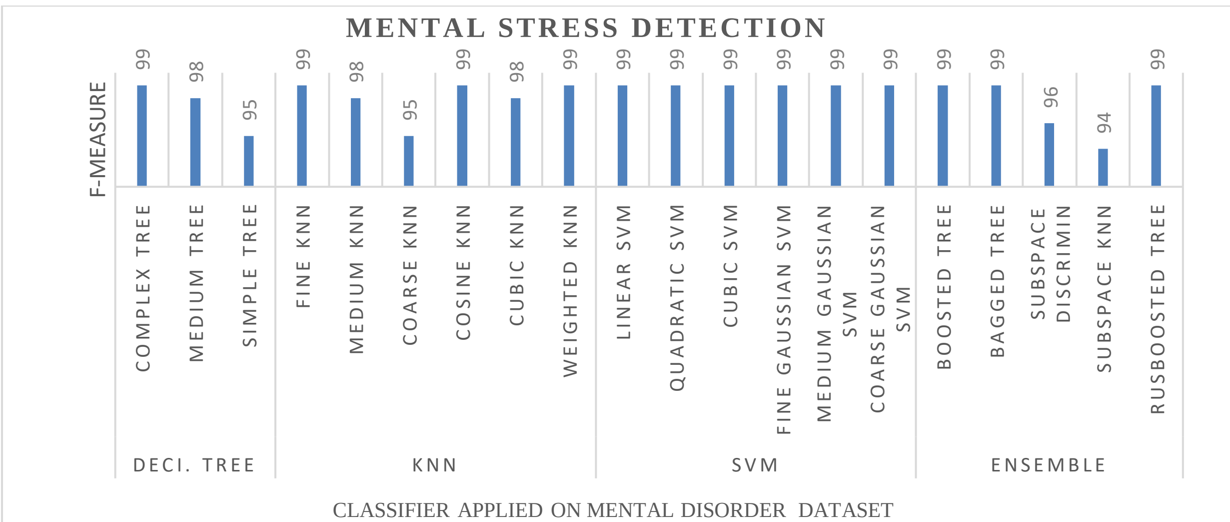

Figure 4.3: Performance comparisons among the classifiers

Moreover, In figures 4.4 and 4.5, we have shown the various range of mental disorders values based on age group and sex, where x-axis represent the various name of the mental disorder and y-axis represent the number of people who are feeling like anxiety and mood.

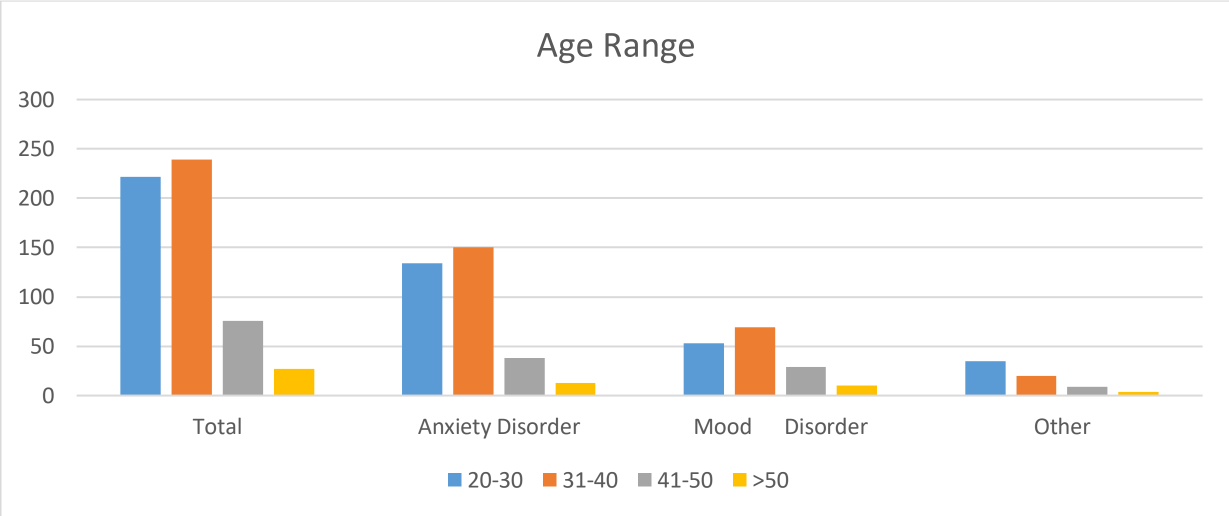

Figure 4.4: The range of age group in IT/Tech workers with mental disorders

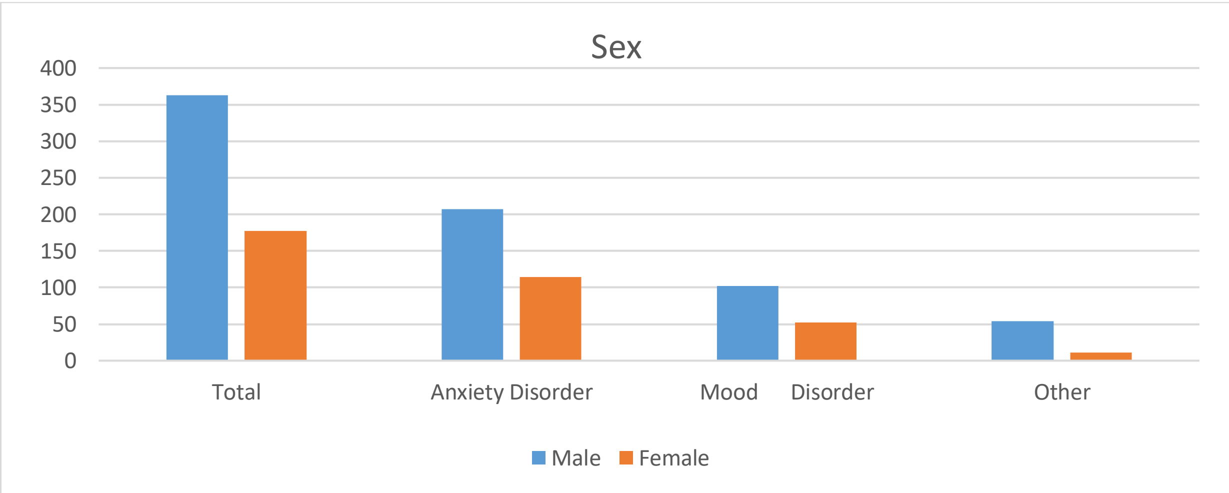

Figure 4.5: The number of male and female in IT/Tech workers with mental disorders

Similarly, Figure 4.6 and 4.7 demonstrated the result of various range of mental disorders values based on occupation such as back-end developer, supervisor/team leader, executive, system admin, dev evangelist, ope person shop, font-end developer, designer, and HR, and country such as USA, UK, Australia, Canada, Germany, Netherland, France etc. where x-axis represents the designation of IT workplace workers and their country respectively and y-axis represent the number of people who are feeling like anxiety, mood etc. In figure 4.6, we examined all of the occupational positions in IT workplace for calculating the mental disorders. It observed that for all of the mental disorder, back-end developers are suffering very much and then respectively.

We trust that our present study has lied the ground for future research on deductions and revelation of new data in view of cause-event correlation.

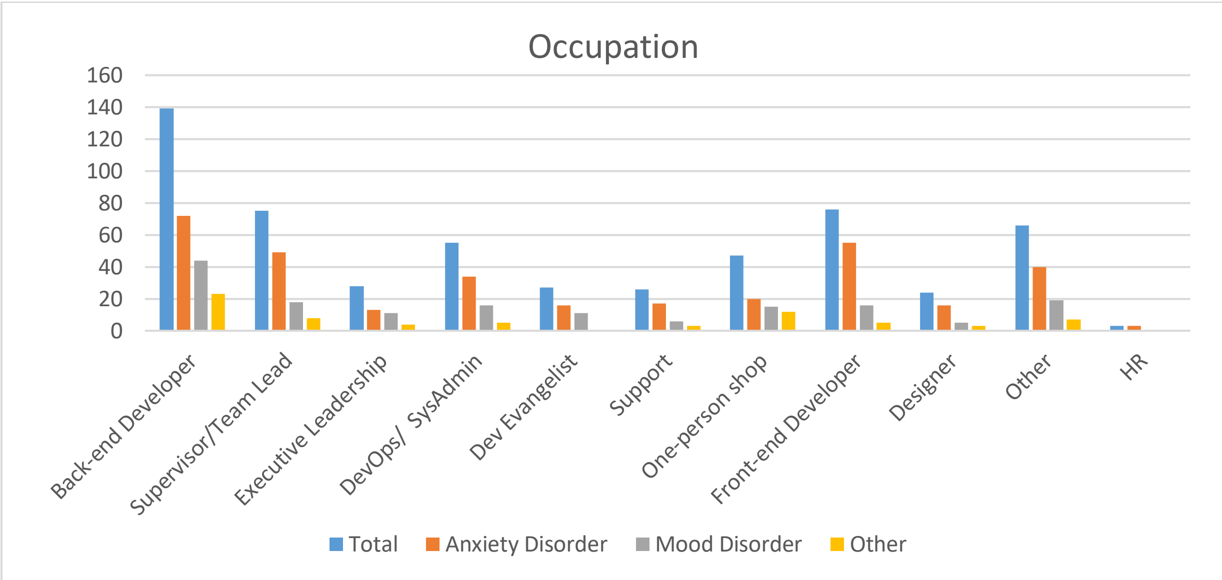

Figure 4.6: The number of people in the different section of IT/Tech workplace with mental disorders.

In figure 4.7, It shows that the populations of some countries who are suffering from different mental disorders in IT/Tech workplace. The country with the larger populations (67%) is USA with anxiety, mood and other disorders whereas Denmark has the smallest population. Apart from USA, the largest countries are UK, Australia, and Canada respectively.

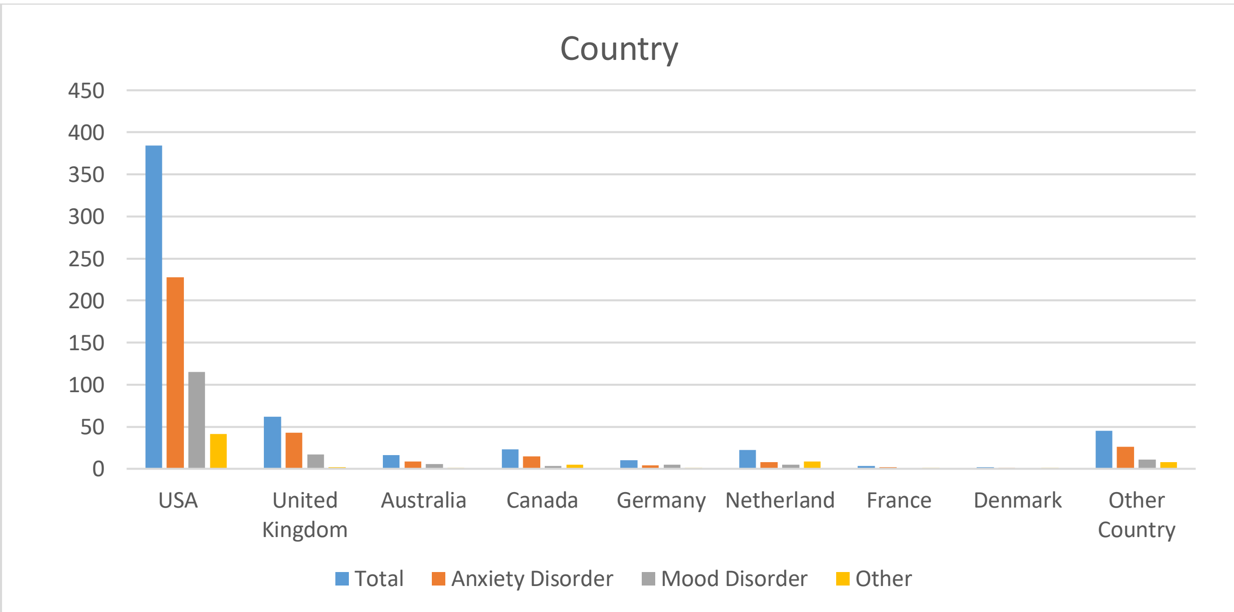

Figure 4.7: The number of people in the different section of IT/Tech workplace with mental disorders

4.4 Second Experimental Analysis

Study 1: In study1, the experiment was performed on the second Facebook dataset of this thesis (see section 3.3.2 for further information). The dataset was crawled over a time period of one and half months. The main aspect of this dataset is that it is a randomly chosen subset of the complete Facebook dataset based on 7100 users. To recap the specifications of this dataset the main characteristics are:

- Emotional process
- Temporal process
- Linguistic style

In contrast to the first dataset, this dataset focuses on Facebook user's depressive information. The results of the experimental evaluation to measure the performance of depression analysis, we have summarized in the Table 4.3 and 4.4.

Table 4.3: Performance Metrics of Machine Learning Classifiers based on emotional process and linguistic style features set.

Feature	Emotional Process			Linguistic Style		
Decision Tree						
Algorithm	Pr	Re	Fm	Pr.	Re.	Fm.
Complex Tree	**.59**	.85	.69	.58	.86	.69
Medium Tree	.58	**.96**	**.73**	.58	**.97**	**.73**
Simple Tree	**.59**	**.97**	**.73**	.58	**.99**	**.73**
Average Value of DT	**.59**	**.93**	**.72**	.58	**.94**	**.72**
K Nearest Neighbors (KNN)						
Fine KNN	**.59**	.59	.59	.58	.58	.58
Medium KNN	**.59**	.59	.59	**.59**	.55	.57
Coarse KNN	**.59**	**.88**	.71	**.59**	.80	.70
Cosine KNN	.58	.59	.58	**.59**	.60	.60
Cubic KNN	.59	.59	.59	.60	.54	.57
Weighted KNN	.58	.62	.60	.59	.65	.62
Average Value of KNN	.59	.64	.61	.59	.62	.61
Support Vector Machine (SVM)						
Linear SVM	.58	1	**.73**	.58	**1**	**.73**
Quadratic SVM	.57	.81	.67	.58	**1**	**.73**
Cubic SVM	.58	.86	.69	.58	.91	.71
Fine Gaussian SVM	.58	.88	.70	**.59**	.90	.71
Medium Gaussian SVM	.58	**.99**	**.73**	.58	**.99**	**.73**
Coarse Gaussian SVM	.58	1	**.73**	.58	1	**.73**
Average Value of SVM	.58	.92	**.71**	.58	.97	**.73**
Ensemble Classifiers						

Ensemble Boosted Tree	.58	**.96**	**.73**	.58	**.99**	**.73**
Ensemble Bagged Tree	**.59**	.68	.63	.58	.68	.63
Ensemble Subspace Discriminant	.58	**.99**	**.73**	.58	**.99**	**.73**
Ensemble Subspace KNN	**.59**	.63	.61	**.59**	.66	.62
Ensemble RUSBoosted Tree	**.62**	.44	.51	**.61**	.40	.48
Average value of Ensemble	**.59**	.74	.64	**.59**	.74	.64

Similarly, we show the results of these prediction models in table 4.4. The outcome shows that the best performing model is DT. For temporal process and all features, KNN and SVM gives almost same the high precision but DT gives the highest result for recall and F-measure relating to the class of depression indicative comments of Facebook user (refer to fig. 4.8).

Table 4.4: Performance Metrics of Machine Learning Classifiers based on temporal process and all features set.

Feature	Temporal process			All features		
Decision Tree						
Algorithm	Pr.	Re.	Fm.	Pr.	Re.	Fm.
ComplexTree	.58	**.90**	.71	**.59**	.84	.69
Medium Tree	.58	**.97**	**.73**	.58	.90	.70
Simple Tree	.58	**.99**	.74	.58	**.98**	**.73**
Average Value of DT	.58	**.95**	.73	.58	**.91**	**.71**
K-Nearest Neighbors (KNN)						
Fine KNN	.57	.58	.58	**.59**	.57	.58
Medium KNN	.58	.57	.57	**.59**	.53	.56
Coarse KNN	.58	.89	.70	**.59**	.77	.67
Cosine KNN	.59	.58	.59	.60	.60	.60
Cubic KNN	.58	.57	.57	**.59**	.52	.55
Weighted KNN	.57	.59	.58	**.59**	.64	.61
Average Value of KNN	.58	.63	.60	**.59**	.61	.60
Support Vector Machine (SVM)						
Linear SVM	.58	1	**.73**	.58	1	**.73**
Quadratic SVM	.58	.76	.66	.58	**.99**	**.73**
Cubic SVM	.57	.81	.67	.58	.70	.63
Fine Gaussian SVM	.58	.86	.69	**.59**	.94	.72
Medium Gaussian SVM	.58	**.99**	.73	.58	**.98**	**.73**
Coarse Gaussian SVM	.58	1	**.73**	.58	1	**.73**
Average Value of SVM	.58	.90	**.70**	.58	.94	**.71**
Ensemble Classifiers						
Ensemble Boosted Tree	.58	**.97**	**.73**	.58	**.96**	.72
Ensemble Bagged Tree	**.59**	.69	.63	**.59**	.70	.64
Ensemble Subspace Discriminant	.58	**.99**	**.73**	.58	**.99**	**.73**
Ensemble Subspace KNN	.58	.65	.61	**.59**	.66	.62

Ensemble RUSBoosted Tree	**.61**	.46	.53	.63	.41	.50
Average Value of Ensemble	**.59**	.75	.65	.59	.74	.64

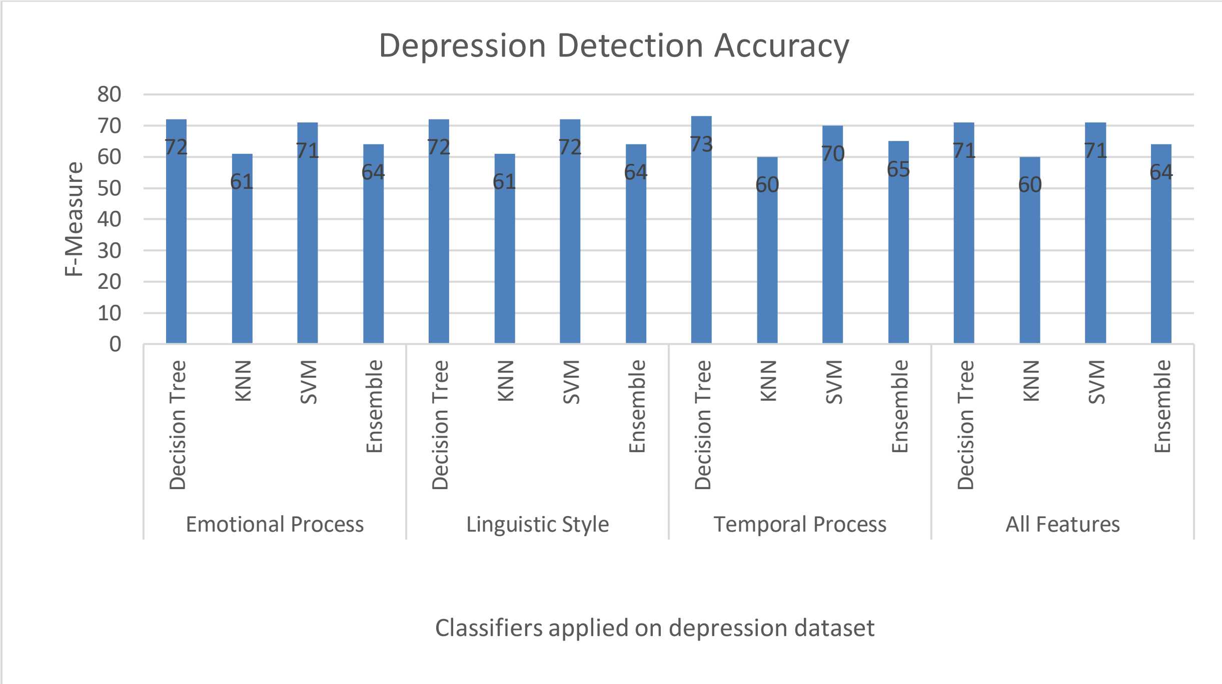

Figure 4.8: Depression Detection Accuracy on depression dataset by different classifier

Study 2: In study 2, we evaluated the performance of the classifiers on top of the constructed of textual feature. The table 4.5 shows that BOW features, extracted from the user posts are performed well. Comparing the LIWC features, Decision Tree classifier provides the better performance by achieving the higher accuracy of 96% with higher recall value of 98%.

Table 4.5: Performance Metrics of Machine Learning Classifiers based on textual Feature sets

Method	Precision	Recall	F-Measure	Accuracy
SVM+ Textual Features	.61	.67	.64	56
KNN+ Textual Features	.58	.55	.57	51
DT + Textual Features	.59	.59	.59	52

Using the above ML classification techniques, we comprehend the significance of different feature types. We demonstrated the results of various characterizations with various proportions of four features. The outcome shows that, the best performing model is Decision Tree (DT). In addition, for all of the features precision, recall and F-measure calculation, DT gives the most outstanding outcome relating to the class of mental disorder indicative of Facebook users.

Chapter 5

Discussion

This thesis presented a construct design for informing the importance of the mental health issues (depression, anxiety, bipolar etc) and possible solution's provisions so a fully-functional decision support application in future can be developed. We promoted an idea of using social network data to assess potential of the factors affecting mental health of Facebook users and machine learning based solution design. We discussed the performance of different ML classifiers in detecting depressive users in a shorter time. The experiment was conducted using MATLAB 2016b. We applied four major ML classifiers such as Decision trees, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Ensemble as mentioned earlier.

We can see the experiment 1 and 2 yielded mixed results suggesting that depression can be judged from Facebook user post and comments. They identified several moderators of accuracy that suggested depression could be assessed from user post and comments. The results from these experiments suggest that depression can be judged when only viewing status updates.

For a better understanding of the general intuition behind depression, in this thesis, we showed that all of these classification techniques based on psycholinguistic (linguistic style, emotional process, temporal process) and textual features are able to successfully extract the depressive emotional result. Table 4.2, 4.3, 4.4 and 4.5 demonstrate the results of various characterizations with various proportions. It can be observed that Decision Tree gives the better outcome. We believe that the current study has laid the ground for future research on inferences and discovery of additional information based on cause-event relation, such as detection of implicit emotion or cause, as well as prediction of public opinion based on cause events, etc.

5.1 Theoretical Implication for Detecting Depression

Depression is a highly evaluative trait, that is, a trait which people are likely to hide from others. In addition, I have argued that impression management processes occur more in public contexts compared to private contexts. Following from these ideas, depression should be easier to detect in private contexts compared to public contexts.

Prior research suggests that depressive symptoms are hard to detect from social interactions [51], however these data demonstrate that a special form of social interactions (e.g. social networking sites) make depression more easily detectable. Social networking sites contain both elements of a public environment (since posted information is available to a network of friends) and elements of a private environment (with the interactions occurring over computer mediated communication) which makes this a unique environment for studying social interactions. It is possible that this type of computer mediated communication creates a sense of security and privacy which allows information relevant to detecting depression to emerge. Because people are not communicating face-to-face in these mediums, they worry less about immediate feedback from their social network and have the chance to formulate and edit the content they post. Theoretically, this creates an environment that is subjectively private yet is objectively public in its purpose or function.

Another important aspect of social networking sites is the degree to which thoughts and feelings are expressed. According to the DSM depressive symptoms manifest themselves primarily in a person's thoughts and feeling (e.g. thoughts of worthlessness, suicide, and inability to concentrate). The results of these studies suggest that the post and comments contains thoughts and feelings which are relevant to detecting depression. Interestingly, the amount of information did not influence how well depression was detected. I argue that it is the longitudinal component of this depressive information which makes easier to detect. In essence, it allows for a baseline positive/negative mood to be established and makes it easier to interpret long and short term changes. In addition, it finds the future levels of depression can be accurately judged. This idea could be further tested by establishing the threshold for changes in mood to determine when a person is becoming depressed. It is possible that sustained negative mood over a short period of time (e.g. one week) is indicative of depression on social networking sites or if the negative mood extends across a longer period of time (e.g. two months).

These studies also provide important evidence that language use is related to depression. Across both experiment 1 and 2, using less positive emotion words and more unhappy related words were related to increased levels of depression. This is consistent with prior research and theories on the role of language use and depression. Experiment 2 also provided additional evidence that other language cues (swear words, negative emotions, meta-physical / religious words) were related to depression. Theoretically, these findings align with prior research linking language and depression.

5.2 Practical Implication for Detecting Depression

The results of these studies raise questions about the practicality of using social networking sites as a means to detect depression. The signal detection analyses point to the practicality of how good strangers are at identifying someone who is depressed from looking at the person's social networking profile. First and foremost, these analyses show that raters often misclassify people who are depressed as being not depressed (i.e. have a low number of true positives and a high number of false negatives). In the current studies, the depressive signal detection analyses were conducted by LIWC and BOW. Practically, this may not be a good cutoff point because people are generally conservative in rating a person as depressed. In support of this idea, in the current studies, the mean accuracy of depression detection was below. The depressive signal detection analyses revealed that raters were very good in detecting people who were non-depressed (i.e. a high number of true negatives and low number of false positives). Overall, these analyses suggest that it is more likely that someone will be classified as non-depressed rather than depressed. These analyses also point to the importance of knowing which pieces of information on social networking profiles are valid cues to depression (e.g. language cues) in order to "zoom" in attention to these diagnostic cues to detect depression.

For friends and family, using social networking sites may be a practical and easy way to easily "check-up" on friends without directly asking the person how they are doing. Since depression is something which people do not readily admit to or discuss freely in everyday life, social network sites may contain other indirect clues to detecting depression. In fact, social networking sites are constructed in a way that allows users to view the activity on their friends' pages. For example, in Facebook, users immediately see a page entitled "news feed" which alerts them to friends' recent activity. Consistent with experiment 2's findings that status updates allow for the detection of depression, one suggestion is to have Facebook users set their Facebook homepage to see only status updates compared to seeing all the activity that is occurring on a friend's Facebook page. This setting allows users to filter out information which is less relevant to detecting depression by focusing specifically on status updates.

Experiment 2 also revealed that several individual differences influence that degree of accuracy in assessing depression. It is harder to judge depression in people who are Extraverted or high on Impression Management. It does not mean, however, that it is impossible to detect depression in friends who are Extraverted or score high on Impression Management. Research indicates that other people can detect Extraversion very easily from brief pieces of information, so a person is likely to correctly identify which of their friends are highly Extraverted. When looking for signs of depression, people may have to pay attention to different pieces information when looking at their Extraverted friends, or “read through the lines”, when looking at their profiles. I argue that Extraverted people post more information about social events and have more social interactions on social networking sites, and based on prior research, the number of social interactions is unrelated to depression. Instead, I argue that depression is related to information related to expressing thoughts and feelings (e.g. status updates, blogs) and users should focus on this information – especially for their extraverted friends.

However, it is harder (if not impossible) to identify friends who score high on Impression Management. People who are dispositionally high on this trait naturally hide information from other people. This is more prevalent though in highly public settings where information is displayed for the world to see [50]. Since the majority of information on social networking sites is available for other people to see, one possibility to gauge depression is to send a message to a friend using the internal messaging system on these sites. This approach capitalizes on creating a truly private environment in which people high on Impression Management may feel less compelled to self-present. These types of messages are confidential between people and it may be more likely a person high in Impression Management might confide in a friend in this one-on-one conversation.

For clinicians, social networking sites may also be helpful addition to diagnosing and treating depression. Social networking sites could be used in conjunction with traditional clinical interviews or recommended to friends and family of depressed patients to use to spot signs of a depressive episode. Friends and family could provide not only their own insight in helping someone cope with depression, but also serve as informants to clinicians as to how depression is manifested on social networking sites.

Chapter 6

Conclusion

Depression is a mental disorder that occurs in many Facebook users. In this thesis, we exhibited the capability of using Facebook as a source for measuring and detecting major depression among its users. First, with the depression and non-depression datasets as well as well-defined discriminative depression-oriented feature groups, we proposed architecture and demonstrated the potential of using LIWC and BOW as a tool for measuring major depression in individuals. We then analyzed the contribution of the psycholinguistic and textual feature modalities and detected depressed users on a large-scale depression-candidate dataset to reveal some underlying online behaviors between depressed users and non-depressed users on Facebook. Our findings showed that individuals with depression show lowered social activity, greater negative emotion, high self attentional focus, increased relational and medicinal concerns, and heightened expression of religious thoughts. Since online behaviors cannot be ignored in modern life, we expect our findings provide more perspectives and insights for depression researches in computer science and psychology. Finally, we leveraged these distinguishing attributes to build classifiers that can detect depression. The classifiers yielded promising results with approximate (65-70) % classification accuracy.

6.1 Future Work

In future work, we will extend the literature in the area of growing attention called 'Mental Health Mining', in particular in the health related knowledge discovery. The suitability of mental health illness data for disease surveillance will be demonstrated along with its potential limitations in a case study of mental health illness in all around the world. Additionally, we will use further techniques to extract paraphrases from more types of mental disorder features. Also, we will be using more dataset to verify the efficiency and effectiveness of ML techniques to subsequently improve the predictive accuracy. This will prove the way for a larger scale investigation into the topic in the upcoming research. As a result, the benefits of high quality input data as well as the innovative visualization tool for patterns exploration will considerably facilitate knowledge discovery. Finally, the literature on social media data mining for mental health applications will be extended with a new practical application to the current matter of concern such as depression, anxiety and bipolar growing prevalence in the world.

Bibliography

[1] Akkın Gürbüz, H.G., et al., Use of social network sites among depressed adolescents. Behaviour & Information Technology, 2017. **36**(5): p. 517-523.

[2] Zafarani, R., M.A. Abbasi, and H. Liu, Social media mining: an introduction. 2014: Cambridge University Press.

[3] Bachrach, Y., et al. Personality and patterns of Facebook usage. in Proceedings of the 4th Annual ACM Web Science Conference. 2012. ACM.

[4] Zhou, J., et al. Measuring Emotion Bifurcation Points for Individuals in Social Media. in System Sciences (HICSS), 2016 49th Hawaii International Conference on. 2016. IEEE.

[5] Yin, J., et al., Using social media to enhance emergency situation awareness. IEEE Intelligent Systems, 2012. **27**(6): p. 52-59.

[6] Weedon, J., W. Nuland, and A. Stamos, Information operations and Facebook. version, 2017. **1**: p. 27.

[7] Conway, M. and D. O'Connor, Social media, big data, and mental health: current advances and ethical implications. Current opinion in psychology, 2016. **9**: p. 77-82.

[8] Greenberg, L.S., Emotion-focused therapy of depression. Person-Centered & Experiential Psychotherapies, 2017: p. 1-12.

[9] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr, "Social Media & Mobile Internet Use among Teens and Young Adults. Millennials," *Pew internet & American life project,* 2010.

[10] N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of computer‐mediated Communication,* vol. 13, pp. 210-230, 2007.

[11] E. Hargittai and Y.-l. P. Hsieh, "Predictors and consequences of differentiated practices on social network sites," *Information, Communication & Society,* vol. 13, pp. 515-536, 2010.

[12] R. Lynn, Learning to like Facebook? Effects of cultural and educational capital on the use of social network sites in a population of university students: University of Missouri-Saint Louis, 2009.

[13] S. Phillips, "A brief history of Facebook," *the Guardian,* vol. 25, 2007.

[14] D. Fletcher, "How Facebook is redefining privacy," ed: Time, 2010.

[15] A. L. Parsons, "Social media from a corporate perspective: A content analysis of official Facebook pages," in *Allied Academies International Conference. Academy of Marketing Studies. Proceedings*, 2011, p. 11.

[16] M. Haskins, "A quick guide to facebook. com," *Retrieved December,* vol. 18, p. 2005, 2005.

[17] A. N. Joinson, "Looking at, looking up or keeping up with people?: motives and use of facebook," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2008, pp. 1027-1036.

[18] K. Lewis, J. Kaufman, and N. Christakis, "The taste for privacy: An analysis of college student privacy settings in an online social network," *Journal of Computer‐Mediated Communication*, vol. 14, pp. 79-100, 2008.

[19] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007, pp. 29-42.

[20] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th Annual ACM Web Science Conference*, 2013, pp. 47-56.

[21] I. Choi, D. N. Milne, N. Glozier, D. Peters, S. B. Harvey, and R. A. Calvo, "Using different Facebook advertisements to recruit men for an online mental health study: engagement and selection bias," *Internet Interventions,* vol. 8, pp. 27-34, 2017.

[22] S. Joyce, M. Modini, H. Christensen, A. Mykletun, R. Bryant, P. B. Mitchell, *et al.*, "Workplace interventions for common mental disorders: a systematic meta-review," *Psychological medicine,* vol. 46, pp. 683-697, 2016.

[23] M.-P. Austin, P. Mitchell, and G. M. Goodwin, "Cognitive deficits in depression: possible implications for functional neuropathology," *The British Journal of Psychiatry,* vol. 178, pp. 200-206, 2001.

[24] R. Purcell, P. Maruff, M. Kyrios, and C. Pantelis, "Neuropsychological function in young patients with unipolar major depression," *Psychological medicine,* vol. 27, pp. 1277-1285, 1997.

[25] D. A. King, C. Cox, J. M. Lyness, and E. D. Caine, "Neuropsychological effects of depression and age in an elderly sample: A confirmatory study," *Neuropsychology,* vol. 9, p. 399, 1995.

[26] M. A. Butters, E. M. Whyte, R. D. Nebes, A. E. Begley, M. A. Dew, B. H. Mulsant, *et al.*, "The nature and determinants of neuropsychological functioning in late-lifedepression," *Archives of general psychiatry,* vol. 61, pp. 587-595, 2004.

[27] J. Chodosh, D. M. Kado, T. E. Seeman, and A. S. Karlamangla, "Depressive symptoms as a predictor of cognitive decline: MacArthur Studies of Successful Aging," *The American journal of geriatric psychiatry,* vol. 15, pp. 406-415, 2007.

[28] N. Sachs-Ericsson, T. Joiner, E. A. Plant, and D. G. Blazer, "The influence of depression on cognitive decline in community-dwelling elderly persons," *The American journal of geriatric psychiatry,* vol. 13, pp. 402-408, 2005.

[29] I. Chi and K.-L. Chou, "Depression predicts cognitive decline in Hong Kong Chinese older adults," *Aging & Mental Health,* vol. 4, pp. 148-157, 2000.

[30] Wee, J., et al., The influence of depression and personality on social networking. Computers in Human Behavior, 2017. **74**: p. 45-52.

[31] Haberler, G., Prosperity and depression: A theoretical analysis of cyclical movements. 2017: Routledge.

[32] Aldarwish, M.M. and H.F. Ahmad. Predicting Depression Levels Using Social Media Posts. in Autonomous Decentralized System (ISADS), 2017 IEEE 13th International Symposium on. 2017. IEEE.

[33] De Choudhury, M., et al., Predicting Depression via Social Media. ICWSM, 2013. **13**: p. 1-10.

[34] G. S. O'Keeffe and K. Clarke-Pearson, "The impact of social media on children, adolescents, and families," *Pediatrics,* vol. 127, pp. 800-804, 2011.

[35] M. H. Selfhout, S. J. Branje, M. Delsing, T. F. ter Bogt, and W. H. Meeus, "Different types of Internet use, depression, and social anxiety: The role of perceived friendship quality," *Journal of adolescence,* vol. 32, pp. 819-833, 2009.

[36] J. Davila, C. B. Stroud, L. R. Starr, M. R. Miller, A. Yoneda, and R. Hershenberg, "Romantic and sexual activities, parent–adolescent stress, and depressive symptoms among early adolescent girls," *Journal of adolescence,* vol. 32, pp. 909-924, 2009.

[37] J. J. Block, "Issues for DSM-V: Internet addiction," ed: Am Psychiatric Assoc, 2008.

[38] I. Stepanikova, N. H. Nie, and X. He, "Time on the Internet at home, loneliness, and life satisfaction: Evidence from panel time-diary data," *Computers in Human Behavior,* vol. 26, pp. 329-338, 2010.

[39] J. P. Harman, C. E. Hansen, M. E. Cochran, and C. R. Lindsey, "Liar, liar: Internet faking but not frequency of use affects social skills, self-esteem, social anxiety, and aggression," *CyberPsychology & Behavior,* vol. 8, pp. 1-6, 2005.

[40] M. A. Moreno, D. A. Christakis, K. G. Egan, L. A. Jelenchick, E. Cox, H. Young*, et al.*, "A pilot evaluation of associations between displayed depression references on Facebook and self-reported depression using a clinical scale," *The journal of behavioral health services & research,* vol. 39, pp. 295-304, 2012.

[41] R. L. Baskerville, M. Kaul, and V. C. Storey, "Genres of Inquiry in Design-Science Research: Justification and Evaluation of Knowledge Production," *Mis Quarterly,* vol. 39, pp. 541-564, 2015.

[42] R. H. Von Alan, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS quarterly,* vol. 28, pp. 75-105, 2004.

[43] https://www.kaggle.com/osmi/mental-health-in-tech-2016

[44] https://github.com/ranju12345/Depression-Anxiety-Facebook-page-Comments-Text

[45] Scott, J., Social network analysis. 2017: Sage

[46] Bazeley, P. and K. Jackson, Qualitative data analysis with NVivo. 2013: Sage Publications Limited.

[47] AlYahmady, H.H. and S.S. Alabri, Using NVivo for data analysis in qualitative research. International Interdisciplinary Journal of Education, 2013. **2**(2): p. 181-186.

[48] Bandara, W. Using Nvivo as a research management tool: A case narrative. in Quality and Impact of Qualitative Research: Proceedings of the 3rd International Conference on Qualitative Research in IT & IT in Qualitative Research. 2006.

[49] Tausczik, Y.R. and J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 2010. **29**(1): p. 24-54.

[50] McDougall, M.A., et al., The effect of social networking sites on the relationship between perceived social support and depression. Psychiatry research, 2016. **246**: p. 223-229.

[51] Serrat, O., Social network analysis, in Knowledge solutions. 2017, Springer. p. 39-43.

Publications:

Journal

[1] Md Rafiqul Islam, Ashad Kabir, Ashir Ahmed and Abu Raihan M. Kamal (2018). Depression Detection from Social Network Data. Health Information Science and Systems**. (Springer) (Accepted).**

[2] Md Rafiqul Islam, Shah Miah, and Abu Raihan M. Kamal (2018). A Design Construct for Developing Approaches to Measure Mental Health Conditions: A design research perspective. Australasian Journal of Information Systems. **(Elsevier) (Accepted).**

Conference

[1] Md Rafiqul Islam, Abu Raihan M. Kamal, Naznin Sultana, Mohammad Ali Moni, Robiul Islam (2017). Depression Detection Using K-Nearest Neighbors (KNN) Classification Technique. International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering. February 8-9, 2018, Rajshahi, Bangladesh. **(IEEE)**