Fall 12-13-2021

Analyzing Tweets For Predicting Mental Health States Using Data Mining And Machine Learning Algorithms

SUDHA TUSHARA Sadasivuni

Recommended Citation

Sadasivuni, SUDHA TUSHARA, "Analyzing Tweets For Predicting Mental Health States Using Data Mining And Machine Learning Algorithms." Dissertation, Georgia State University, 2021.
doi: https://doi.org/10.57709/26633473

ANALYZING TWEETS FOR PREDICTING MENTAL HEALTH STATES USING DATA MINING AND MACHINE LEARNING ALGORITHMS

by

Sudha Tushara Sadasivuni

Under the Direction of Yanqing Zhang PhD

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2021

ABSTRACT

Tweets are usually the outcome of peoples' feelings on various topics. Twitter allows users to post casual and emotional thoughts to share in real-time. Around 20% of U.S. adults use Twitter. Using the word-frequency and singular value decomposition methods, we identified the behavior of individuals through their tweets. We graded depressive and anti-depressive keywords using the tweet time-series, time-window, and time-stamp methods. We have collected around four million tweets since 2018. A parameter (Depressive Index) is computed using the F1 score and Mathews correlation coefficient (MCC) to indicate the depressive level. A framework showing the Depressive Index and the Happiness Index is prepared with the time, location, and keywords and delivers F1 Score, MCC, and CI values.

COVID-19 changed the routines of most peoples' lives and affected mental health. We studied the tweets and compared them with the COVID-19 growth. The Happiness Index from our work and World Happiness Report for Georgia, New York, and Sri Lanka is compared. An interactive framework is prepared to analyze the tweets, depict the happiness index, and compare it. Bad words in tweets are analyzed, and a map showing the Happiness Index is computed for all the US states and was compared with WalletHub data. We add tweets continuously and a framework delivering an atlas of maps based on the Happiness Index and make these maps available for further study.

We forecasted tweets with real-time data. Our results of tweets and COVID-19 reports (WHO) are in a similar pattern. A new moving average method was presented; this unique process gave perfect results at peaks of the function and improved the error percentage.

An interactive GUI portal computes the Happiness Index, depression index, feel-good-factors, prediction of the keywords, and prepares a Happiness Index map. We plan to create a public web portal to facilitate users to get these results. Upon completing the proposed GUI application, the users can get the Happiness Index, Depression Index values, Happiness map, and prediction of keywords of the desired dates and geographical locations instantaneously.

INDEX WORDS: Tweets, Mental Health, Hedonometrics, Data Mining, Machine Learning

ANALYZING TWEETS FOR PREDICTING MENTAL HEALTH STATES USING DATA MINING AND MACHINE LEARNING ALGORITHMS

by

Sudha Tushara Sadasivuni

Committee Chair: Yanqing Zhang

Committee: Xiaojun Cao

Raj Sunderraman

Ying Zhu

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

December 2021

DEDICATION

My doctoral degree has always been my dream, and I couldn't have achieved it without the unconditional love I received from my family.

My dad, Dr. Lakshminarayana, has always been my inspiration. He trusted in my abilities even when there were times that I could not. His faith and belief made me who I am today. My mom, Aruna Kumari, helped to motivate me. In fact, she used to listen to all the research work I was doing and encouraged me more. My Uncle, Prasad, and Radhika aunty supported me emotionally and were very determined to make me cross all the hurdles. My shoulder of faith and belief has always been my brother, Sudarsan. He always stood by me in all the stages of life. I would also like to thank all my friends who supported and encouraged me.

ACKNOWLEDGEMENTS

It is my genuine pleasure to thank Dr. Yanqing Zhang for his invaluable guidance in my research. Dr. Zhang inspired me with his immense wealth of knowledge and perspective towards life. I had the freedom to experiment with new ideas and always had a positive approach to my work. Dr. Zhang always heard my views with patience and suggested the best ways to forward.

I would also like to thank my committee members: Dr. Raj Sunderaman, who continuously motivated me with my work and career. I would also want to thank Dr. Xiaojun Cao and Dr.Ying Zhu for their implicit feedback, time, and effort. My work was extremely advanced from the feedback of all the committee members.

I would also like to thank the entire Computer Science department and Neuroscience Institute of Georgia State University for granting the Brains and Behavior (B&B) Fellowship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACF	Auto Correlation Function
API	Application Programming Interface
AUC	Area Under the Curve
B2B	Business to Business
B2C	Business to Consumer
COVID-19	Corona Virus Disease 2019
DI	Depression Index
GUI	Graphic User Interface
FOMO	Fear Of Missing Out
HI	Happiness Index
KCF	Keyword Contributing Factor
KNN	k- nearest neighbors
LDA	Linear Discriminant Analysis
MCC	Mathews Correlation Coefficient
MUNSH	Memorial University of Newfoundland Scale of Happiness
PACF	Partial Auto Correlation Function
PCC	Pearson Correlation Coefficient
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TDM	Term Document Matrix
WHO	World Health Organization
WHR	World Happiness Report

PREFACE

This work is the outcome after understanding that "Mental illness is reported as one of the leading diseases in coming years to be addressed by the stakeholders." To understand, identify a mental illness person, we need to follow their activities. Social networks have the power to attract human feelings, emotions, and expressions. In our words, social network posts are the remanences and direct consequence of ones' mind on the situation, time, and location.

Tweets from Twitter are believed to be the output of human feelings, and we considered them as input for our study. More than 2.3 million tweets related to depression are taken from Kessler's works, and anti-depression words are taken as base data. Using the word-frequency method, time-series, time-window, time-stamping methods, these tweets were analyzed. Depressive Index, a new parameter, is exhumed that identifies the levels of depression in comparison with other areas with time. Clustering of these words showed an interrelationship among these tweet keywords. The impact of tweets on Mental health is calculated with an example of Space tourism tweets. Hedonometrics, a study of happiness, is relooked with the tweets, as we hypothesize that people's posts are the outcome of the human mind's behavior. The happiness index is computed using the tweets. Revisiting the simple moving average method, we could get better accuracies in forecasting.

In truth, I could not have achieved the current level of success without solid support: my parents, uncle, and brother, who supported me with love and understanding. And secondly, my committee members, each of whom has provided patient advice and guidance throughout the research process. Thank you all for your unwavering support.

1. INTRODUCTION

1.1 Problem Statement

Mental illness is observed through the actions/behavior of an individual, emotions, and expressions towards a situation. American Psychiatric Association indicates that 19% of people experience some form of mental illness. Nearly 4.1% of people are seriously affected by mental Illness. [1]. In 2001 the WHO reported that 450 million people suffer from mental disorders. With the technological growth, and affordable internet access, Social Media usage and impact increased in society.

Social media facilitated people to express their feelings in the public domain. Participating in and sharing information on social media became a daily routine for many people. Around 42% of people participate in Twitter postings at least once a day. Such participation created large volumes of data. This data has geographical location, timestamp and leaves markers of the individual who generated the content. Individuals are also habituated to mark the social media site addresses in all their communications to invite others to follow, view, and comment. Such intensified growth allowed people to know about others, communicate, and offer suggestions. In view of many people who are directly and indirectly suffering from mental illness, it is essential to study the methods that help in the identification of mental illness, track and predict the emerging mental illness strategies.

1.2 The early history of depression

United States mental illness history illustrates how psychiatry and cultural understanding of mental illness influence national policy and attitudes towards mental health. During the stone age, people used to drill holes into the skull to get rid of evil spirits, which was believed to cause mental illness. Around 400 B.C., Hippocrates treated the mentally suffered people, assuming

mental illness was an issue to the human body rather than a God's punishment. Bethlem Hospital opened in 1377 in London for mental illness people. During 1600, imprisonment was levied as a control mechanism for the mentally ill. Asylums started in 1850, and an experimental psychology lab was set up at the University of Leipzig in Germany in 1879. Around 1920, modern treatment was introduced for stress disorders. In 1938 Electro-shock therapy for schizophrenia and depression was a method to treat mental illness people. Thorazine was discovered in 1952 for psychosis. Behavior therapy was started in 1950, and new generation drugs were found in the 1980s.

In 1840 Dorothea Dix tried to extend better living conditions for mentally depressed people, and during the late 18th century, the USA built 32 psychiatric hospitals. Later, people were treated out-patient method. During 1950 there were 560,000 patients reported, whereas this number was reduced to 130,000 by 1980 [2]. Clifford founded Mental Health America in 1909. National Mental Act was passed in 1946 and allocated funds towards research in this direction. Support for education, advocacy was started, and government programs with welfare events have improved mental health care. There were 339 beds for 100,000 in 1955 for treating mental illness people, whereas this number reduced to 22 in the year 2000.

Depression is a common mental illness, and the visible symptoms are staying in a sad mood, losing interest in daily activities that are pleasurable. People also noticed weight gain or loss, fatigue, struggle in concentrating, and feelings of self-destruction. Depression causes problems at the workplace and in relationships. Depression has also been correlated positively with adverse health behaviors, including smoking, alcohol abuse, physical inactivity, and sleep disturbances. People also opined that poor nutrition, stressful events, poverty, and war could be a few factors for depressive behavior in general.

1.3 Motivation

Around 40-45% of people use Social Networks and spend a lot of time. This tendency of use of Social Networks is increasing day by day in human life. It has now come to a stage that an instance, function, event, or feeling is not finished until its details are posted on social networks. There is a lot of swings in people's behavior in these areas. It is found that there are several adverse effects of using social media. The increase of depression, anxiety, cyberbullying, Fear Of Missing Out (FOMO), unrealistic expectations, unhealthy sleep patterns, increased negative emotions, and general addition is severe impacts of social media use. But at the same time, it is wrong to conclude that social media is a bad idea because it has benefits to add to our lives.

Social Media is a new phenomenon that allows social interaction between people known/unknown using Internet-based applications. The recent increase in suicides and the Internet's role, mainly social media, draws attention to many researchers. It is not established how much social media contributes to an increase in suicidal behavior in human beings. Still, there is certainly a positive correlation between suicides and social media. The internet also provided ways to know the suicide descriptions and lethal means to kill themselves. Social networking sites can also facilitate ways to avoid such depressive thoughts by providing connections among peers of similar experiences and getting help from society. More research is needed to understand the role of social network influence in suicidal attempts.

Identifying a depressive individual is easy if we associate for a longer time and follow the actions. But in today's world, people are busy with their works. We have thus chosen the tweets of an individual, which is one of the feeling *outcomes* of a person, can be easily tracked in social networks and analyzed for its characteristics. Any attempt to develop means and methods that help or reduce suicides is a positive contribution to humanity. We have thus chosen this area of research

to contribute a few new techniques and algorithms. Using the word-frequency method, singular value decomposition method, time series, time window, and time stamp methods, we analyzed the depressive and anti-depressive tweets. An attempt is made to categorize the tweet keywords with these tweets.

We also scaled the depression levels, clustered the depressive keywords, and studied the impact on mental health. We extended our studies to find the Happiness Index of a geographical area, a crucial parameter in Hedonometrics. Also, we attempted to analyze the tweets at a location, country and delineate the degree of depressive Index of the people on that day. In the end, we implemented a new moving average method to forecast the depressive and anti-depressive tweets.

The present growth of mental illness among the children, students, youth, families, and older citizens motivated to contribute a few new methods for detecting mental illness in society, adding the computing parameters of Hedonometrics, and expanding new tools for better living. The support from governments towards this cause by instituting many units and national suicide prevention helplines (1-800-273-8255), crisis text line (74174), Trevor lifeline (1-866-488-7386), Trans lifeline (1-877-565-8860), and many more [3] motivated my thoughts to focus in this direction.

1.4 Outline

The ream of the dissertation is structured as follows. Section 2 presents the past related works along with a review of literature for the relevant area. Section 2 reports several methods applied in our study. Section 3, the data collection and cleaning methods are described.

In section 4, the analysis of tweets is described. Section 4.3.1 describes how analyzing Tweets using word frequency could identify the Twitter user's mental health status. This section 4.3.2 presents the results using the Singular Value Decomposition (SVD) method and the chances

of individuals tweeting at least three of the other keywords if they are mentally depressed. Time Series, Time window, and Time stamp studies were described in sections 4.3.3 to 4.3.5. Classification of the tweet keywords is done and stated in section 4.3.6. It is also witnessed that the tweeting pattern of 'tired', 'restless' exhibited a different pattern to other depressive keywords. The anti-depressive tweets followed the pattern of 'failure', 'hopeless', 'nervous', and 'worthless' depressive tweets. Section 4.3.4 describes the results that are obtained within the depressive and anti-depressive tweet keywords within ten minutes of interval and beyond.

In Section 5, a new parameter with F1-Score and Matthew's Correlation Coefficient is formed, and the outcomes are illustrated. Section 6 describes the clustering method results with the tweets obtained during the Sri Lanka Bomb blasts. Section 7 describes the impact of depressive and anti-depressive tweets on mental health. Space tourism tweets for a month were collected are analyzed to delineate the results.

Hedonometrics, a branch of the study of happiness, was discussed in section 8, where we revisited a new parameter, 'happiness index,' using tweets. Section 8.4 deals with the study of the happiness index and depressive index using bad words in tweets.

Section 9 describes the simple moving average methods, weighed moving averages, and exponential smoothing. We came out with a new process of moving averages, which gave better results than existing methods. These results are applied using the ARIMA model on the COVID-19 epidemic data set for forecasting the tweets. Section 10 describes the conclusions. Whereas Section 11 illustrates the future works.

2 BACKGROUND WORK

2.1 Background Work

It is estimated that by the year 2020, depression will become the second leading cause of disease burden, and depression and schizophrenia are the main reasons for most suicides due to psychiatric disorders [4], [5], [6].

De Choudhury et al. [7] Park, M et al., [8], Nadeem, M. et al. [9], and many other computational social scientists worked in this area to predict the levels and identify depression using Twitter postings. Gwynn, R. C. et al. estimated that more than 45% of major depression cases are undiagnosed [10]. Jianhong Luo et al. [11] suggested that suicidal prevention can be achieved through a systematic assessment of the behavior from Twitter posts. Suicide prevention to mental disorder individuals can thus be of benefiting society at large. The complex problem is to identify the individual with mental illness. People habited to share information using social media, and this became a communication platform. People discuss many issues in social media like politics [12], disasters [13]. Of late, people are sharing health tips, success stories, and help peers [14]. Scanfeld et al. [15] mentioned that social media sites offer a means of health information sharing. Seeman [16] mentioned that web surveys reveal mental illness and the size of the problem. Mental disorders impact the global economy and are estimated at the cost of US $ 2.5 trillion in 2010 and are estimated to be the US $6 trillion by 2030 [17]. Given these complicated issues, any attempt the study and address mental health illness will help the human at large.

Identification and research leading development for better health among the people is the need of the hour as one in five experience mental illness. New events are detected from tweet streams [18]. Dou et al. [19] worked on event detection, event tracking, and event association in

streaming data. Zhu and Laptev [20] studied deep and confident predictions for time series data and presented an end-to-end neural network architecture for uncertainty estimation. Sato, Junbo Wang, and Zixue Cheng [21] used extended Hybrid TF-IDF and Remarkable Word Detecting Methods to quantify the importance of words accurately and evaluate the quantified values dynamically. Radoslaw Michalski, Kazienko, and Dawid [22] [23] applied machine learning techniques to determine social network prediction's suitability using the time series forecasting and classification approaches.

Doulamis, Anastasios, Kokkinos, and Varvarigos Doulamis used a multi-assignment graph partitioning algorithm for event detection in Twitter Micro-blogging [24]. Several scientists also worked in semantic reasoning and event classification [25], and average group clustering methods [26] for event detection. Ishikawa, Y. Arakawa, S. Tagashira, and A. Fukuda used burst detection in the data for event detection [27]. Prediction of future events using the social media data was also attempted [28] [29]. Researchers also endeavored to use social media to identify and predict mental issues [30]. Resnik, A. Garron, and R. Resnik used topic modeling to improve the prediction in depression [31]. Choudhury, Gamon, Counts, and Horvitz [32], Tushara, and Zhang [33] attempted to discover Twitter users' mental health status by word-frequency method.

Social media is a cheap and economical way to share information with present-day technology. People communicate using social media about wellbeing suggestions, success tales, and assist colleagues [34]. With the technological growth and affordable Internet access, Social Media usage and its impact on daily life increased. Park [35] Sho et al. [36], and many other computational social scientists worked in this area to predict the levels and identify depression using Twitter postings.

Many researchers [37] [38] [39] have reported a direct link between Social Media usage and Mental Health issues. Tushara and Zhang [40] [41] identified individuals' depressive Twitter and behavior through the tweets and the relation between Illness and tweets. Spending more time on social media is likely to be depressed more than non-users [42]. Studies revealed that Internet usage directly impacted happiness and observed that heavy internet users are twice as unhappy as others [43]. The depressed people recognized a higher ratio of negative to neutral words. Earlier works reported that depressed participants took longer than controls to remember neutral words but did not differ in response times to negative words [44]. Suicide prediction tools are beneficial to relatives and friends of an individual, so that intervention of a Mental Health Specialist will be placed to address the issue on time. Hence the Social Media data is a high value for machine learning and data mining research. Evaluation Performance is one of the critical parameters in any model, and Data Mining is not exceptional. Many researchers are using this value as one of the parameters for the assessment of data. Classification Accuracy, logarithmic loss, confusion matrix, Area under Curve, Precision, Recall, F-1 Score, and Matthews Correlation Coefficient (MCC) are a few indicators that researchers are working on depicting and interpreting the data [45]. In case there are more tweets related to ‘fire,’ we can assume that there could be a fire accident. If people tweet with words related to happiness, fun, joy, excitement, hope, and delight, etc., we can conclude that people are happy in that area/time. More people tweet with depression, failure, hopelessness, and tired or similar words; we can imagine that there is unhappiness around. Many researchers worked on the ‘Happiness Index.’ Lane et al. [46] Rabbi, Ali, Choudhury, and Berke [47] mentioned that mental health is one of the criteria for happiness inhuman. Marcelo et al. [48] study Twitter users’ response methods and behavior during an earthquake. Forest fire incidents were discovered through tweets [49]. Hossny and Mitchell [50] works detected the

events with word counts. Liu et al. [51] worked with Twitter tweets and found a method to discover the smaller-scale local events. Florian and Antal [52] have found indications to recognize events using the relative frequency of tweets. Yasuyuki et al. [53] found depression in humans after the Nagasaki bombing. Yasmin and Maria [54] studied the level of anxiety, depression, and stress after the bombing. Depressed and Anti-depressed tweets were analyzed to cluster based on Twitter events reported during Sri Lanka Bomb Blasts [55]. These studies indicate that Twitter users post information related to current events. Stephen and Paris [56] analyzed the tweets during Syndey Siege (December 201. They mentioned that these emotional tweets expressed on social media help understand the reactions and why the emotional reactions occur as they do. Sykora et al. [57] studied several datasets and their relationship with different events.

There were no studies earlier reported to categorize the emotional keywords in association with the event. We attempted to analyze the tweets data collected during the Srilanka bombing attacks on 21st April 2019. There were a few hundred human deaths reported during these bomb attacks, and hence certainly, there will be considerable emotions in people around the geographical areas. Studies by Haewoon et al. [58] revealed that more than 85% of tweets are related to current news. We attempted to study the tweets during a selected period to correlate depression/anti-depression tweets with bombing tweets. We categorized these tweet keywords concerning the event. In recent months, the COVID-19 spread has negatively affected so many people's mental health. According to a recent KFF poll, 45% of adults' mental health is impacted due to stress over COVID-19 in the United States [59]. The longer COVID-19 exists and people are isolated at home, the more likely they have mental problems due to anxiety, depression, fear, pressure, etc. Quantitative studies about understanding the depression changes were carried out on quarantined people [60]. Similar studies revealed that health professionals during the SARS outbreak had

depression, anxiety, fear, and frustration [61]. Socioeconomic distress is also a reason for psychological disorders and anxiety [62]. Social media sites offer a means of sharing health information [15]. Since the World Health Organization (WHO) determined the outbreak of novel coronavirus disease, COVID-19, to be a "public health emergency of international concern," the stress levels everywhere have continued to mount [63]. We developed a method to predict people's future mental health status during the spread of COVID-19 and then provide an early warning with psychological problems and other people with anxiety, depression, fear, and pressure. In addition to tweets, the COVID-19 data, such as daily confirmed COVID-19 cases and daily mortality rates [64], are also used to build the machine learning-based time series prediction system.

Stigma [65], lack of education [66] , limited availability of Mental health professionals, limited affordability [67], and policy limitations [68] are regarded as a barrier to mental illness identification and care. World Health Organization (WHO) recommended effectively developing methods to invest in mental care, add more workforce, and adopt best practices and human rights protection to overcome mental disorders [69].

Mental illness is a leading cause of disability worldwide. It is estimated that nearly 300 million people suffer from depression (World Health Organization, 2001). Reports on lifetime prevalence show high variance, with 3% reported in Japan to 17% in the US. In North America, the probability of having a major depressive episode within one year is 3–5% for males and 8–10% for females. However, major governments allocate funds towards the Mental Health improvement programs, more than 28% of governments [70]. Nearly one in six children and teenagers admitted to psychiatric hospitals has an intake diagnosis of depressive illness [71]. The depression rates among children and adolescents are rising [72].

3 DATA COLLECTION AND PROCESSING

3.1 Introduction

We are motivated to study the mental health status of humans and address the challenges using their social network posts. Twitter launched on 15-June-2006, and it is an excellent social microblogging platform that caters few million individuals. Twitter is one social network site chosen for our study and a perfect example of a micro blogging social network site. Twitter connects millions of users with tweets. Users can write about any topic in text format within the 280 characters limit and follow other users on Twitter to get an update. We used the R programming language to get the tweets from the Twitter API. Twitter takes authentication from the user and accessURL, api_key, apisecret, authURL, consumerkey, consumentsecret, my_oauth, requestURL, token, token_secret keys are generated.

Twitter delivers a maximum of 10,000 tweets online at any time of our API request. Twitter API often returns less than the requested number of tweets due to a fixed spanned time frame of twitter policies. The # (hash) tag is the prefix of each keyword that a particular group of people tweets. We represent the data collection diagrammatically as follows:

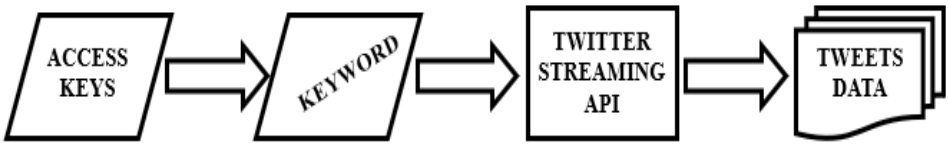

Figure 3.1 Data collection from Twitter using API (Keyword)

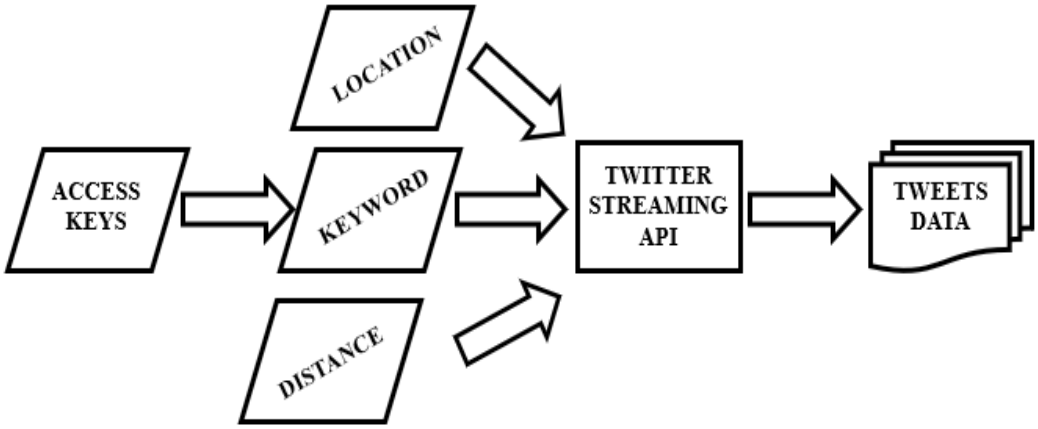

Figure 3.2 Data collection from Twitter using API (Keyword. Location, and distance around the location

3.2 R-Studio

Ross Ihaka and Robert Gentleman in 1990, developed a statistical programming language called 'R' reflecting the creating first names of the creators. RStudio is an open-source integrated development environment (IDE) for data manipulation, calculation, and graphical display. It includes an R console, a code editor, file browser, help files, and graphical display. We used R-Studio for our data collection version 1.2.5001. R-studio was downloaded from https://cran.r-project.org/bin/windows/base/ and installed in the system.

NLP, plyr, ggplot2, wordcloud, wordcloud2, ROAuth, stringR, tm are the libraires that are to be installed before use. In the next section, the code is mentioned to collect for one keyword.

3.3 Methods

The following is the code in R that collects the tweets from twitter.com

```
library(twitteR)
library(ROAuth)
#R interface with OAuth and needs the following keys
```

```
# consumerKey: The consumer key provided by your application

#consumerSecret: The consumer secret provided by your application

#needsVerifier: Whether this OAuth needs the verification step. Defaults to TRUE

#handshakeComplete: Whether the handshaking was successfully completed

#requestURL: The URL provided for retrieving request tokens

#authURL: The URL provided for authorization/verification purposes

#accessURL: The URL provided for retrieving access tokens

#oauthKey: For internal use

#oauthSecret: For internal use verifier: For internal use

#signMethod: For internal use

# Declare Twitter API Credentials from dev.twitter.com

api_key <- "1kHELrmPmtHcfPNPnYW43Xdbz"

apisecret <- "qjF7FSXVfqpCXrlePCINSXeN8J5br5LcIezWgc9xmio7Duf9fG"

token <- "874554929982472961-rQJDJHovgLgZtYfApewO0hCc67UbYTN"

token_secret <- "Ph3K6hUke9XBi5MYM3AE5TzATzsav1QDlLROdBOBfvtvs"

#Create Twitter Connection

setup_twitter_oauth(api_key, apisecret, token, token_secret)

library(NLP)

library(plyr)

library(ggplot2)

library(wordcloud)

library(wordcloud2)

library(ROAuth)
```

```
library(stringr)
library(tm)
tweets <- searchTwitter("#depress",n=9999, geocode="7.8742,80.6511,500km",lang =
"en")
tweets.df <- twListToDF(tweets)
tweets.df
#tweets will be copied to a file in CSV format.
write.csv(tweets.df,file = "D:/Tushara/depress_27_09_2021.csv")
# In the above example, 'depress' tweets are collected and stored in local system.
```

3.4 Study and data areas

We have been collecting tweets since 2018 for depressive, anti-depressive, and event tweets related to Kessler's depressive keywords, their antonyms, and few events. Corona-related tweets have been collected since March 2020 daily. In 2019, the Sri Lanka Bombing tweets were collected 500 Kms around the '*Dambulla"* central location. This helped us to understand the depressive symptoms of Srilanka people during and after the '*Bombing"*. Results are discussed in the following chapters. In 2020, in the Burevi cyclone that affected Srilanka, we collected the tweets related to "*Burevi"* 500Kms around '*Dambulla"* to cover the entire Srilanka. Similarly, we collected tweets around Mumbai for the study of the "*Tautkae'* cyclone. Tweets related to 'Space tourism' also collected from 06-July-2021 to 06-August-2021 for the keywords 'unity 22', 'blue origin', 'new Shepard'.

Table 3.1 Distribution of tweets during April 2020 to March 2021 (Sample)

S. No	Keyword	No. of Tweets
1	Depress	915
2	Failure	122486
3	Hopeless	12649
4	Nervous	3602
5	Restless	3602
6	Tired	76596
7	Worthless	13262
8	Active	84697
9	Calm	202773
10	Comfort	92358
11	Delight	43351
12	Excite	1689
13	Hopeful	40488
14	Peaceful	141490

In addition, we collected tweets on the location basis related to '*Atlanta'* and '*New York'* for the location-based studies. Similarly, tweets are collected in the years 2019, 2020, and 2021 also. We collected tweets from 6-July-2021 to 6-August-2021 for the keywords Unity22, Blue Origin, New Shepard.

3.5 The framework using Data Mining and Machine Learning for Predicting Mental Health States

Our work is to collect the tweets daily for the keywords 'depress', 'failure', 'hopeless', 'nervous', 'restless', 'tired', 'worthless', 'active', 'calm', 'comfort', 'delight', 'excite', 'hopeful', 'peaceful', and 'corona' using the Twitter API methods. Depending on the use and requirement, we add to collect the tweets related to a few additional keywords. We are collecting tweets related to 'Space tourism', 'cyclone', and 'corona' at present and are used for our works. This framework will use the following functions:

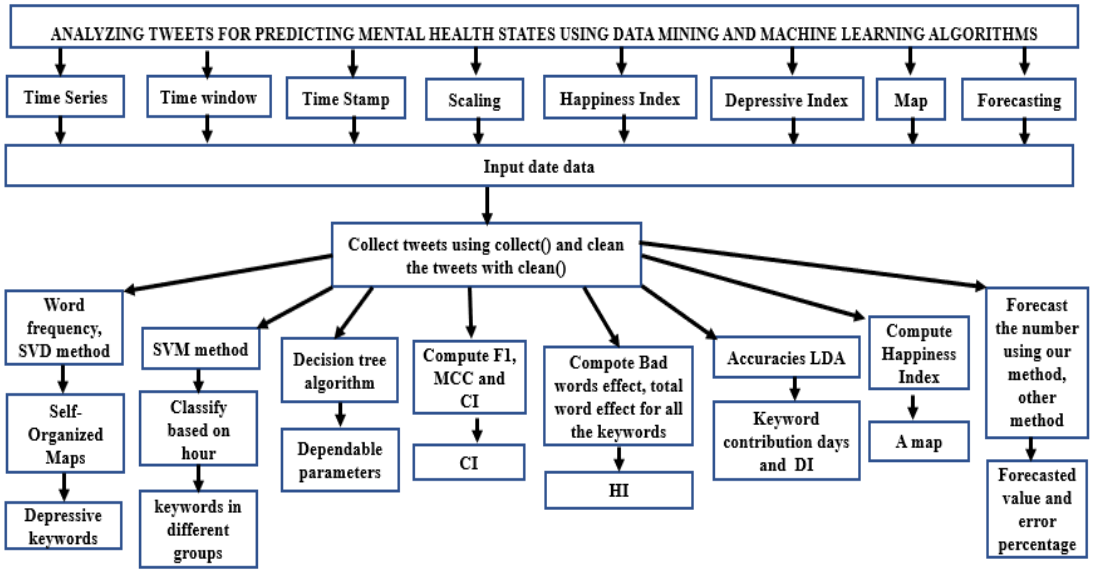

Figure 3.3 Framework diagram of our works

Figure 3.3 shows the framework of our work. The schematic flow charts for the collection of tweets from Twitter are shown in Figure 3.4. Flow charts for cleaning tweets and search of tweets are shown in Figures 3.5 and 3.6, respectively.

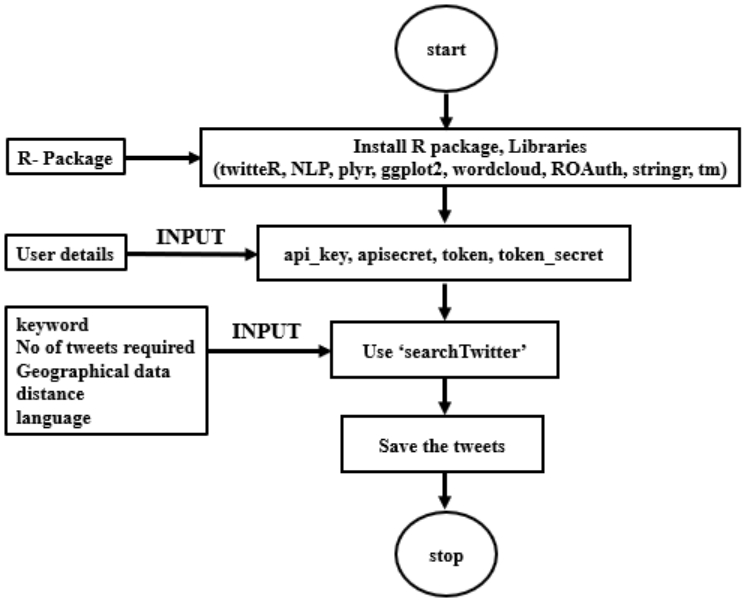

Figure 3.4 Schematic flow chart – Collection of tweets from Twitter

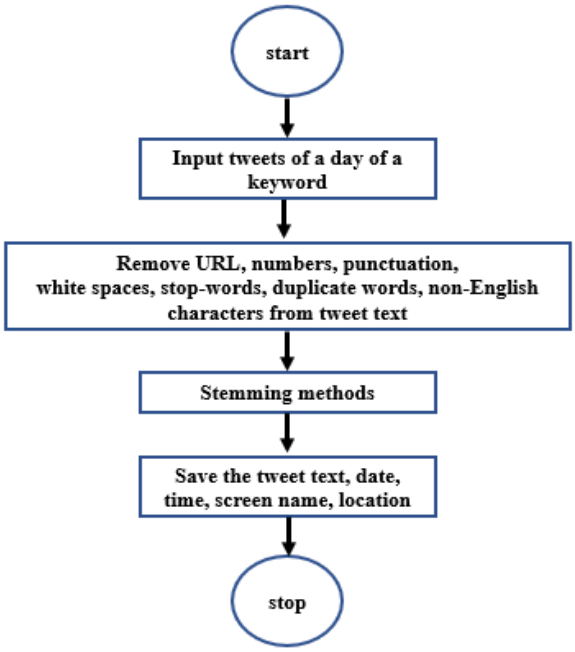

Figure 3.5 Flow chart – Cleaning of tweets

3.5.1 Collect function: Collect()

A function 'Collect' is prepared to collect the tweets as follows:

Input: Date of requirement, Keywords, Time duration (optional) to be collected.

Output: This function interacts with our database and collects the tweets belonging to the keywords and date. In the event time duration is chosen, the tweets will be limited to the required duration of the day.

3.5.2 Clean function: Clean()

This function, if applied, will remove all the URLs, remove numbers, non-English characteristics, and white spaces in the tweet text.

Input: Tweet data set

Output: Cleaned tweet data set

3.5.3 Search function: Search()

This function, if applied, will collect the number of times the keyword appeared in the in the tweet text.

Input: Tweet data set, Keyword

Output: Number that shows the keyword appears in the tweet data.

This function is used to find the True Positives, True Negatives, False Positives, and False Negatives in preparation for the Confusion Matrix

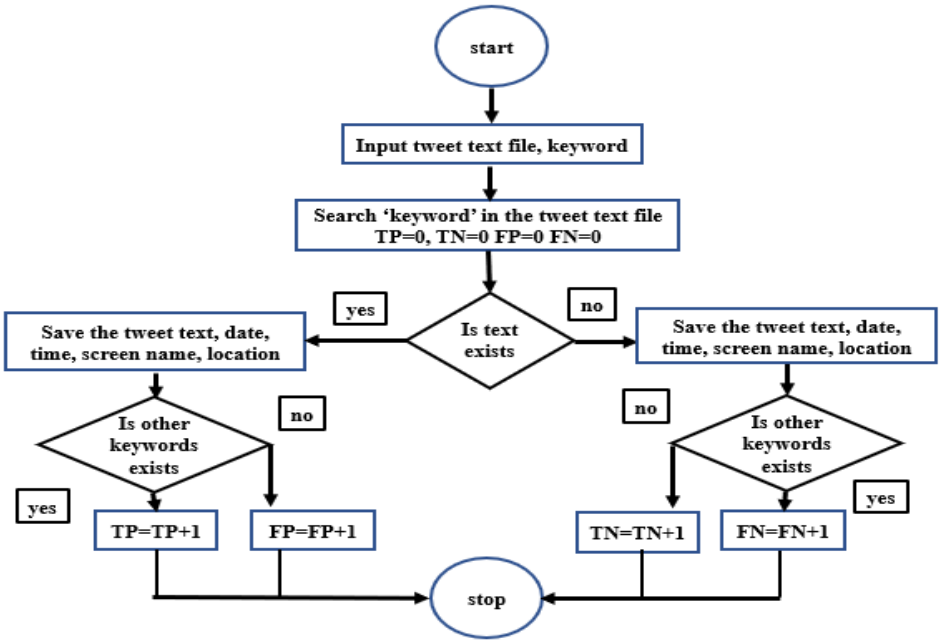

Figure 3.6 Flow chart – Search of tweet keywords

3.5.4 Happiness Index function: computeHI()

Input: Depressive and anti-depressive tweet data set of a day

Output: Happiness Index Number

This function computes the Happiness Index as described in section 8.3

3.5.5 Depression Index function: computeDI()

Input: Depressive and anti-depressive tweet data set of a day

Output: Depressive Index Number as described in section 8.4.1

3.5.6 Feel-Good-Factor function: feel_good()

Input: Depressive and anti-depressive tweet data set of a day

Output: Keywords in the rank order of the contribution on that day as described in section 8.4.2

3.5.7 Happiness Map function: create_map()

Input: Depressive and anti-depressive tweet data set of a day, Bad words set, date, and WalletHub data.

Output: A map showing the Happiness Index map along with the WalletHub Happiness Index

3.5.8 Prediction of keywords: predict()

Input: Depressive and anti-depressive tweet data set of our database. At present, it is from 01-April-2020 onwards. A keyword that is to be predicted. Date to be predicted

Output: A plot showing the forecast data. The predicted value of the required day and previous days error percentage

3.6 Graphic User Interface

An interactive GUI portal to calculate mental health states which processes the collect(), clean(), search(), computeHI(), computeDI(), feel_good(), creatMap(), predict() functions. We plan to create a public portal to facilitate users to interact and get the results. Upon completing the proposed GUI application, the users can get the Happiness Index values, Depression Index values, Happiness map, and prediction of keywords of the desired dates of geographical locations instantaneously.

4 ANALYSING TWEETS

4.1 Introduction

Analyzing tweets is an established research area, where several researchers implemented the known algorithms to deduce the results. Research is ongoing to verify the results to real-world issues. Six hundred tweets are added every second in diverse matters. Topic combinations and subjects demand numerous solutions for these applications. Some of the works are mentioned in the chapter - 2 of this dissertation. The tweets are analyzed to find the basis of the popularity of a person and reasons to support [73], emotions [74], disaster management with Support Vector Machine (SVM) detect the impact [75], Opinion mining [76], level of frustration [77], detecting events [78], and many more applications.

The technological growth, ease of availability of the internet networking devices such as mobile phones, and GUIs of social network sites improved people's behavior to interact with known and unknown people through posting text and pictures about the emotions, events, and valuable tips. Researchers analyzed the social network data to reveal mental health disorders, suicidal and depressive behavior [79]. Twitter is one of the most popular social network sites that facilitates many people building relationships with experts in many disciplines, promoting research, product, and feedback. The tweet is a composition of 280 characters, with a maximum of four photos up to 5MB on mobile and 15MB using the web interface restricted to GIF, JPEG, and PNG formats. Twitter also provides APIs that allow the collection of tweets data much simpler than other social network platforms.

Different formats of tweets are posted to gain the attention of more users in advertising products [80]. Social networks inflicted the weekly pattern and rhythm in our activities, where

cultures are deeply experienced [81] and adjust social behavior [82]. In social networks, Twitter has gained popularity in reaching the masses with its features. Twitter supports 300 million active users with more than 500 million tweets daily and will likely reach 340 million users by 2024 [83]. Using tweets, researchers studied the real-time events [84], behavior in health changes [85], smoking habits, advertising formats [80], depression studies [32] [55] [23], behavior activation during morningness-eveningness depression [86], engagement of users [87] and misinformation using social networks [88]. Juntunen et al. studied to integrate B2B advertising with social media [89], and these marketers are inclined to utilize more emotional than functional requests in their tweets [90]. B2C research showed the positive effect of Twitter activity on information irregularity [91]. We analyzed the time series data, depressive, and anti-depressive tweets to identify and categorize the keywords.

4.2 Data collection and Cleaning

We identified seven keywords from the Kessler [92] ten-point questionnaire, the most used method to find the individual Psychological Distress scale. We obtained the keyword-related tweets from twitter.com using API from April to July 2018. These tweets using #depress, #failure, #hopeless, #nervous, #restless, #tired, #worthless were collected. We cleaned the data as mentioned in our previous chapters. Using ‘tm’ package in R, the numbers, special characters, symbols, white spaces, stop words, and ‘http’ links were removed from tweets.

4.3 Methods

4.3.1 Word Frequency Method

A Term Document matrix is obtained, and a word frequency table is prepared. We chose the top 24 high-frequency words from each keyword (#hash tag). The same process was repeated to all the keywords separately. A set of tweets shows 202 different words in the collection of word

sets of all the keyword lists. Table 4.1 shows the top-24 words and the number of times that appeared in the combined tweet set of the keywords.

The significance of the word within the words and keyword within the keywords is computed using the formula WS (Word Significance) = $A(i)/\Sigma A(i)$, where A (i = all the keywords) is the frequency from Table 4.1. KS (Keyword Significance) = $B(i)/\Sigma B(i)$, where B (i = all the frequent words) is the frequency from Table 4.1. Each of these values (WS and KS) will be within 0 to 1, and we then computed the $SV_{i,j}$ (Significance Value) = $WS_{i,j}$ * $KS_{i,j,}$ and shown in Fig 4.1. Each $SV_{i,j}$ value iterated again to compute WS and KS. Thus we obtained SV i,j and shown in Fig. 4.2

Word Frequency Algorithm

Step 1: Identify the keywords (i=1,7; i_1=depress, i_2=failure, i_3=hopeless, i_4=nervous, i_5= restless, i_6= tired, i_7=worthless) from Kessler work [92] .

Step 2: T_i = Set of tweets of for each keyword (i = 1 to 7) // Collection of tweets using each #hash tag (T_i = Tweet text)

Step 3: //For all the keywords Depress, Failure, Hopeless, Nervous, Restless, Tired and Worthless//

For i=1 to 7

Do

{

T_i = {T_i – Stopwords} // Remove the stop words from all the set of tweets

WF_j = (W_j, F_j) //Wi is the word from all the tweets T_i and F_i is the frequency of the word in the tweet set

// Sort (Wj, Fj) with Fj as a sorting parameter.

Select ((Wj, Fj) such that Fj>Fj+1 for j= 1,24 // Selection of top 25 cited keywords in the tweet text.

}

Step 4: {Wi, j; Fi, j} = Union of WF_j; j = 1,7 //Top 25 ordered keywords in the combined tweet set

Step 5: Select (Wi, j; Fi, j) such that Fi,j > Fi,j+1 for j=1,24 //Selection of top 24 keywords from the combined set.

Step 6: Ai,j = Fi,j for all i,j

Step 7: WS (Word Significance) = $A(i,j)/\Sigma A(i,j)$

//, where A (i = all the keywords) is the frequency, is computed, j is the keyword

Step 8: KS (Keyword Significance) = $B(i,j)/\Sigma B(i,j)$

// where B (i = all the frequent words) is computed, j is the keyword

Step 9: $SV_{i,j}$ (Significance Value) = $WS_{i,j}$ * $KS_{i,j}$ (Fig 2.1)

Step 10: Ai,j = $SV_{i,j}$ for all i,j.

Step 11: Repeat steps 7 to 10 to get second degree $SV_{i,j}$ (Significance Values) (Fig 4.2)

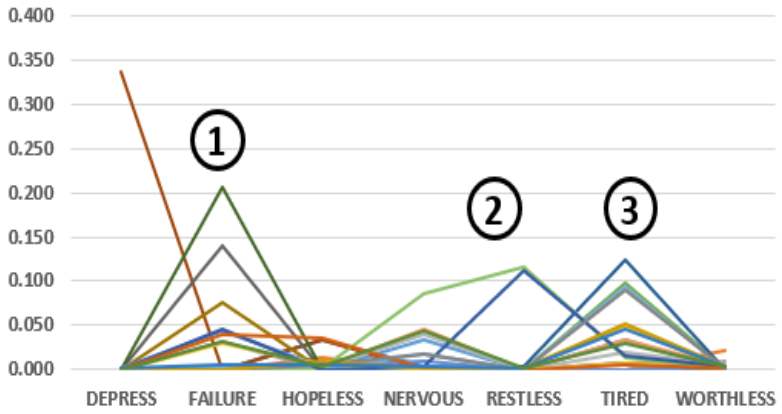

Figure 4.1 Significance Factor obtained at algorithm step 9

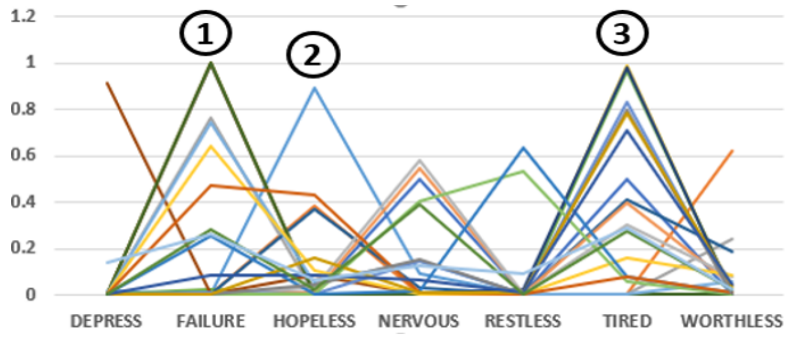

Figure 4.2 Significance Factor obtained after one iteration step 11

Table 4.1 Distribution of tweets from April 2020 to March 2021

	Depress	Failure	Hopeless	Nervous	Restless	Tired	Worthless
Hope	9	0	220	134	0	0	0
Know	0	0	143	0	0	0	258
Realdonald trump	0	920	0	0	0	0	305
Night	0	0	0	0	72	1823	0
Today	0	0	0	1280	0	1984	0
Work	0	529	0	0	0	4042	160
One	0	0	159	85	0	504	161
Depress	246	0	253	0	0	0	0
Never	11	3338	0	0	119	0	85
Succeed	0	1686	0	0	0	0	0
Sleep	0	0	0	0	246	4602	18
Fail	0	4648	0	0	0	0	0
Will	15	1594	169	576	0	0	256
Go	0	0	133	1745	0	2292	365
Like	0	0	179	1445	29	1619	393
Don’t	14	1320	222	0	21	828	284
Day	16	473	139	1274	17	4687	4
Now	0	0	126	3012	1949	1775	237
Time	0	2814	81	589	2022	1998	376
Can	12	1888	753	231	44	944	188
Im	10	0	343	1261	43	4461	148
Feel	0	0	374	209	110	2511	255
Just	27	781	324	554	129	2852	329
Get	0	2505	322	2372	95	3069	442

Out of the seven keywords considered for observation, three peak patterns are found in our study. It is noticed that people who tweet with one depressed keyword is automatically using the other three depressed keywords [33].

4.3.2 Singular Value Decomposition Method

Singular Value Decomposition (SVD) is an important tool in the area of Information Retrieval. Similar techniques are used to address problems in an approximation of the keyword-document matrix using its SVD [93]. Correlation coefficients were found to these keywords from the frequency lists. We applied the Singular Value Decomposition method [94] to identify the number of parameters contributing to this process. Understanding data and extraction of features using SVD method delivered good results [94], and this method is also used for data dimension reduction [95]. In SVD method, the given matrix A (nxm) is decomposed into A= Unxn, V mxm, and Γ nxm matrices with matrix have elements in the diagonal. The elements Γ i,j, where i=j are called the singular values of the matrix. We can also note that Γi,j values are always in descending order. Feature extraction was also carried out using the SVD method [96]. The singular values extracted in the SVD process capture the essential features in the data. We have applied the SVD method and found that the first three matrix elements comprise more than 63% of this decomposition. Singular values from SVD Method resulted as 1.862, 1.370, 1.216, 0.910, 0.775, 0.488 and 0.379 from these data [33]

4.3.3 Time Series Method

We also attempted to analyze the tweets collected at their time of posting. Table. 4.2 shows the time of the tweet and the number of tweets in our corpus. It is observed that a minimum number of tweets (12499) were posted during the 8-10 hrs time, and the highest number of tweets were posted during the 16-18 hrs time. People tweeted more during the afternoon to midnight than

midnight to the afternoon. Tweets related to ‘Failure’ appeared 64013 times, and a minimum of 382 tweets appeared with the ‘depress’ hashtag in our corpus. (In Table 4.2: 0-1 indicates 00:00 to 1:59 hrs). We computed the average percentage of each #hash (keyword) and its contribution to the corpus. The tweet data consist of 0.2 million used as a resource base for our further analysis.

Table 4.2 Distribution of tweets from April 2020 to March 2021

Keyword/ Time	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	16-17	18-19	20-21	22-23	Total
Depress	22	17	48	22	26	24	40	49	49	26	33	26	382
Failure	5250	4584	3894	4125	3642	4516	6011	7192	7284	6251	5925	5339	64013
Hopeless	1114	1062	982	1158	1085	1272	1375	1528	1921	1569	1670	1432	16168
Nervous	3516	3225	1975	2121	2127	2330	3069	2930	3273	3511	3088	3447	34612
Restless	738	730	637	615	523	659	892	827	800	646	697	585	8349
Tired	5018	5391	5529	5023	3661	4404	5628	5942	5912	5826	6173	5595	64102
Worthless	1439	1251	1291	1444	1435	1689	1919	1749	1953	1853	1925	1482	19430
TOTAL	17097	16260	14356	14508	12499	14894	18934	20217	21192	19682	19511	17906	207056

Kessler's [92] works are one of the acceptable methods to uncover personal Psychological Distress levels. Seven keywords are identified from Kessler’s questionnaire, and we collected tweets with these seven #hash (keywords) on 25-October-2018 to identify any abnormality on that day. Fig 4.3 indicates the depression keywords and their anomaly with the corpus data. We found the difference between the average corpus data to the present day of observation.

While comparing the day data (25-October-2018), we found an anomaly at 10-12 hrs in Depress, 6-8 hrs at Restless, and 10-12 in worthless tweets. The difference between the average corpus value in today's observation varied between 8.61 to -14.41. We took the absolute values and categorized them into four different categories. ‘A’ is very high, ‘B’ is High, ‘C’ is given to abnormal, and ‘D’ is for the normal category. While ‘A’ needs urgent attention, ‘B’ needs caution, ‘C’ category needs to be monitored, while ‘D’ is normal.

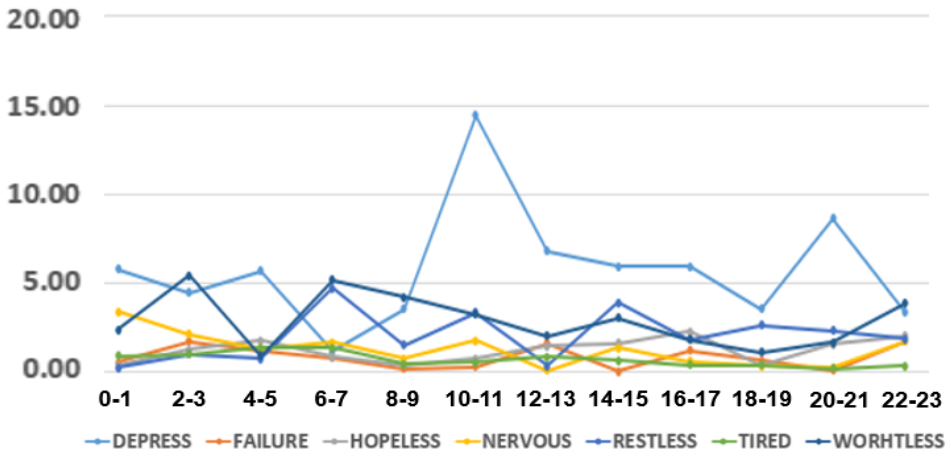

Figure 4.3 Significance Factor obtained after one iteration step 11

Table 4.3 Distribution of tweets during April 2020 to March 2021

Keyword/ Time	0-1	2-3	4-5	6-7	8-9	10-11	12-13	14-15	16-17	18-19	20-21	22-23
Depress	C	C	C	D	D	A	C	C	C	D	B	D
Failure	D	D	D	D	D	D	D	D	D	D	D	D
Hopeless	D	D	D	D	D	D	D	D	D	D	D	D
Nervous	D	D	D	D	D	D	D	D	D	D	D	D
Restless	D	D	D	D	D	D	D	D	D	D	D	D
Tired	D	D	D	D	D	D	D	D	D	D	D	D
Worthless	D	C	D	C	C	D	D	D	D	D	D	D

It is observed that depression at 10-11 hrs was abnormal, and between 20-21 hrs too found different. Such charts are made for several days for our study. A combination of such grade charts indicates the different grade timings. [40].

4.3.4 Time Window Method

We have collected more than 200,000 tweets for all these keywords (#hash tag) for this work. We removed the URLs and their links in the tweets. We removed the numbers, punctuation characters, and white spaces. We used the stemming method and removed all the stop-words. In tweets, URLs, links, numbers, punctuations, and other stop-words (unnecessary words) are removed in the tweet datasets. We found each word frequency in each day's tweets, and then we chose the 25 high-frequency words from each keyword (#hash tag) related tweets. The

identification of frequently used words was repeated to all the keywords for all the days separately. We found that there are 202 different words in common in our data corpus. The workflow is as follows:

Algorithm

Step 1: N = number of depression keywords taken from Kessler work

Step 2: M= number of days work

Step 3: i=0; j= 0 //initialization

Step 4: Collect #hashtag tweets using twitter API

Add to a corpus

i = i+1

if i ≤ M go to Step 4

j= j+1

If j ≤ N go to step 4

Else

Stop

//Data Preparation Phase

// $t_{i,j}$ is the tweet collected on the i th day for the j th keyword//

Step 5: ∀j= 1 to N,

Do

{

∀ $t_{i,j}$ Remove numbers, URLs, punctuations, blank spaces, stop-words.

Apply stemming

Create a table of word frequency

Sort the table

Collect 25 top frequently used words from this table

}

Step 6: ∀j= 1 to M

Do

{

Create a combined corpus of all frequently used keywords

}

Step 7: Find top 25 (keywords) from this combined list.

//data processing stage//

//$rt_{i,j}$ is the time of the tweet $t_{i,j}$ tweeted.

Step 8: p=0; q=0; k=0 // initialization

{

Search #keyword in the $t_{i,j}$

If #keyword exists in the text of $t_{i,j}$

{

find $rt_{i,j}$

K=K+1

if $k \leq 25$ go to step 8

}

Else

}

//Separation of tweets that are within 10 minutes duration and above//

Step 9:

// $d_{i,j}$ is the time difference between successive usage of the keyword in the tweets//

∀ k = 1 to 25

Do

{

$d_{i,j} = rt_{i+1,j+1} - rt_{i,j}$

if $d_{i,j} \leq 10$ p=p+1

else q=q+1;

}

We also collected tweets related to 'anti-depressant' and many other drugs used for reducing depression. We repeated the same process, computed the frequency of the top-25 words available in the sets, and analyzed further for our study. Using MATLAB, we calculated the accuracies of people who tweeted with the anti-depressant medicine #hashtags with the other identified keyword tweets. The results from the top 25 frequently used words were discussed from the set of 202 words.

The similarity values of the frequently used words that are posted in a ten-minute interval and beyond in depression and anti-depressant tweets are computed. They are shown in Figures 4.4 and 4.5. We have established the trend of the words in a graphical pattern in Figures 4.4 and 4.5. Fig 4.4 shows the depression and anti-depressant trend lines for the frequent words used in the tweets that are posted within 10 minutes. Fig 4.5 shows the depression and anti-depressant trend lines for the frequent words used in the tweets that are posted beyond 10 minutes. The dotted line in blue indicates the anti-depressant tweets frequency in the selected set of tweets. Other lines

show the trend lines of ‘failure,’ ‘hopeless,’ ‘nervous,’ ’restless,’ ’tired,’ and ‘worthless. Most of these words followed a similar trend except at one point of observation.

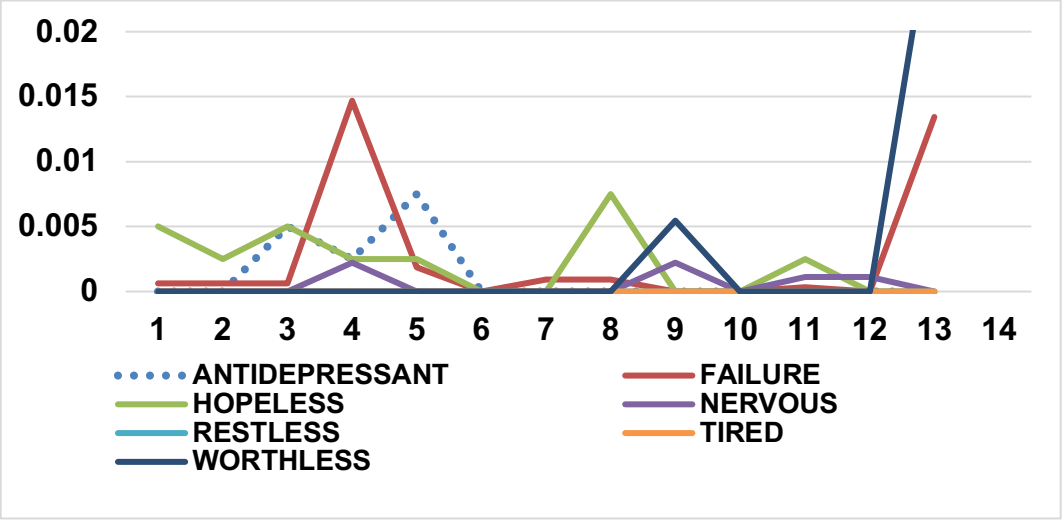

Figure 4.4 Trend of Anti-depressant (blue) and other Kessler depression words (<10 minutes)

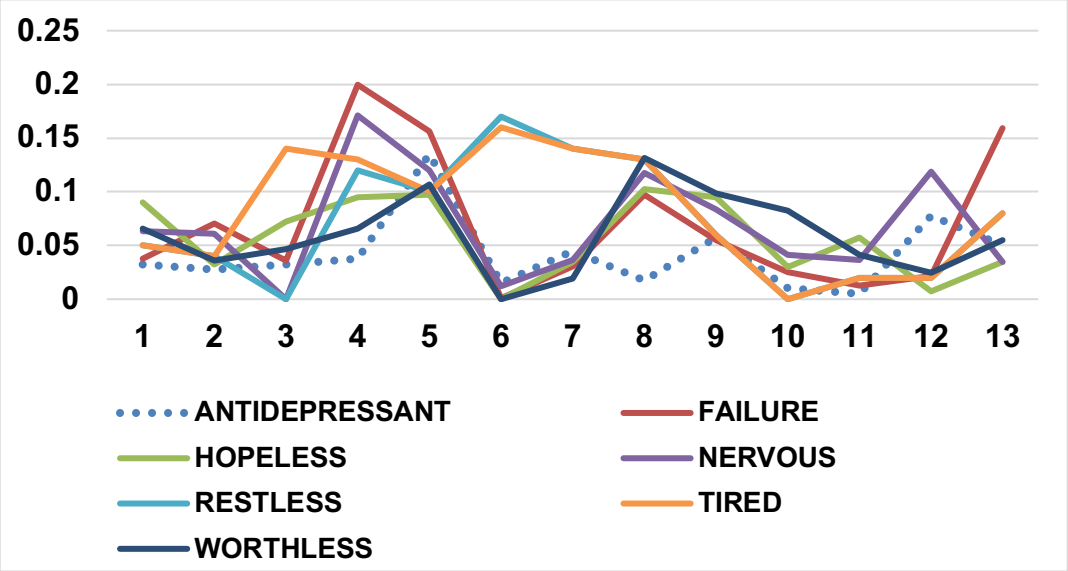

Figure 4.5 Trend of Anti-depressant (blue) and other Kessler depression words (>10 minutes)

More similarity is seen in the pattern with the frequent words of the tweets that are collected beyond 10 minutes. The X-axis indicated the top frequently appeared word set from the dataset. Some of the tweet sets do not represent these words, and hence we eliminated them for our analysis. We analyzed the data using the Fine-Tree algorithm [97], SVM [98], and KNN methods [99] and found 20-40, 16-40, and 8-40 accuracies. The similarity values within ten minutes and above ten minutes revealed similarity in some depressive and anti-depressive keyword hashtag tweets. The study of the tweeting pattern of depressive and anti-depressive tweets showed a similar trend in some of the tweet keywords. The tweet pattern of 'tired' and 'restless' exhibited a similar pattern, and others showed a different pattern. The anti-depressive tweets followed the pattern of 'failure,' 'hopeless,' 'nervous,' and 'worthless' tweets. Fig 4.4 and Fig 4.5 show that there is a similarity in the tweet keywords.

4.3.5 Time Stamp Method

Analyzing the tweets' timestamps, the researchers attempted to comprehend user percentage compared to other time slots and focused on timing when the tweet impact was high [100] [101]. Modeling of time-of-day and day-of-week behavior that influences a customer was studied [102], and daily patterns of such tweets were analyzed [103]. Studies show that the best times to post tweets on Twitter were, in general, 8-10 am and 6-9 pm. It is also reported that the best times for B2C companies were 8-10 am, 12 pm, and 7-9 pm, and to get more retweets, it was 5-6 pm. The best days for B2C are weekends, and for B2B, they are weekdays [104].

The best times for tweets posting of various sectors were studied and observed that key days for media companies to post were Thursday and Friday. It is reported that it is beneficial to post tweets around 5-6 pm on Saturdays for the education sector. It is suggested to post tweets for non-profit and charity organizations around 7 am on Wednesdays, tech companies on Tuesdays

and Wednesdays, health care companies between 8 am and 2 pm on weekdays, and finance companies between 1 am and 5 am on Tuesday. At 2 pm on Sunday and during Fridays are suggested for the recreation industry [105]. A similar study [106] exhibited a peak volume of tweets from different geographical areas globally: retweets, replies, and feedback times discussed in this work.

So far, not much work has been done using the mental health-oriented tweet data. We analyzed the multi-domain tweets related to depressive, anti-depressive, and COVID-19 to discover tweeting patterns, timings, and days of tweets by using the date and timestamp information. We also compare newly discovered results with the COVID-19 confirmed cases published by World Health Organization (WHO) to verify our discovery.

Kessler's Psychological Distress scale [92] is one of the methods to understand an individual's mental health status. We used Twitter API and collected the tweets data with seven keywords from Kessler's questionnaire. The seven keywords chosen from Kessler's works are depress, failure, hopeless, nervous, restless, tired, and worthless. We also used active, calm, comfort, delight, excitement, hopeful and peaceful, the antonyms of these seven keywords, to collect the tweets. In addition, we collected tweets with the keyword 'corona.'

The tweet data used for our study belong from 01-April-2020 to 01-April-2021 (366 days). 2.3 million tweets are used. The number of tweets in a keyword set during a particular hour is computed using the date and timestamps of each tweet. This process is repeated for all the depressive, anti-depressive, and corona keywords.

WHO documented the corona confirmed cases data that were available in the public domain [107]. Random weekly data of corona confirmed cases were obtained during July 4-10, 2021 (WHO-1), June 6-12, 2021(WHO-2), and May 2-8, 2021 (WHO-3).

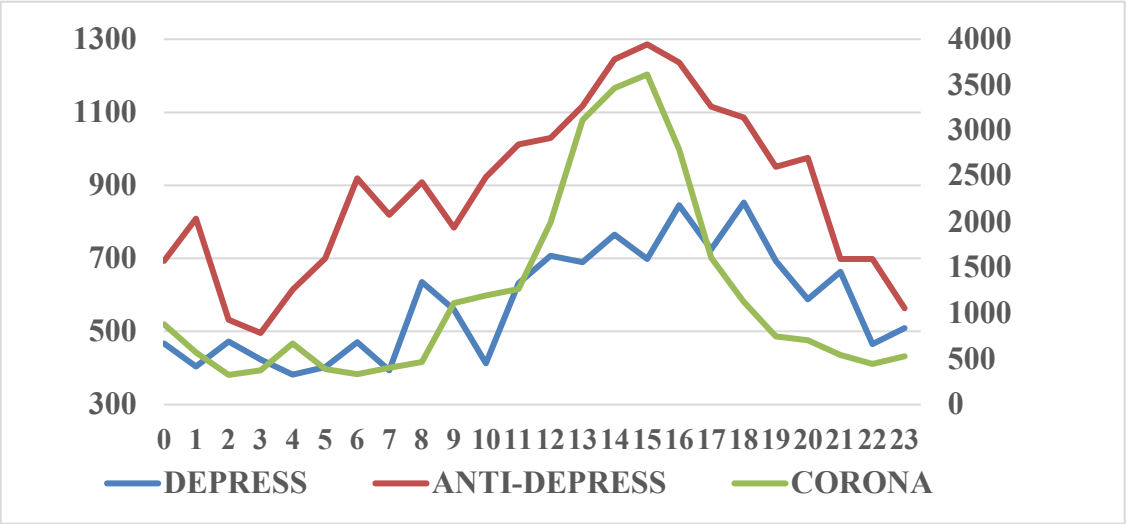

Figure 4.6 Depressive, Anti-depressive, and Corona tweets and with timestamp

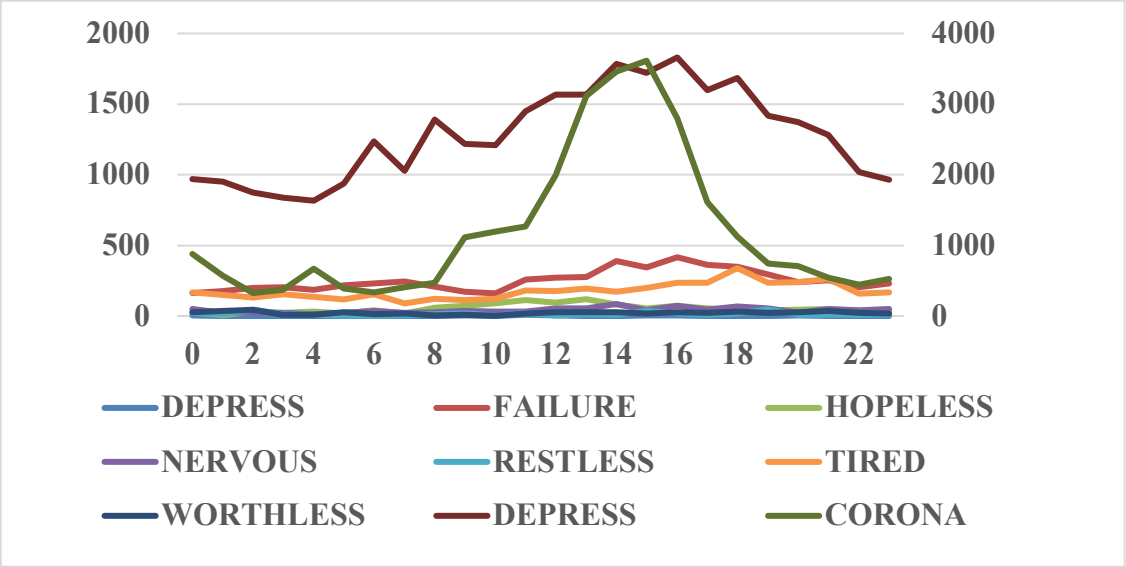

Figure 4.7 Depressive keywords and Corona tweets and with timestamp

Fig. 4.6 shows the depressive, anti-depressive tweets, and corona tweets along with their timestamps. Fig. 4.7 shows each depressive keyword (depress, failure, hopeless, nervous, restless, tired, and worthless) and corona keyword along with relevant timestamp information.

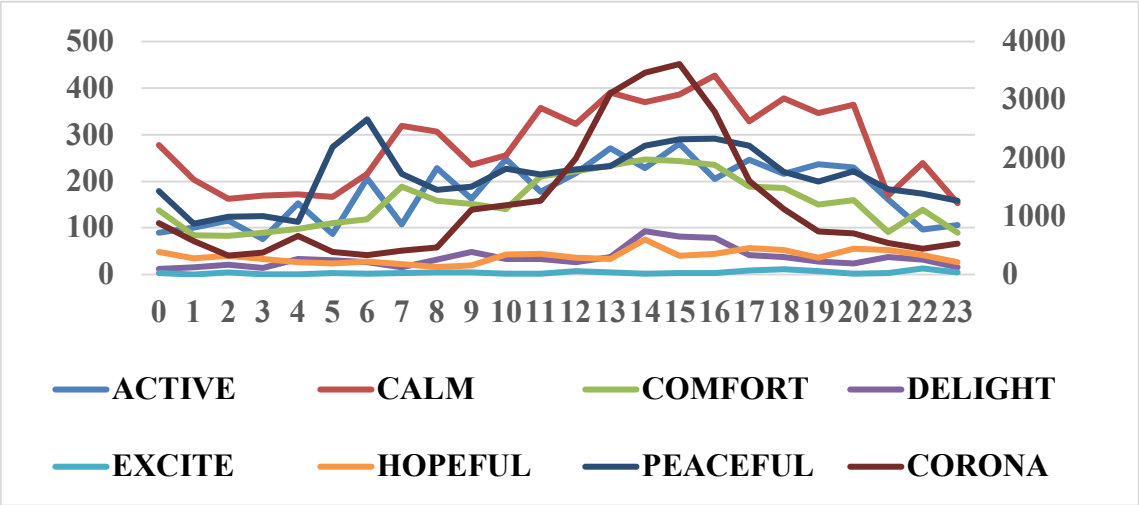

Figure 4.8 Anti-depressive keywords and Corona tweets and with timestamp

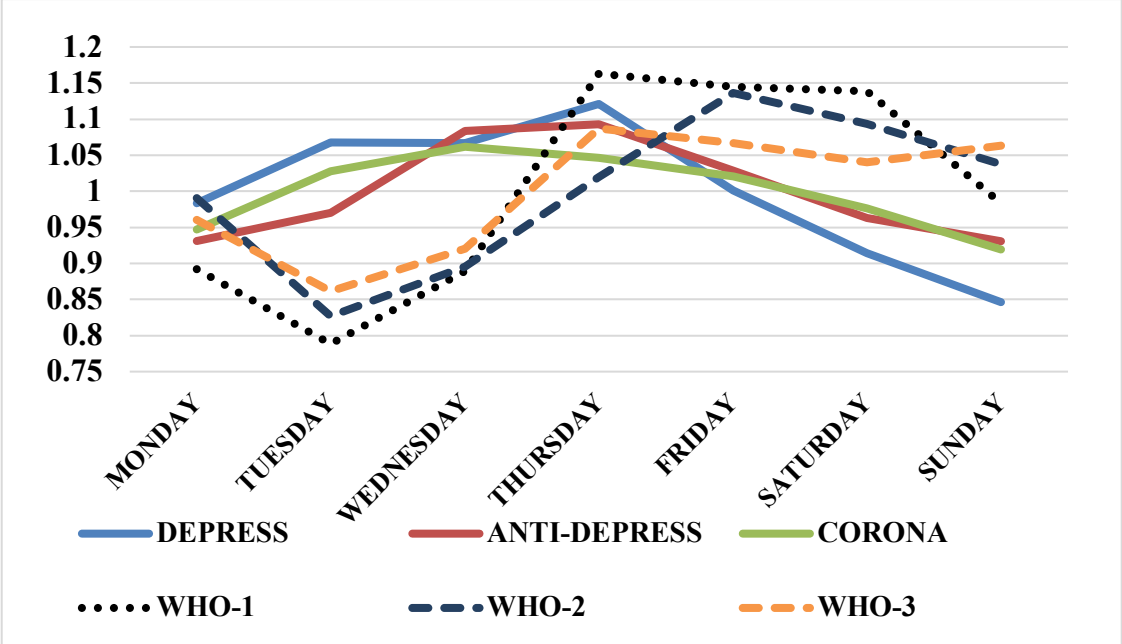

Figure 4.9 Depressive, Anti-depressive keywords, and their days of posting

Similarly, Fig. 4.8 shows the anti-depressive keywords (active, calm, comfort, delight, excite, hopeful and peaceful) along with relevant timestamp information. Depressive, Anti-depressive, and corona keyword tweets collected from 01-April-2020 to 01-April-2021 (366 days) are analyzed. Fig. 4.9 shows normalized average values of the depressive, anti-depressive, and corona tweets data along with the WHO data. Fig. 4.10 shows the rate of change of the number of

tweets from the previous hour. Figures 4.11 to 4.15 show the depress, nervous, worthless, active, and calm tweets patterns in a day, respectively

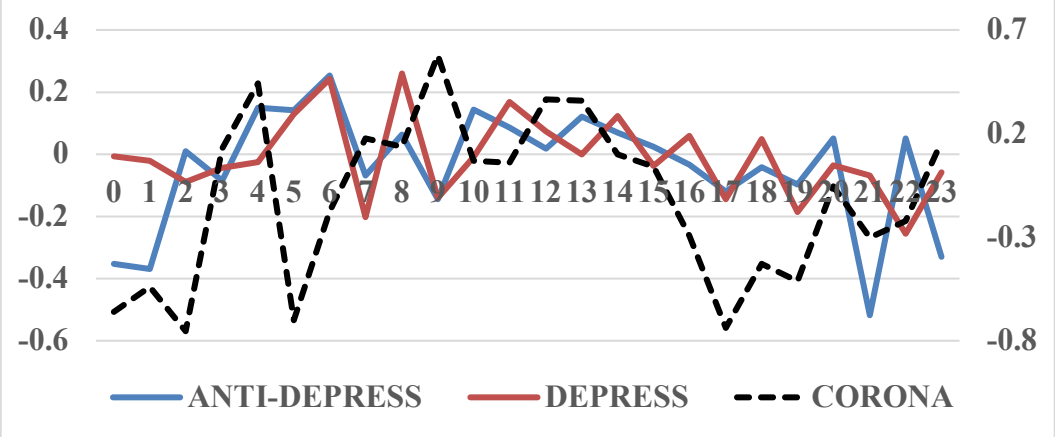

Figure 4.10 Rage of Depressive, Anti-depressive keywords, corona tweets from the previous hour

4.3.6 Classification of Tweet Key words

The tweets collected from 01-April-2020 to 01-April-2021 are organized in an hour-wise and noticed pattern appeared in a waveform.

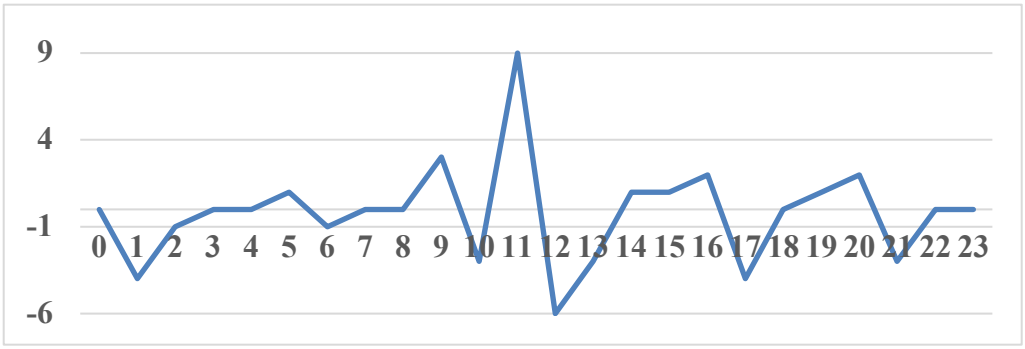

Figure 4.11 Depress keyword tweet pattern in a day

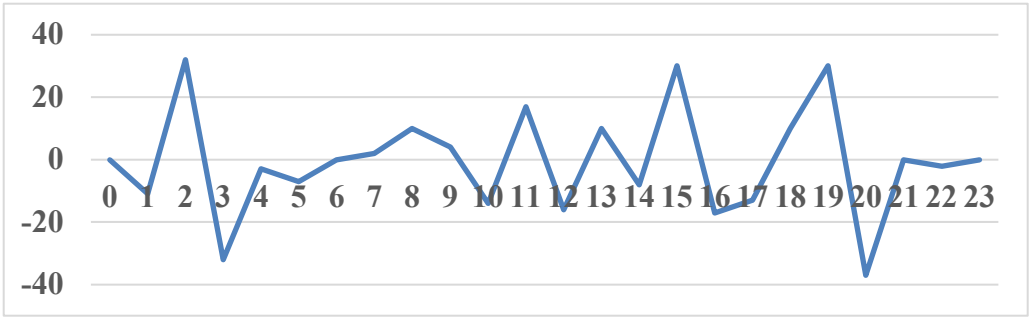

Figure 4.12 Nervous keyword tweet pattern in a day

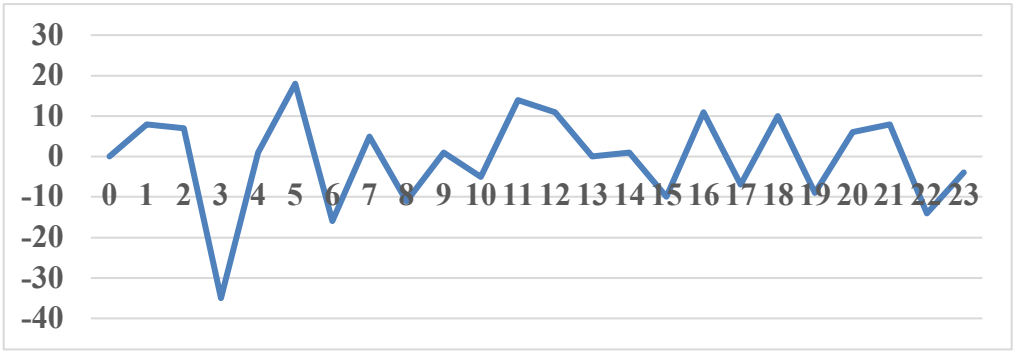

Figure 4.13 Worthless keyword tweet pattern in a day

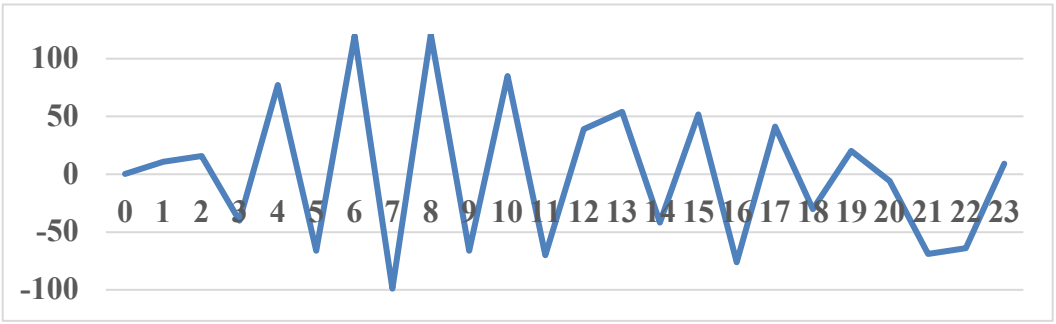

Figure 4.14 'Active' keyword tweet pattern in a day

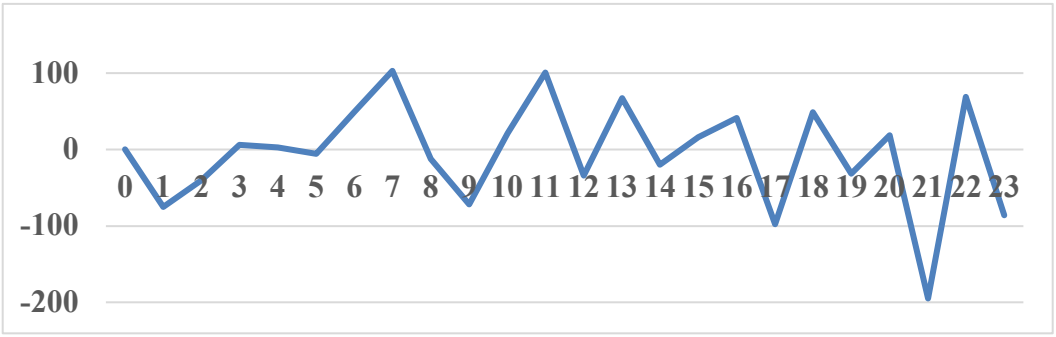

Figure 4.15 'Calm' keyword tweet pattern in a day

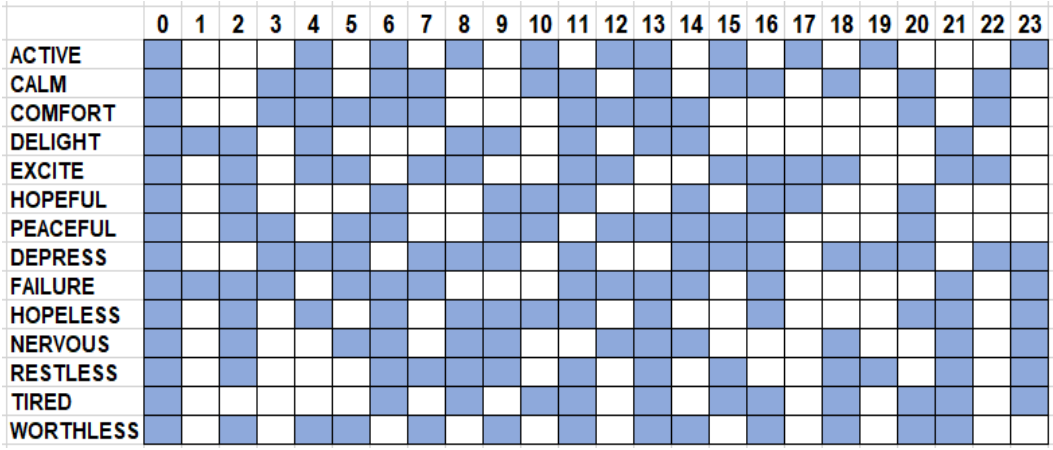

Figure 4.16 Tweet keyword map pattern in a day

The gradient of tweets is computed and shown dark if the rise is positive and white if it is negative in Fig. 4.16. The frequencies and amplitude of each keyword are also studied.

Table 4.4 Classification of Keywords

Sl. No	Tweet Keyword	Frequency Type	Amplitude Type	Polynomial degree
1	Active	Low	High	3
2	Calm	Low	High	3
3	Comfort	High	Med	3
4	Delight	High	Low	2
5	Excite	Med	Low	2
6	Hopeful	Med	Low	3
7	Peaceful	High	High	3
8	Depress	Med	Low	2
9	Failure	High	High	3
10	Hopeless	Low	Low	3
11	Nervous	Low	Low	2
12	Restless	Low	Low	1
13	Tired	Low	High	3
14	Worthless	Low	Low	3
15	Covid	High	High	4

We classified the tweet keywords into three categories with their frequencies as Low (2.4-3.2), Medium (3.2-4.0), and High frequency (4.0-4.8). Similarly, the wave pattern was studied for

its amplitude and classified into three categories as Low (0-40), Medium (40-75), and High (75-110). Table 4.4 shows these details.

Polynomial trend lines for the keyword curve patterns are obtained, and their degrees are shown in Table 4.4. The value of the function, first and its second derivative values, are shown in Figures 4.17, 4.18, and 4.19 for the function values 1, 14, and 15 hours.

2.3 million tweets collected from 01-April-2020 to 01-April-2021 (366 days) were analyzed with hour-wise data, resulting in significant discovery. In depressive tweets, it is noticed that the larger number of tweets posted during 12-1 pm and 2-3 pm, thus partially coincides with the previously reported results [103]. More anti-depressive tweets are posted during 2-5 pm. Similarly, we observed that corona-related tweets are posted more during 2-4 pm. The rate of change of tweets related to corona tweets is varied primarily during 0-6 am and 3-8 pm, whereas depressive tweets have 6-9 am. Anti-depressive tweets have similar changes during 8-10 pm, agreeing with the results [103].

It is observed that more depressive tweets are posted on Thursday, followed by Tuesday and Wednesday. On Monday, Saturday, and Sunday, the depressive tweets are posted less than the average tweets in the depressive category. More anti-depressive tweets are posted Thursday, followed by Wednesday and Friday. On Monday, Tuesday, Saturday, and Sunday, the anti-depressive tweets are less than the average tweets posted during the week. Corona tweets are posted more on Wednesdays and followed by Thursday and Tuesday. On Monday, Saturday, and Sunday, the corona-related tweets are posted less than the average number of week tweets.

On Thursdays, the depressive, anti-depressive tweets, and corona confirmed cases data obtained from WHO show an increase from Wednesday. A similar trend is followed on Friday, Saturday, and Sunday too. Our new study supports the previous works [105] [106].

We observed that the tweets related to Hopeless, Nervous, Restless, and Worthless are low in frequency and amplitude. The tweets with keywords Failure, peaceful and covid showed a high frequency and amplitude. Other keywords have shown a different nature. The trend line equations to fit curves and their polynomial degrees are obtained (Table 4.4)

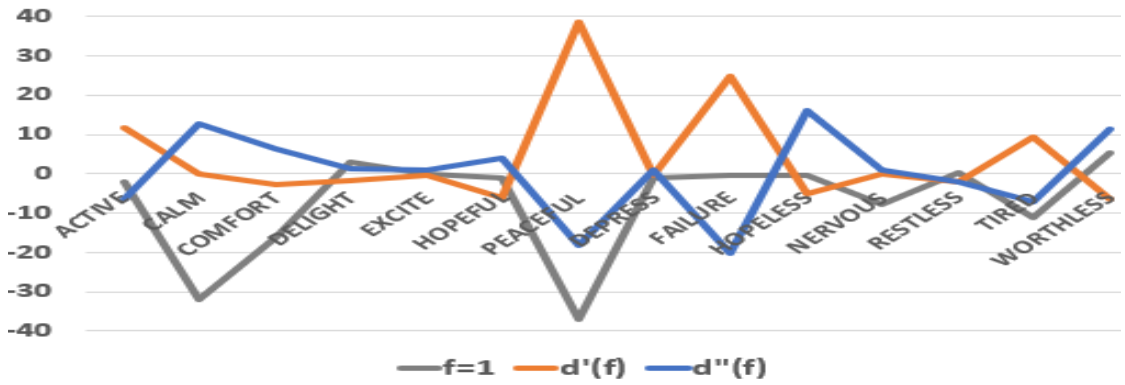

Figure 4.17 Function value (=1), first and second derivatives

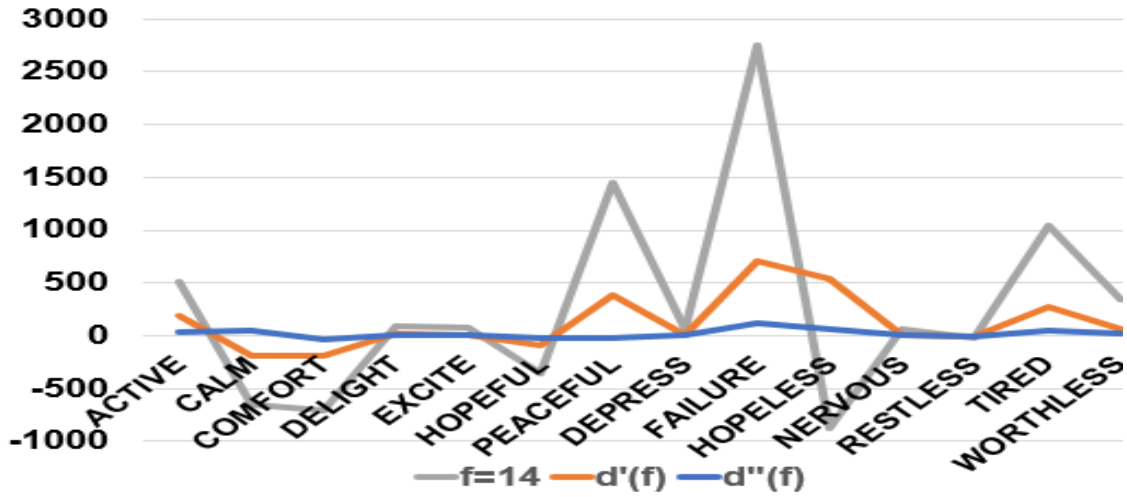

Figure 4.18 Function value (=14) first and second derivatives

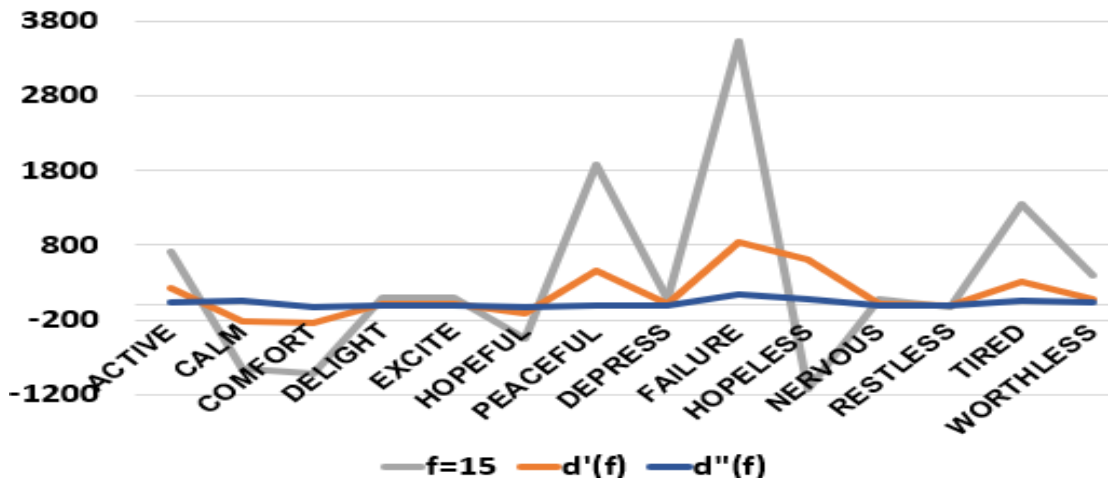

Figure 4.19 Function value (=15), first and second derivatives

4.4 Discussion

We observed that there are three peaks from the significance values from Figures 4.1 and 4.2. Out of the seven keywords considered for observation, three peak patterns are found in our study. It is noticed that people who tweet with one depressed keyword are automatically using the other three depressed keywords. The frequency of words usage depicted the word usage pattern with keywords and resulted in an individual in a depressed stage may soon fall into another depressed phase. The results obtained from the Singular Value Decomposition (SVD) method resulted in a similar result showing a complement of 63% contribution to the first three pivotal values. The SVD method also supported our results, and we observe that there is always a chance of individual tweeting at least three of these keywords if he/she is mentally depressed. Everitt and Dunn [108] proposed an alternative approach based on comparing the component contribution of these diagonal elements as almost 64%, which is in line with our results. In all these cases, an individual with a symptom is automatically going into the other three situations. The results will be an indicator to identify and suggest developing a mechanism for social websites to help mental health illness people. Continuous monitoring of tweets of an individual who tweeted with one

keyword will identify whether he/she tweets with other depressive keywords so that we can conclude that the individual is prone to Mental health Illness soon.

It is observed that depression at 10-11 hrs. was abnormal, and between 20-21 hrs. too found different. A combination of such grade charts indicates the different grade timings. Such categorization of tweet time will help predict the time-series data, where an abnormality is observed. Identification of tweets of a duration to be abnormal with other timings will help to isolate for any event during that time. Studying individual tweets from this abnormal part of tweets will result in user-level identification. Once the user is identified for such abnormality, diagnosis can be demonstrated with the help of a Mental health Specialist.

The values of second derivatives of these curves remained positive in many cases except peaceful and restless, which showed negative throughout. During the 3-5 hours, the second derivative changed its slope to negative, indicating a local maximum. The second derivative of the corona-related polynomial also turned negative between 4-5 hours. Our study gives a detailed timestamp analysis to understand the depressive, anti-depressive, and corona tweets postings. Tweeting patterns are depicted, and abnormalities are identified. People post tweets more on Thursdays compared with the other days. The tweeting pattern falls until Sundays and picks up from Monday. During the day, there is a significant rise in tweets during 2-6 am and 10 am to 2 pm. The number of tweets is retarded from 2 pm onwards till the day ends. Most of the depressive tweets follow the same tweet diurnal pattern. Anti-depressive tweets follow a similar trend with intermediate aberrations. Classification of these tweet keywords resulted in that Hopeless, Nervous, Restless, and Worthless fall under one category. Active, calm, comfort, delight, excite, hopeful, depress, and tired are under the second category. Failure, peaceful and covid fall under the third category. The classification will allow the researchers to group them in future mental

health studies. The first and second derivative data of the function of all these keywords at 14 and 15 hours show a similar pattern and supports our classification of these keywords and supports earlier results [103]. The second derivative of seven keywords remained positive, two were negative, and five changed their sign during the day cycle, supporting our study.

The word frequency method resulted in the word usage pattern that an individual in a depressed stage may soon fall into another depressed phase. The singular Value Decomposition (SVD) method resulted in a compliment of 63% contribution to the first three pivotal values. The SVD method also supported our results, and we observe that there is always a chance of individuals tweeting at least three of these keywords if they are mentally depressed. Everitt and Dunn [108] proposed an alternative approach based on comparing the component contribution of these diagonal elements to almost 64%, which is in line with our results. We could identify the need for urgent attention, caution and monitor it with the Time series method. Such categorization of the tweet time will improve prediction for the time-series data, where the abnormality is observed. Our accuracy in the Time window method resulted in 20-40, 16-40, and 8-40 in Fine-tree algorithm, SVM methods, and KNN methods. In our Time stamp methods, the tweets followed a similar trend (Fig 4.9). Our study of Classification of tweets resulted in the depressive, anti-depressive tweets, and corona confirmed cases (WHO data) showing an increase from Wednesday, which supported previous researchers' works.

5 SCALING DEPRESSION USING TWEETS

5.1 Introduction

Mental Health Illness will become the second leading cause of disease burden to the stakeholders and the government in coming years. Studies to understand and gauge depression levels will help in addressing the mental illness. Cramer and Becky [109] have reported a direct link between Social Media usage and Mental health issues. They reported a 70% increase in anxiety and depression in young people with the use of social media. Lin et al. [38] reported a significant relation between Social Media usage and increased depression. Passos et al. [39] have mentioned that there is a high prevalence between suicide and depressive symptoms. We [33] [41] [45] identified the depressive Twitter and behavior of individuals through the tweets and the relationship between illness and tweets. Suicide prediction tools are beneficial to relatives and friends of an individual, so that intervention of a Mental health Specialist will be placed to address the issue on time. Hence the Social Media data is a high value for machine learning and data mining research. Facebook and Twitter are top-ranked social media websites in today's world.

Several researchers attempted to estimate the levels of depression, and Kessler's method [92] is most popular among them a questionnaire, and these keywords were used in our study.

5.2 Data Collection

Tweets related to Kessler's questionnaire keywords are being collected every day and 11,953 tweets on a single instance are used for this study. The most popular antonyms for these seven keywords are too identified, and a set of 6,132 tweets related to these were collected using the Twitter API. Depressive and anti-depressive tweets data for the days of 20-March-2019, 27-

March-2019, and 4-April-2019. Tweets are cleaned as mentioned in Chapter -3 and these tweets are employed in this analysis.

5.3 Methods

5.3.1 Confusion Matrix

The confusion matrix is a table used to describe a classification on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand. There are four terms to understand. True positives (TP) are the set of data that are predicted true and actually true. True negatives (TN) are the set of data that we predicted negative and actually negative too. False positives (FP) are the set that we predicted positive, but they are negative. False negatives (FN) are the set that we predicted negative, actually positive.

We identified Predicted Positive tweets to Kessler's words (Predicted positive and actual positive - TP) and identified the antonym Kessler's keyword tweets (Predicted Negative and actual Negative – TN). We computed the number of antonym tweets present in the predicted Kessler keyword tweets (predicted positive but negative - FN) from the Kessler keyword tweets. Similarly, we also calculated Kessler keyword-related tweets in antonym Kessler tweets (predicted negative but actual positive – FP). Table – 5.1 gives the actual number of tweets in each category.

Table 5.1 Confusion Matric (March 20, 2019)

	ACTUAL	
PREDICTED	11731(TP)	134 (FN)
	218 (FP)	6002 (TN)

Table 5.2 Confusion Matrix (March 27, 2019)

	ACTUAL	
PREDICTED	8516 (TP)	24 (FN)
	147 (FP)	9706 (TN)

Table 5.3 Confusion Matrix (April 4,2019)

	ACTUAL	
PREDICTED	8016 (TP)	145 (FN)
	211 (FP)	21580 (TN)

Table 5.4 Statistical Parameters

Sl. No	ITEM	March 20, 2019	March 27,2019	April 4, 2019
1	Accuracy = TP / (TP + FN)	98.05	99.07	98.80
2	Sensitivity	0.9818	0.9830	0.9819
3	Specificity = TN / (FP + TN)	0.9782	0.9975	0.9902
4	Precision = TP / (TP + FP)	0.9887	0.9972	0.9739
5	Negative Predictive Rate = TN / (TN + FN)	0.0182	0.9851	0.9933
6	False Positive Rate (FPR) = FP / (FP + TN)	0.0218	0.0025	0.0098
7	False Discovery Rate = FP / (FP + TP)	0.0113	0.0028	0.0261
8	F-1 Score = 2TP / (2TP + FP + FN)	0.9852	0.9901	0.9779
9	Matthews Correlation Coefficient (MCC) = TP*TN-FP*FN/sqrt((TP+FP) *(TP+FN) *(TN+FP) * (TN+FN))	0.9568	0.9814	0.9696
10	No. of Tweets	18085	18393	29596

Similar data was collected on March 27, 2019, and again on April 4, 2019, to compare the days' tweet data. We found the values and showed them in Tables: 5.2 and 5.3.

5.3.2 F-1 Score and Matthews Correlation Coefficient

We computed accuracy, sensitivity, specificity, precision, negative predictive rate, false-positive rate, false discovery rate, F-1 score, and Matthews Correlation Coefficient from these data using standard formulae (Table 5.4). F1 score is the harmonic average of precision and recall and attains 1 (if perfect precision and recall) and will be zero, in the worst cases. MCC have values ranging between [-1,1]. High accuracy does not necessarily characterize a good classifier. F1-score does not consider TNs and hence does not depend on the full confusion matrix. Matthew's

correlation coefficient (MCC) depends on the full confusion matrix, and the value ranges from –1 (when the classification is always wrong) to 0 (when it is no better than random) to 1 (when it is always correct) [110] (Fig 5.1)

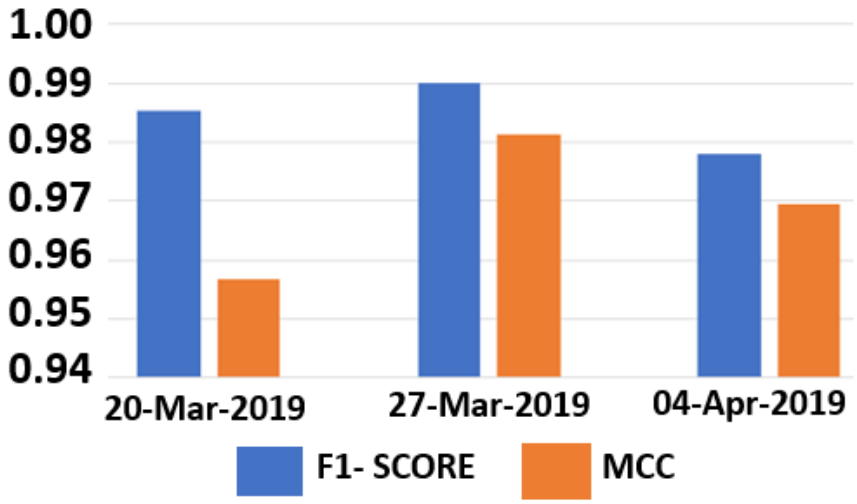

Figure 5.1 F-Score and MCC Value

5.3.3 Classification Index (C_I)

A new metric (Classification Index, C_I) computed using the formula $C_I = \frac{F1}{(1-MCC)}$.

We consider F1 as the numerator, close to 1 if there is perfect precision and recall in the data. We find the gap between 'always correct' to 'practical value' i., e. (1-MCC), which will be a smaller value if the data is close to correctness. This value represents the distance between the MCC value to 100% correctness. The new parameter formulated as $(C_I) = \frac{F1}{(1-MCC)}$ which is a quotient of the harmonic average of precision and recall with the slit between the perfect precision.

The values of C_I are presented in Table 5.5 in extreme conditions of F1 and MCC. F1 value becomes zero if the TP value is close to zero, and FP and FN values are very high. In such cases, the accuracy and precision values will be low. A good trade-off must exist for better

classification, or else our data results in false alarms or will fall under one category. The C_I is observed in the range of 0.5 to ∞ but varies with the accuracy and precision. We also identified the C_I value as a comparative classification index between the time series data collected over the period. The C_I value is suitable to compare data sets and for classification.

Table 5.5 F1-Score, MCC and C_I Value (Extreme Conditions)

F1	MCC	C_I
0	-1	0
0	0	0
0	1	∞
1	-1	0.5
1	0	1
1	1	∞

The statistical results like accuracy, sensitivity, specificity, precision is in a better range for classification performance in this case. MCC is the one that correctly considers the ratio of the confusion matrix size. Especially on imbalanced datasets, MCC can appraise the prediction evaluation is going well or not, while accuracy or F1 Score would not. MCC is used in machine learning as a measure of the quality of binary classifications. It considers true and false positives and negatives and is generally regarded as an equal measure. MCC can be called a correlation coefficient between the observed and predicted classifications. MCC also returns a value that indicates perfect prediction/random prediction or disagreement between forecast and observation. While F-1 scores in these two days' data are similar, the MCC values show a considerable change (Table 2.8). Zhu et al. [111] mentioned that feature selection improved using MCC values optimization methods. MCC is one stable method and yields the best results [112]. Researchers [113] studied and showed that F-Score and MCC gave better feature selection in their works.

Table 5.6 The Values of the new parameter C_I

Date	F1	MCC	$\frac{F1}{(1-MCC)}$
20-Mar-2019	0.9852	0.9568	22.80
27-Mar-2019	0.9901	0.9814	53.23
4-Apr-2016	0.9779	0.9696	32.16

The study demonstrated a significant impact on the value of C_I value. Table 5.6 shows the C_I values computed using the depression and anti-depression keywords. This analysis suggests that the new parameter 'C_I value' has an association with the tweeted data. F1 and MCC features are discussed above, and the parameter C_I is a combination of F1 and MCC; thus, it takes F1 and MCC properties. We considered the positive for depressed tweets and the negative for the anti-depressed tweets in the confusion matrix and computed the C_I value. Hence, the higher the CI value, is higher the depression on the tweets' day [45]. A software framework is set up to compute the classification index for given dates. To accomplish this process, we created a database with the tweets data on day-wise and keyword-wise. An interactive framework is prepared to take input date, collect the data from our database. This will exhibit the F1, MCC, and CI values of the desired date.

5.4 Discussion

We conclude that the higher the 'C_I value' value relates to the greater the depressing day. We thus also conclude that people are more depressed on 27-March-2019 than 20-March-2019. People on 4-April-2019 are less depressed than 27th March and more depressed while compared with 20th March. (Table 5.7)

Table 5.7 The Values of the new parameter C_I

Depression Index Levels (Comparison)
20-Mar-2019 < 04-Apr-2019 < 27-Mar-2019

The collection of tweets from a geographical area and comparing it with other areas is conceivable in case the tweet locations are too considered and analyzed. These studies will help identify the people's psychological effects while in natural calamities like fire, earthquakes, and floods. Using the framework, we can find the C_I value of each day and compare to the other days.

We can compare and thus determine the degree of distress among the people during that period, which helps to study psychotherapy (afterburn) and 'agitated depression.' Interpretation of these results could give economists and visionaries insight to plan and execute the societal programs in those geographical areas.

6 CLUSTERING OF TWEETS

6.1 Introduction

Clustering is a data analysis technique for discovering exciting patterns by dividing the data points into a few groups. Data points in the same groups are more like other data points in the same group and dissimilar to those in other groups. This allows classifying the data into structures that are more easily recognized and used. Topical clustering of tweets was attempted by Kevin et al. [114], Clusters are characterized through the most representative words, and association rules are used to highlight correlations among these words [115], hashtags have been used to cluster the tweets [116], Labelled the unlabeled tweets using clustering methods [117], and many other researchers worked in clustering the tweets and deduced the results.

6.2 Data Collection

We collected tweets from '*Dambulla,'* a central place of Srilanka, and its 500 Kms radius using Twitter API. This will suffice that all the tweets from Srilanka are collected. We identified seven keywords from Kessler [92], which were used in several Psychological Distress Surveys. Kessler, in his questionnaire, used the words depress, failure, hopeless, nervous, restless, tired, and worthless in the survey from the user and thereby measures depression levels. We also collected tweets related to seven antonyms for these keywords. 33,366 tweets were collected during April 14-30, 2019, with the hashtags #depress, #failure, #hopeless, #nervous, #restless, #tired, #worthless, #active, #calm, #comfort, #delight, #excite, #hopeful, #peaceful and #bomb. We have also collected Anti-depression tweets and bombing-related tweets from April 14-30, 2019. We compared our results with 18,096 tweets from a normal period (May 14-25, 2019), related to all the same keywords.

6.3 Methods

6.3.1 Gradient-Based Method

The contributing ratio of a depressive tweet with '*bomb'* tweets are calculated using the formula for each keyword *i*, $C_i = \frac{(ni/\Sigma ni)}{(nb/\Sigma nb)}$, where ni is the number of tweets in that day, Σni is the total number of tweets of the keyword for all days, nb is the number of tweets related to '*bomb'* on that day, and Σnb is the total number of tweets of '*bomb´* keywords for all days. Fig. 6.1 shows these ratios of each day. On the 'Bombing days,' i.e., 21st and 23rd April 2019, all the keywords merge at one point.

Figure 6.1 Ratio of depressive tweets with bombing tweets

Figure 6.2 Ratio of depressive and Bombing tweets with total tweets

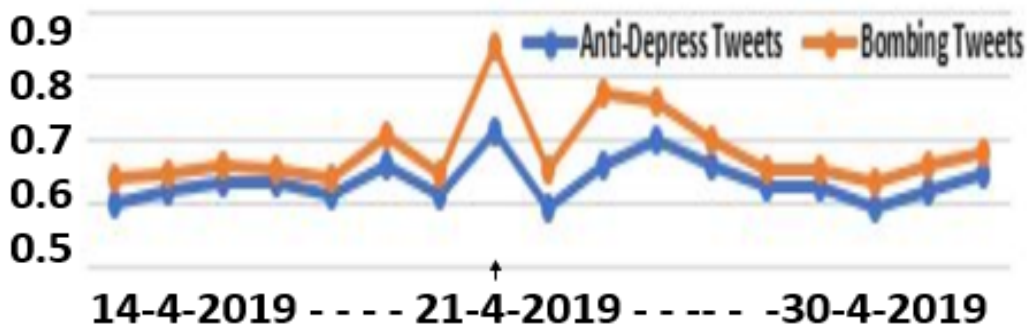

Figure 6.3 Antidepression and Bombing tweet rations to the total tweets

The ratio of tweets related to the depressing category and bombing category is shown in Fig. 6.2. Similarly, the ratio of tweets associated with the anti-depressant category and bombing category is shown in Fig. 6.3. Pearson Correlation coefficient (PCC) was computed between the two sets of tweets with 'bomb' tweets collected during April 14-30, 2019, and May 14-24, 2019.

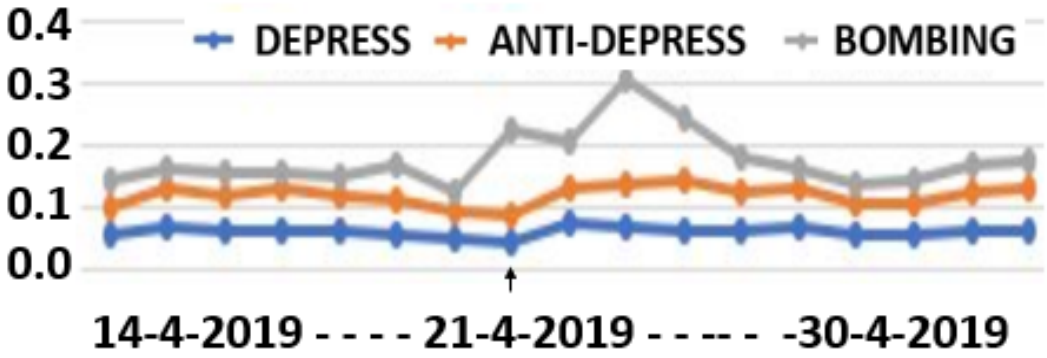

Figure 6.4 Probability of Depressed tweets, Anti-depressed and Bombing tweets

$PCC = [i*\sum_{i=1}^{n} k*b_i - \sum_{i=1}^{n} k*\sum_{i=1}^{n} k\ b_i]/\sqrt{(i*\sum_{i=1}^{n} k\ i^2 - (\sum_{i=1}^{n} k)^2 * (i*\sum_{i=1}^{n} bi^2 * (\sum_{i=1}^{n} bi)^2}$ for each i^{th} day, K the keyword, and B is the Bomb keyword. We calculated the probability of a tweet to be considered as 'depressed,', 'anti-depressed,' and 'bombing' category. The results are shown in Fig. 6.4. We considered, if $((PCC_k)_{d1})$ and $(PCC_k)_{d2})$ both positive or negative, then $k \in A_{new}$ or else $K \in B_{new}$, where k is the keyword, and d_1 and d_2 are two periods of tweet collection. We found a positive correlation of tweets of some keywords with bomb tweets, and others showed a negative correlation. We noticed that Failure, Nervous, Comfort, Delight, and Peaceful are in the

similarity group (A_{new}), and Depress, Hopeless, Restless, Tired, Worthless, Active, Calm, Excite, and Hopeful are in the dissimilarity group (B_{new}). We attempted to support our clustering by using the Learning Quotient method, Keyword Contribution factor, Text mining Methods (Term Document Matrix), Confusion Matrix, Accuracies, and Association factor.

6.3.2 Learning Quotient Method [Q]

We computed a Learning Quotient value [Q_i] for each day, for similarity category (S_i) and dissimilarity category (DS_i) as below. $Q_i = P_i * ln\ (P_i)$; where $P_i = S_i / (\sum_{i=1}^{n}(S_i) + \sum_{i=1}^{n}(DS_i))$. The learning quotient of the two categories is shown in Figures 6.5 and 6.6.

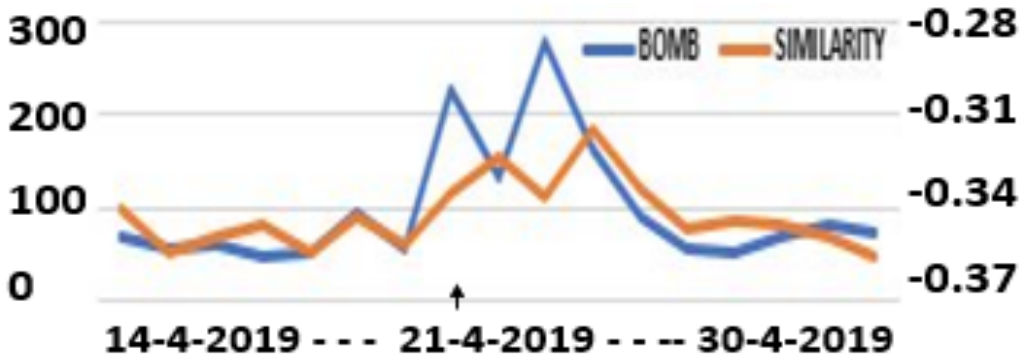

Figure 6.5 Bomb tweets Vs. Q_i in Similarity keyword set

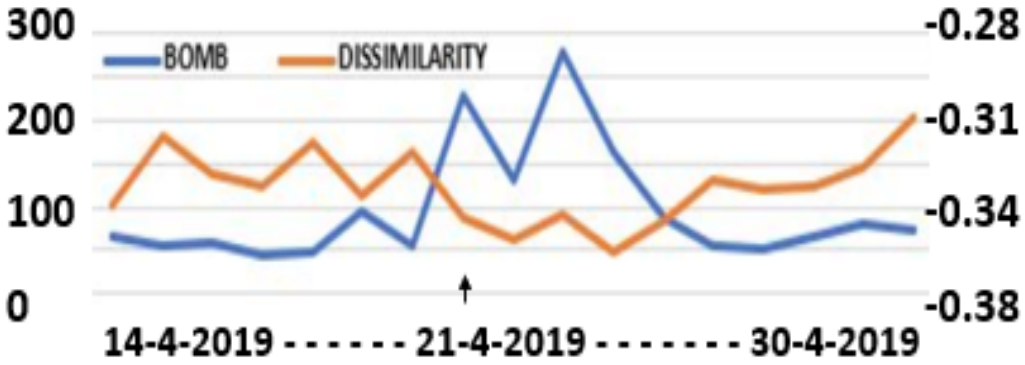

Figure 6.6 Bomb tweets Vs. Q_i in Dissimilarity keyword set

The similarity tweet keyword group resulted in a good correlation pattern (Fig. 6.5), whereas the dissimilarity set of tweet keyword groups did not show any pattern with '*bomb'* tweets

(Fig. 6.6). We noticed that Failure, Nervous, Comfort, Delight, and Peaceful are in the similarity group and Depress, Hopeless, Restless, Tired, Worthless, Active, Calm, Excite, and Hopeful fall in the dissimilarity group.

6.3.3 Keyword Contribution Factor (KCF)

We considered two input sets A with keywords {depress, failure, hopeless, nervous, restless, tired, worthless} and set B with keywords {active, calm, comfort, delight, excite, hopeful, peaceful}. Using the following formula, we computed the keyword contribution factor (KCF) of the set A and B. KCF of A = A_i * $\sum_{i=1}^{n} A_i$ / ($\sum_{i=1}^{n} A_i + \sum_{i=1}^{n} B_i$) for each i^{th} day. KCF values of the sets A, B, A_{new}, B_{new} are 0.38, 0.62, 0.52, 0.48 respectively.

6.3.4 Text Mining Methods

Tweets representing #depress, #failure, #hopeless, #nervous, #restless, #tired, #worthless, #active, #calm, #comfort, #delight, #excite, #hopeful, #peaceful and #bomb during April 14-30, 2019, and May 14-25, 2019, from Twitter API were collected. Tweets originated from '*Dambulla'* were also collected for the same keywords. We also collected '*bomb'* related tweets too during April 14-30, 2019. These five sets of tweets for each keyword were analyzed using NLP methods. We created a *corpus* for each of these selections and removed the 'punctuations', 'blank spaces', 'converted to lowercase', 'stop-words' removed, and word-stemming is applied using the '*tm'* R-package. A Term-Document-Matrix (TDM) was obtained for each set of tweet data.

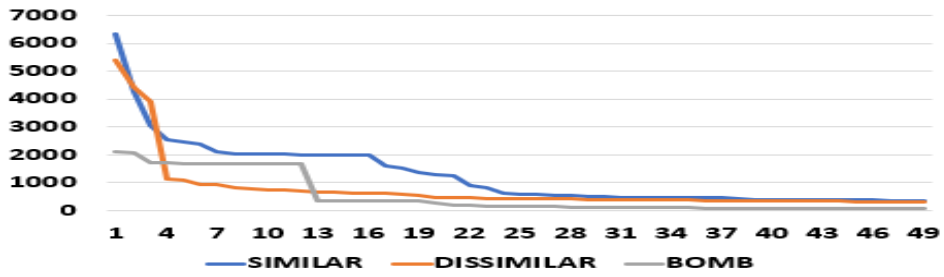

Figure 6.7 Bomb tweets Vs. Q_i Keyword set

We grouped the TDM data of similar and dissimilar sets separately and ranked with their word frequency. We considered the 300 frequently used words from these sets and compared them with the bombing tweet word set. Fig. 6.7 shows the pattern of similarity group, dissimilarity group, and bomb tweets. We computed a Confusion Matrix for this data and presented it in Table 6.1.

Table 6.1 Confusion Matrix from TDM data

	Actual	
Predicted	TP=27	FP=273
	FN=118	TN=7262

We calculated accuracy from the confusion matrix using the formula Accuracy $AC = \frac{(TP+TN)}{(TP+FN+FP+TN)}$ and observed that the accuracy is 94%.

6.4 Association Factor (AF)

We computed a factor *AF = (A*B)/(A^2+b^2-A*B)* between a Similar set and Bomb Set and Dissimilar set with a bomb set, and the results are presented in Fig. 6.8.

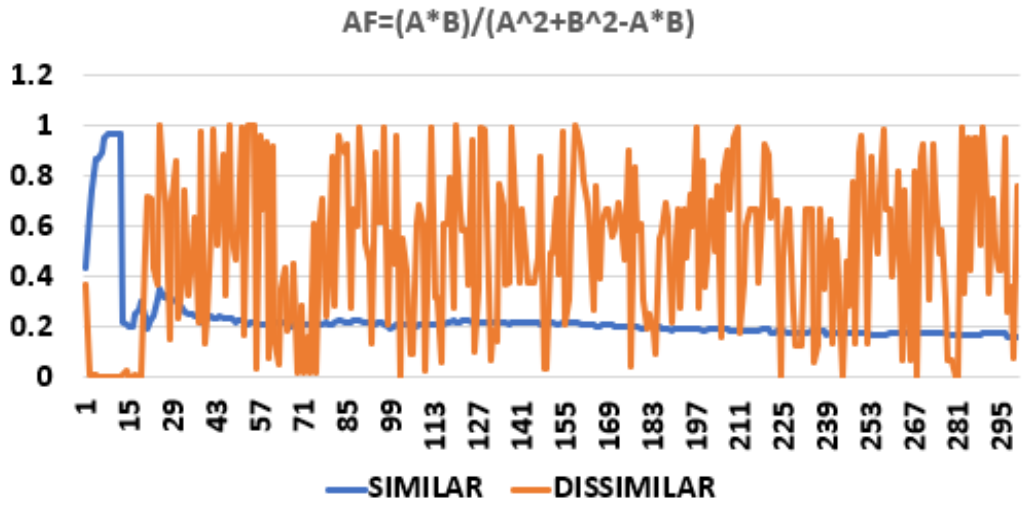

Figure 6.8 Association factor between Similarity/dissimilarity and Bombing set tweets

6.5 Discussion

The merger of ratios in Fig. 6.1 reflects the contribution of the keyword on Bombing Day. There is a visible indication that the proportions of depressing tweets and bombing tweets on 21st April 2019 (Bombing Day) are merging. Since the bomb-related Tweets are many, the keywords combine near low values compared to the other keywords on the bombing day. We conclude that more depressive keywords have a similar phenomenon. There is also a significant change in the ratio of Depress-related tweets to Total tweets collected during the bombing days (Figures 6.2 and 6.3). Probability of Depressed Tweets, Anti-depressed Tweets, and bombing Tweets are shown in Fig. 6.4. There is a similarity in the tweets in depressed and anti-depressed tweets in all the days. However, it is noticed a change in bombing tweets on the day of the bombings. Ratios of depressive tweets with Bomb related tweets are shown in Fig. 6.1, Depressive and Bombing tweets with the total number of tweets ratio computation in Fig. 6.2, Anti-depression and Bombing tweets with the total number of tweets ratio computations in Fig. 6.3, and Probability of Depressive and Anti-depressive tweets with bombing tweets in Fig. 6.4. All these supported that these keywords are in good correlation with Bombing hashtag tweets in our observation. We noticed that the hashtags with the keywords Failure, Nervous, Comfort, Delight, Peaceful, Depress, Hopeless, Restless, Tired, Worthless, Active, Calm, Excite, and Hopeful are associated with the ‘Bomb’ hashtag keyword. Pearson Correlation coefficient was calculated within the keywords with the ‘Bomb’ keyword, and we grouped them into two clusters.

Learning Quotient with ‘bomb’ tweets with the similarity keyword set and dissimilarity keyword set during April 14-30, 2019 (Bombing Day is 21st April 2019) is shown in Figures 6.5 and 6.6, respectively. Failure, Nervous, comfort, delight, and peaceful showed similarities, and depress, hopeless, restless, tired, worthless, active, calm, excite, hopeful showed dissimilarities in

correlation in the two spells of our observation. Our assumption of grouping, Failure, Nervous, Comfort, Delight, and Peaceful into one category and Depress, Hopeless, Restless, Tired, Worthless, Active, Calm, Excite, and Hopeful into another group was supported with our results shown in Figures 6.5 and 6.6. The similarity tweet set followed the bombing assimilation tweet pattern in almost all the days of observation, while the dissimilarity tweet set did not follow the same pattern. The keyword Contribution Factor (KCF) values of sets A, B, A_{new}, and B_{new} are 0.38, 0.62, 0.52, and 0.48, respectively. In our clustering method, the KCF value of A moved to 0.38 to A_{new} 0.52; B value from 0.62 to B_{new} to 0.48. Also, the change of KCF value in A and B resulted in the balancing of the sets. Such balanced clustering of sets helps to yield the highest-level accuracy (the average of True Positive Rate and True Negative Rate) [118]. Term document matrix data is analyzed to study the pattern of the tweets in these two groups. A good pattern match is noticed in the frequently used word pattern in the similarity group, dissimilarity group, and bombing tweets (Fig 6.7). The accuracy obtained from the clustering from the confusion matrix is 94%. The results obtained from the association factor value are computed and presented in Fig. 6.2 indicated an excellent pattern match between the similarity and the dissimilarity groups. The above results support the study of the clustering of tweet keywords into two groups (similarity and dissimilarity).

The correlation values of ’Keywords’ with ’Bombing’ tweets resulted in clustering among the keywords into similarity and dissimilarity groups. We could identify that the five keywords (failure, nervous, comfort, delight, and peaceful) are in a similar set, and nine keywords (depress, hopeless, restless, tired, worthless, active, calm, excite, hopeful) are dissimilar set. People tweeted more with depressive tweets than non-depressive tweets during bombing days. The results obtained and presented in Figures 6.5 to 6.8 show an excellent match to our clustering of keywords

from the combined depressed and anti-depressed keywords. We conclude that most people who tweeted with the hashtag ‘bomb’ coincides with a set of similar tweets. Our work considered two sets of keywords belonging to depression and anti-depression and clustered into two balanced sets as similar and dissimilar in association with the ’event.’

Monitoring and real-time interfacing similarity set users' tweet data help in extending psychological assistance. Micro-blogging sites need to develop a framework to capture tweets that are going abnormally and counteract them with socio-medical aids. The clustering algorithms and studies can be extended to understand the relations within the keywords. Studies can also be taken with ubiquitous sets for clustering to improve partition-based models' applications. We clustered on the PCC values and depending on the requirement. We can cluster them on many output sets by defining different ranges between -1 to +1.

7 IMPACT OF TWEETS ON MENTAL HEALTH

7.1 Introduction

Social network usage has increased in recent years because there is support from all ages and groups worldwide. Several social networking sites provide numerous GUI applications to the users. Twitter supports 300 million users with more than 500 million tweets in a day. Messages posted as tweets on Twitter have many lifestyles, emotions, business, political, health tips, festivals, and events. Users post messages to express their views, and these social networking sites become a platform for expression. Mental illness disability is one significant cause worldwide in the future [4], [119], [120]. Several researchers [121] [122] [123] [38] [124] showed an association between social media use and depressive symptoms, and tweets from Twitter are sources to study and analyze insight about mental health [125] [124] [126]. Researchers studied the behavior of individuals through tweets, and the relationship between illness and tweets was studied [33]. Depressive and anti-depressive keywords are categorized using learning and correlation coefficients of the tweet in an event [55] [127]. Tweets are used to find real-time events [84] during earthquakes. Many researchers used event detection techniques and addressed using n-gram analysis [128] [129], Latent Dirichlet Allocation methods [130] [131], and bag-of-words methods [132], [133].

We chose the Sri Lanka bomb blasts in 2019, Burevi Cyclone in 2020, and Tauktae cyclone in 2021 as events for our study. All these events have significantly traumatized the families both financially and emotionally in Sri Lanka. Srilanka witnessed bomb blasts in April 2019, and nearly 300 people died, and many were injured. Tourism in Sri Lanka lost $1.5 billion due to this event [134] [135] . The Burevi cyclone has a wind speed of 60 miles per hour and gusts up to 70

miles [136]. Tauktae is reported to be the fifth strongest storm observed in the Arabian sea since 1998 by the U.S. Joint Typhoon Warning Center, with sustained winds of 125 miles per hour and gusts up to 145 miles [137]. Classification of events using tweets was carried out by several researchers in the past [138] [139] [140].

7.2 Data collection and cleaning

Kessler's psychological distress scale [92] is one of the most used methods to understand individuals' mental health status. We considered seven keywords from Kessler's questionnaire and collected the tweets related to keywords using #hash tag with a specified set of keys/tokens of Twitter. Depress, failure, hopeless, nervous, restless, tired, and worthless are the keywords chosen from Kessler's works. We identified the antonyms of these seven keywords (active, calm, comfort, delight, excite, hopeful and peaceful) and the event keyword. 90,649 tweets are collected for this study from 500 Kms around *Dambulla*, thus covering the entire Sri Lanka geographical area. 34,555 tweets during Srilanka bombing time, 17,384 during Burevi, and 38,710 during Tauktae time. We cleaned the tweets by removing the blank tweets and having html links.

7.3 Methods

We grouped the tweets of depress, failure, hopeless, nervous, restless, tired, and worthless as '*depressive set*' and active, calm, comfort, delight, excite, hopeful and peaceful as '*anti-depressive set*' and '*event*' as separate sets. The gradients of the tweets data are shown in Figures 7.1 to 7.3 for Sri Lanka Bomb blasts, Burevi cyclone, and Tauktae cyclone. We normalized these data set values between 0 and 1. Normalized depressive and anti-depressive tweet values during the Srilanka bombing, Burevi cyclone, and Tauktae cyclone are shown in Figures 7.4 to 7.6. The linear equation of these data sets is retrieved. The Area Under the Curve (AUC) is computed and

shown in Fig. 7.7. We computed the correlation coefficient between the normalized depressive and anti-depressive tweet data with the event data (Table 7.1).

Table 7.1 Correlation coefficient of Sri Lanka Bomb blasts, Burevi cyclone and Tauktae cyclone with the event

	Depressive Tweets	**Anti-Depressive Tweets**
Srilanka Bomb Blasts	0.7294	0.4125
Burevi Cyclone	0.8362	0.3106
Tauktae Cyclone	0.8322	0.6012

The tweets data shown in Figures 7.1 to 7.3 for Sri Lanka Bomb blasts, Burevi cyclone, and Tauktae cyclone indicate a remarkable change in depressive tweets and anti-depressive tweets patterns. An online live data plot analysis can predict such a change and may lead to identifying an event. Normalized tweet data of depressive, anti-depressive, and event tweets along with date wise (Figures 7.4, to 7.6) are considered for computation of AUC with 'Total Period' which is under regular conditions, 'before the event,' 'during the event' and 'after the event.' The change in the AUC of Event vs. depressive and anti-depressive tweets in the four different states is computed (Fig. 7.7), revealing the event's effect at normal conditions, before, during, and after the event.

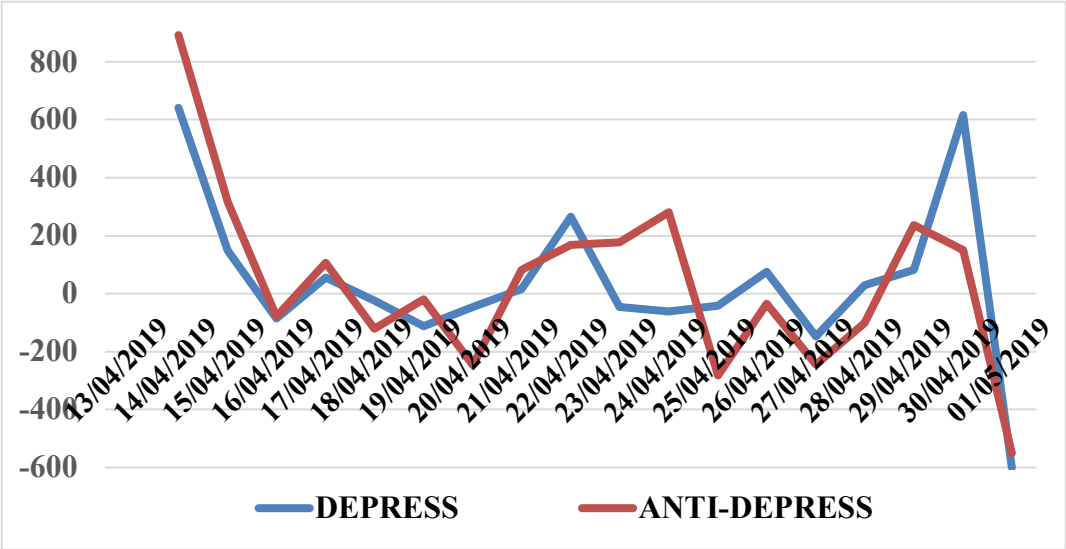

Figure 7.1 Gradient values of Depressive and Anti-depressive tweets during Sri Lanka Bomb blasts

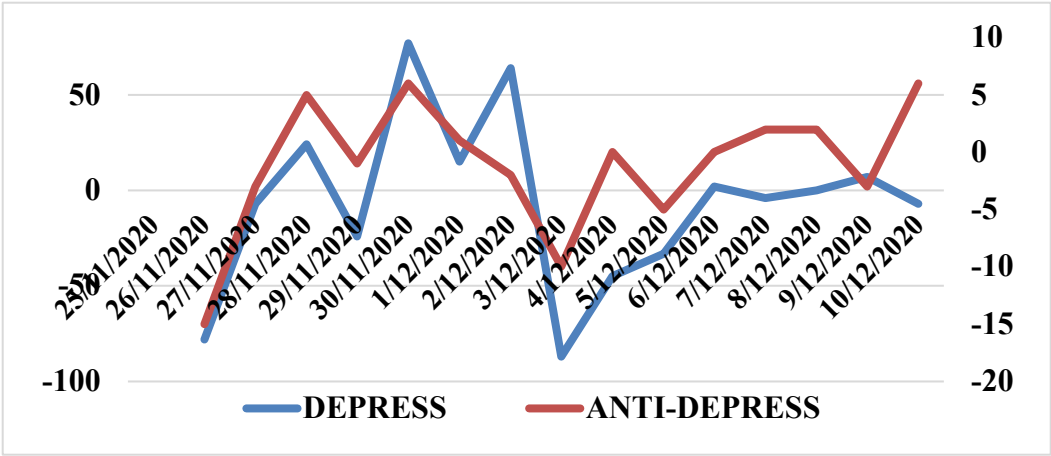

Figure 7.2 Gradient values of Depressive and Anti-depressive tweets during Burevi cyclone

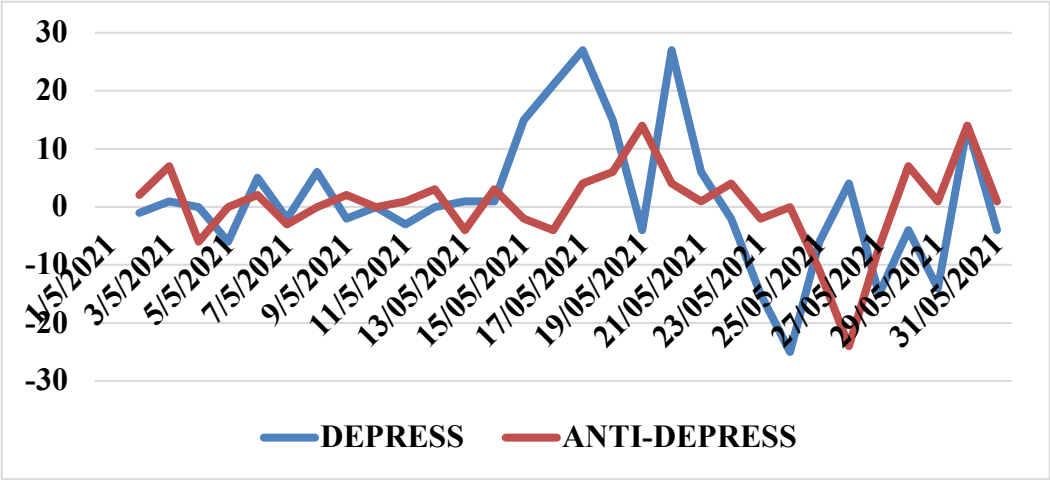

Figure 7.3 Gradient values of Depressive and Anti-depressive tweets during Tauktae cyclone

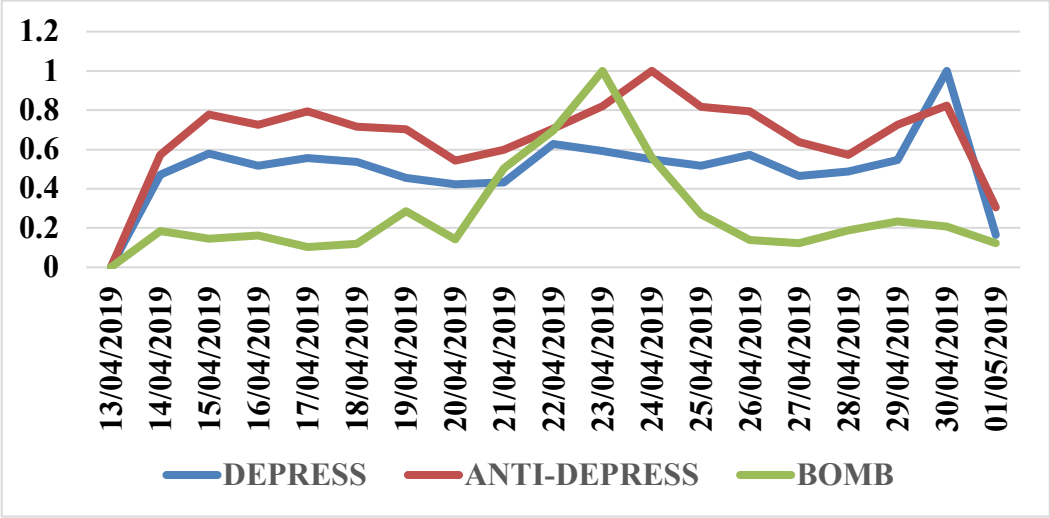

Figure 7.4 Depressive and Anti-depressive tweet curves during Sri Lanka Bombing

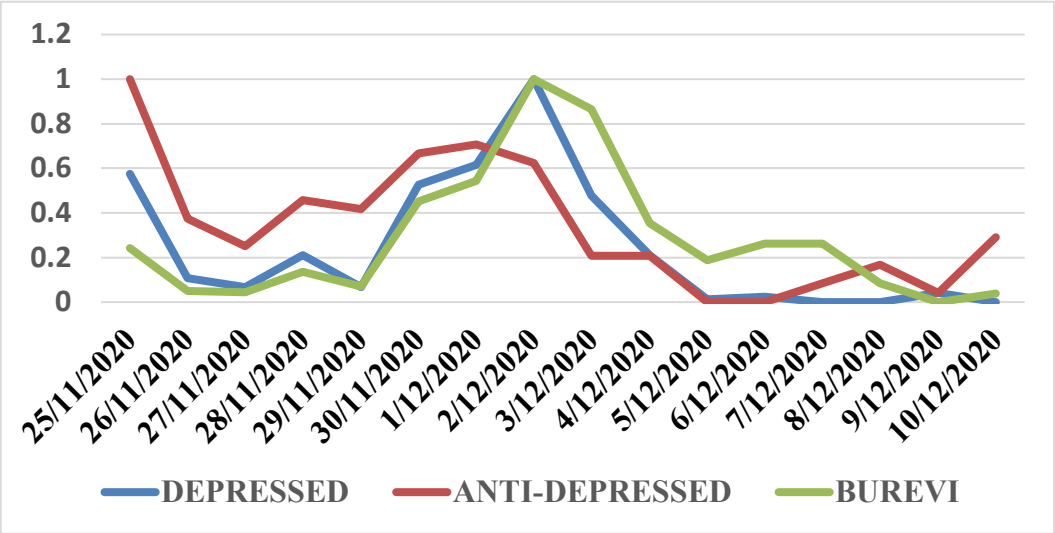

Figure 7.5 Depressive and Anti-depressive tweet curves during Burevi cyclone

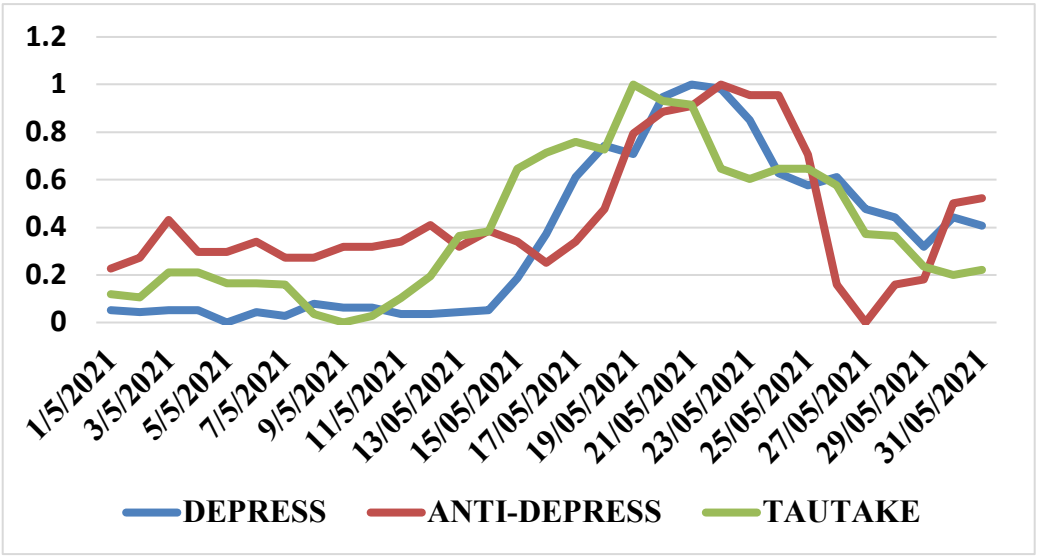

Figure 7.6 Depressive and Anti-depressive tweet curves during Tauktae cyclone

The rate of change from 'before the event' to 'during the event' is shown in Table 7.2. It indicates the event's effect on the depressive tweets and the people affected by the event. Similarly, the rate of change from 'during the event' to 'after the event' in Table 7.2 affects tweets to show the people's recovery rate.

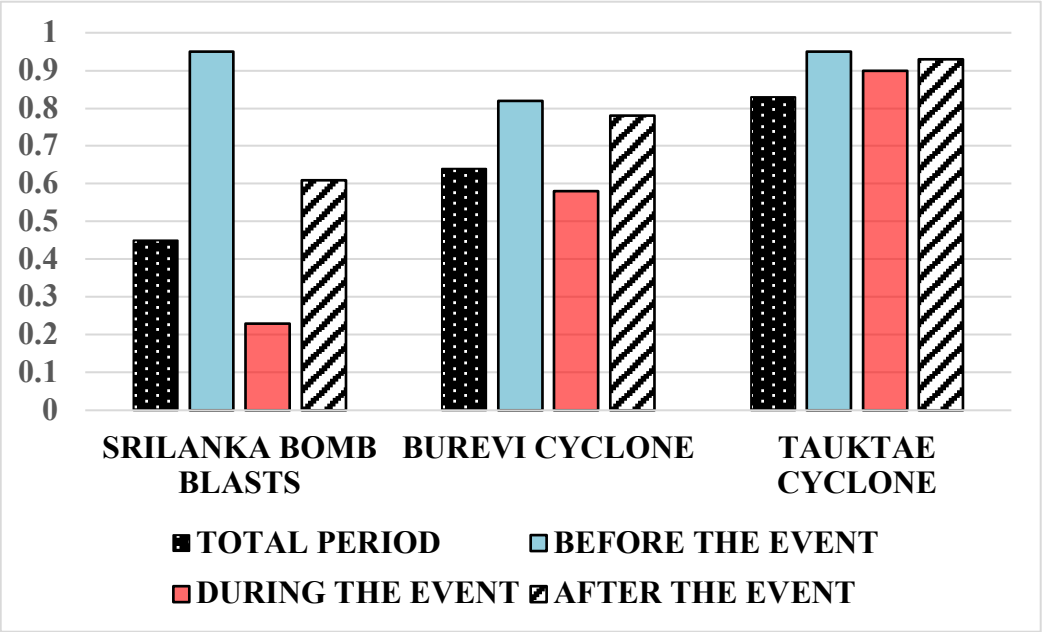

Figure 7.7 Impact of the event during Sri Lanka bomb blasts, Burevi and Tauktae cyclones

Table 7.2 Effect of Sri Lanka Bomb blasts, Burevi cyclone and Tauktae cyclone from our study and other resources

	Our Study		***Other Sources***
	Effect	***Recovery***	
Srilanka Bomb Blasts	24%	37%	GDP growth reduced 28% [135]
Burevi Cyclone	70%	74%	Winds 60 miles per hour and gusts up to 70 miles [136] 75,000 people evacuated [139]
Tauktae Cyclone	94%	96%	Sustained winds of 125 miles per hour and gusts up to 145 miles [137] 2,00,000 people evacuated and loss of US 2.1 billion [140]

7.4 Space Tourism Tweets

The space tourism aircraft, Unity 22, of Virgin Galactic, made its space travel mission on 11-July-2021. Blue Origin has its maiden tourist launch with New Shepard on 20-July-2021. Depressive tweets and Anti-depressive tweets are collected using Twitter API from 6-July-2021

to 6-August-2021. Tweets with 'Corona', 'unit22', 'Blue Origin', 'New Shepard' were also collected. Covid-19 confirmed cases are obtained from WHO website [141]. The tweets belonging to a combination of depress, failure, hopeless, nervous, restless, tired, and worthless are shown as '*depressive*' (blue line) and active, calm, comfort, delight, excite, hopeful, and peaceful as '*anti-depressive* (red line) in Fig 7.8.

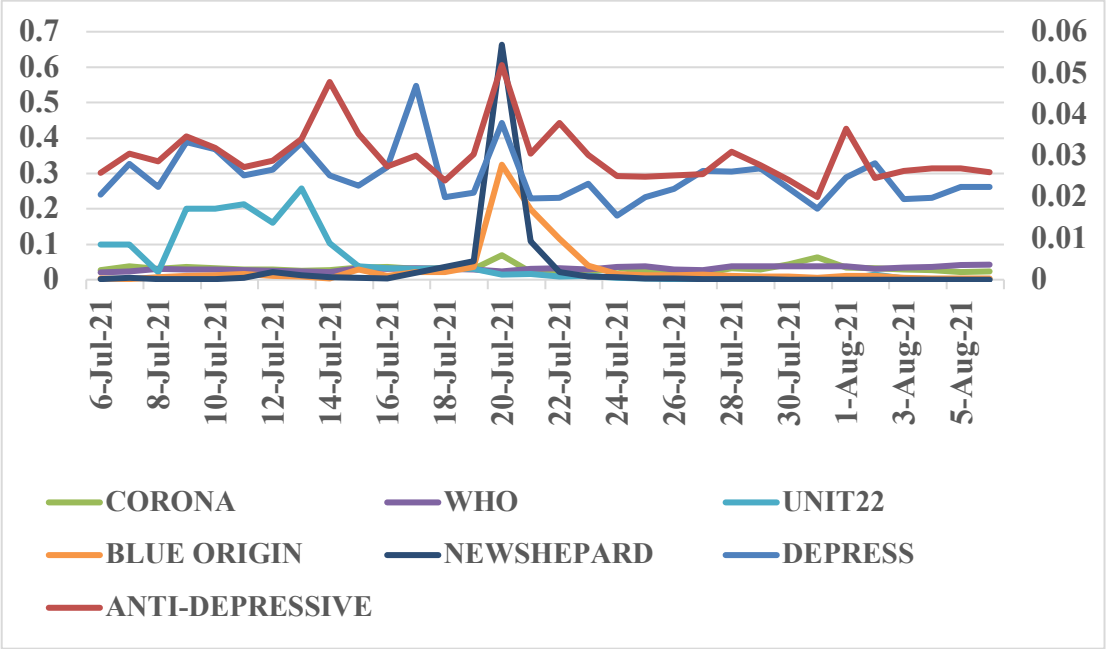

Figure 7.8 Tweets during 6-July-2021 to 6-August-2021

Depressive and anti-depressive tweets are considered a tool to understand people's mindset about Unit 22, Blue Origin, and New Shepard.

Table 7.3 Correlation coefficient of depressive and anti-depressive tweets with the space tourism

	Unit 22	Blue Origin	New Shepard	Corona	WHO data
Depress	0.38	0.20	0.35	0.22	-0.34
Anti-depress	0.21	0.58	0.61	0.26	-0.41

Unit 22 is the first space craft deployed, and people are more tensed about the success and hence shown a high correlation with 'depress' tweets and low correlation with anti-depressive

tweets. The higher correlation coefficient between the anti-depressive (happiness) and 'blue origin/New Shepard' shows that people are confident in predicting a success story (Table 7.3). During these times, corona and Covid confirmed cases do not affect depression and anti-depressive tweets.

7.5 Discussion

It is observed that the Tauktae cyclone created more havoc than the Burevi cyclone [142] [143]. Our results also correlate with the effect of events reported by other sources [136] [137] [142] [143]. The Pearson correlation confirms that there is a perfect correlation between depressive tweets and events. Collection of tweets online and computation of the impact will help governmental, non-governmental bodies plan and support the disaster preparedness and emergency team management in rescue operations during cyclones, forest fires of a particular geographical area,

Our study shows that the Tauktae cyclone impacted more than the Burevi cyclone, and the Burevi cyclone affected more than Srilanka bomb blasts incident in peoples' mental health (Table 7.2). Similar results are noticed from Fig.7.7 about these three events. The impacts '*before,' 'during'* and *'after'* the events are higher in the Tauktae cyclone than in the Burevi cyclone and Bomb Blasts incidents. The works are done by other researchers [135] [136] [137] [139] [140] also support our findings. Our findings help scale the impact of an event using mental health-related tweets.

Space tourism is a classic example to delineate people's behavior. During Unit 22 space travel, many people showed a tense mood and were more depressed than the timings of Blue origin/New Shepard.

8 HEDONOMETRICS

8.1 Introduction

The ease of availability of the internet networking devices and their affordability increased the interaction with the social network sites considerably in the daily life of human beings. People take pleasure in posting text and pictures about their emotions, events, and valuable tips. Researchers analyzed the social network data to reveal mental health disorders, suicidal and depressive behavior [40] [41] [79]. Twitter is one of the most popular social network sites that facilitates people building relationships with experts in many disciplines, promoting research, product, and feedback. The Twitter tweet consists of a maximum of 280 characters. Twitter also provides APIs that allow the collection of tweets data to be much more straightforward than other social network platforms.

The study of wellbeing (happiness) has a long history in human evolution. Aristippus, a Greek philosopher, thought that the goal of life is to understand and feel happiness. Quantitative determination of Euphoria, a state of happiness [144], was attempted in 1935 itself. Life Satisfaction Index [145], Rosenberg Self Esteem Measure [146], Geriatric center Morale scale [147] are earlier scales that are considered to measure the happiness or wellbeing of humans.

Happiness is defined as the outcome of the pursuit of pleasure over pain [148]. Shin and Johnson [149] categorized resources to scale happiness into three types: (a) data belonging to age, sex, and race, (b) data related to income and education, and (c) social relations (family and friends). With time and technological growth, the parameters that are considered to measure happiness are restructured. Substantial progress in measuring happiness using self-reports and finding how wellbeing is affected by various life factors is reported earlier [150]. Social Media use has increased

multifold in recent years as various smart mobile devices connect the Internet, and such services are affordable. In 2005 social media adoption was just 5% of American adults, increased to 50% by 2011, and today 72% of the public participate in social media activities [151]. The extensive use of social media is aggravating mental health problems [152]. Sharing opinions, expressing feelings, getting in touch with family and friends, and participating in social media for business and other recent issues drove a person to spend more time with social media sites. Social media became a vital communication tool to reach people, in touch and lost in the past with its technological fuse in Internet-based applications. Mental health and Social Media use are interrelated. It observed that more time spent with social media is prone to psychological distress, and extreme usage may lead to suicidal ideation [152].

Many social media posts are related to emotional issues, feelings, and individual reactions to recent topics. Thus, Happiness and Unhappiness are two such parameters that take a significant role that an individual tends to post. Happiness appears to be abstract from several centuries of human life. The issues, circumstances, and events in the recent past reflect Happiness in one's life. Gross National Happiness (GNH) evaluates the quality of a country with different values and considers that the development takes place with material and spiritual growth [153].

Depression is one of the psychological behaviors that human being represents in their actions and relates to Happiness. Several researchers [92] [154], [155], [156] [157] [158] worked in scaling the depression levels, and out of these, Kessler's works [92] are the most used methods in mental health depression measurements [159]. Davitz suggested in his report [160] that Happiness was most often associated with 'pleasant mood-states.' Bradburn [161] reported that Happiness influenced mental health both positively and negatively. Irwin [162] mentioned that Happiness was measurable by asking people how happy they were. Kammann [163] and

Underwood [164] identified multi-parameter scales, and Oswald and Wu [165] revealed the positive correlation between objective life satisfaction measurement in one's life. Mental disorder is the leading cause of unhappiness in modern society, and investment in mental health care is likely to add to average Happiness [166]. Galati et al. [167] studied to identify the subjective components of Happiness and analyze their degree of attainment in two countries, Italy and Cuba.

Income, education, occupational prestige, social ranking, appreciation within society, ethical values, children, physical and mental health, and living conditions are few parameters responsible for the Happiness of a human being [153]. Socio-economic circumstances also impact the likelihood and causes of developing a mental disorder. Lower socio-economic status has a more prevalence of mental disorders, depression, and anxiety [168]. We need to measure the positive and negative emotions separately to appreciate people's Happiness [169]. Dodds et al. [170] analyzed the tweets from Twitter and explained the variations in Happiness and information levels. People's Happiness depends on the others with whom they are associated [171]. Kircanski et al. [172] and Barrett [173] explored the implications of language's role in emotion concept acquisition and use for emotional experiences and perceptions. Wright et al. [174] and Rose et al. [175] used labels to study mental disorders. Data mining techniques are applied, and the keywords are categorized into different groups with the tweets from Twitter data [55]. People use bad words when they are depressed and sensitive to contextual emotions [176]. Such tweets contain few bad words as text, and we studied these tweets with bad words as a contributing factor for the feelings.

There are many happiness indexes scales in which the Subjective Happiness Scale [177] and Steen happiness index [178] became popular. Subjective Happiness Scale (SHS) measures an individual's happiness through self-evaluation [177]. SHS considers a four-item with a seven-point scale. Steen Happiness Index follows the principle of the Beck model [179] in computing

the changes in happiness. SHI takes twenty survey items and considers the pleasant, engaged, and meaningful life to measure depression [178]. Many North Americans think about happiness at least once each day [180].

Cantril's self-anchoring scale [181] takes eleven steps of understanding of best and worst experiences, Affect Balance Scale [182] considers five positive effects and five negative effects to compute the well-being of people. Gurin uses the perceived problems [183], and the Memorial University of Newfoundland Scale of Happiness (MUNSH) index utilizes more new items to measure happiness in people. Wellbeing is often investigated as the outcome or dependent variable in these studies but rarely studied as a forecaster.

All the above works depend on a survey and/or answering a questionnaire or interaction with people. The number of questions varies, and the content framework varies in different Happiness Index Scales. The accuracy of the Happiness Index depends on the input values of the user interactively. Subjective Happiness Index uses four items in 0-7 scale [177], Steen Happiness Index scales twenty items in 0-5 scale [178], Cantril self-anchoring scales combination of three groups (thriving, struggling, and suffering) with different scales [181]. Affect Balance scale [182] considers ten items with negative and positive values for measurement. Such varied scaled parameter measurements will impact the Happiness Index with a slight change in input data.

Online surveys have advantages like the ease of data gathering, minimal costs, input data automation, the flexibility of parameters, remotely administrated, and many respondents. Still, a few disadvantages exist like the non-suitability for open-ended questions, absence of interviewer, inability to interact in challenging stage, a respondent may not provide an accurate and honest answer, and survey frauds. In addition, there could be technology-originated frauds like robot answering, non-compatibility with the IoT devices and network issues, etc. Accuracies of web

survey limitations were discussed [184], and researchers found that online surveys may mislead if the sample is contaminated population [185]. Advantages of online surveys are discussed [186] in normal conditions, but authors agree that different benchmarks, different scaling factors will impact the result undependable [187].

Face-to-Face survey data collection allows more in-depth comprehensive data collection, accurate screening of the respondent, and can capture verbal and non-verbal emotions and behavior information. However, there are disadvantages like more time-consuming, biased responses, expensive and limits to sample size. Pitfalls and oversights in survey data collection can be avoided using good practices [188]. The methods and the survey patterns mentioned above are 'input' mode estimates. Happiness is an outcome over pain [148], and results will be more appropriate if we compute the Happiness Index from 'output' parameters. Outcome measurements will impact the best match in clinical trials [189], and it is critical in underlying happiness [190].

So far, no work has been reported earlier to scale Happiness Index using the mental health-oriented tweet data. Depression is considered as unhappiness, and Kessler [92] discovered a method to scale the depression levels with a questionnaire with depress, failure, hopeless, nervous, restless, tired, and worthless keywords.

We analyze the tweets related to depressive and anti-depressive to discover tweeting patterns and their impact on happiness. We also compare this new Happiness Index parameter to 'Sri Lanka,' 'New York', and 'Georgia' state geographical areas for two different periods, and the results are in good correlation.

8.2 Data

The tweet data used for our study consists of 2.3 million tweets belong to the period from 01-April-2020 to 01-April-2021 (366 days). Our hedonometric study considers two weeks of data for the period of 4-July-2021 to 10-July-2021 (HI-1) and 25-July-2021 to 31-July-2021 (HI-2). In addition to these data, Depressive and Anti-depressive Tweets belonging to the graphical areas of Srilanka, New York, and Georgia states are collected using the Twitter API. A 500 Km radius from '*Dambulla,'* a central place, is chosen (Latitude 19.0760N, Longitude 72.8777E) for the collection of Sri Lanka tweets, Tweets from 600Kms area from 40.785091N,73.968285W, and 500 km area from 33.771629N, 84.418553W are considered as tweets of New York and Georgia, respectively. The radius is chosen to cover the entire geographical area in that state. Tweets are cleaned, and duplicate tweets are removed. We also collected depressive and anti-depressive data for one day for all the 50 states of the USA separately to prepare a Happiness Map.

8.3 Methods

8.3.1 Happiness Index

Kessler's Psychological Distress scale [92] is one of the methods to understand an individual's mental health status. We used Twitter API and collected the tweets data with seven keywords from Kessler's questionnaire. The seven keywords chosen from Kessler's works are depress, failure, hopeless, nervous, restless, tired, and worthless characterized as depressive. The active, calm, comfort, delight, excitement, hopeful and peaceful antonyms of these seven keywords as the Anti-depressive category. If more depressive tweets are posted, it shows that people of that part of the geographical area are depressed. Similarly, if there are more posts with the happiness (anti-depressive) category are posted, we can assume that the people are happy in that area.

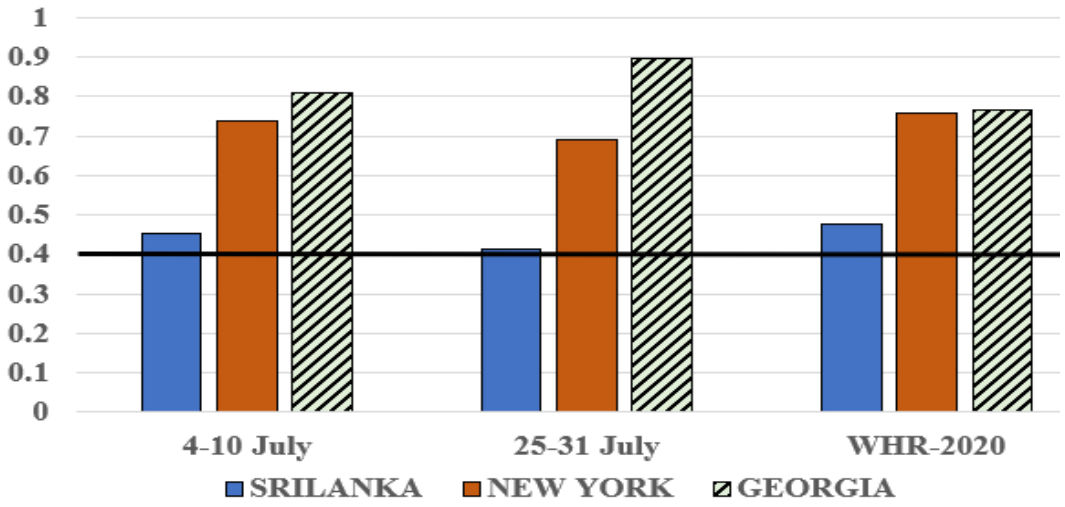

Figure 8.1 Happiness Index of Sri Lanka, New York, and Georgia along with the average lines

Happiness Index (HI) is defined by the ratio of anti-depressive tweets to the sum of the depressive tweets of the same duration. $HI = [\sum_{n=1}^{p} \bar{d}_n / \sum_{n=1}^{q} (d_n + \bar{d}_n)]$ for $0 \leq HI \leq 1$, where p is the total number of depressive keywords, q is the total number of anti-depressive keywords. d_n and $\bar{d}_n$ represent the depressive and anti-depressive keywords, respectively. In our case, p and q are equal to seven.

Fig. 8.1 shows the Happiness Index of Sri Lanka, New York, Georgia, and the World Happiness Report data [191]. The happiness index computed from the 366 days tweets data (01-April-2020 to 01-April-2021) is the horizontal black line. Figures 8.2 and 8.3 show the depressive and anti-depressive tweets data and seven days average values of the same period, and 366 days data to exhibit the change in each tweet keyword.

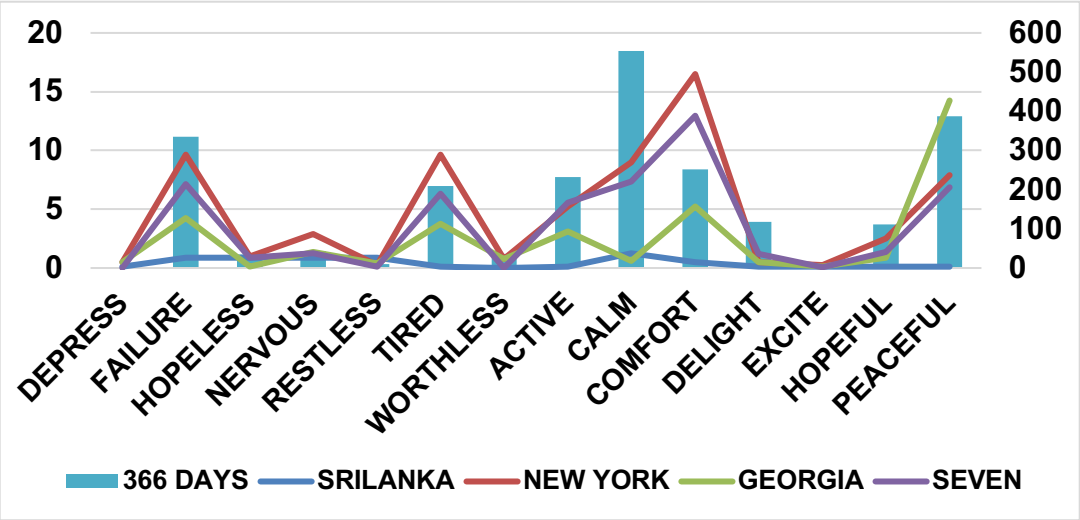

Figure 8.2 Happiness Index of Sri Lanka, New York, and Georgia along with the average lines

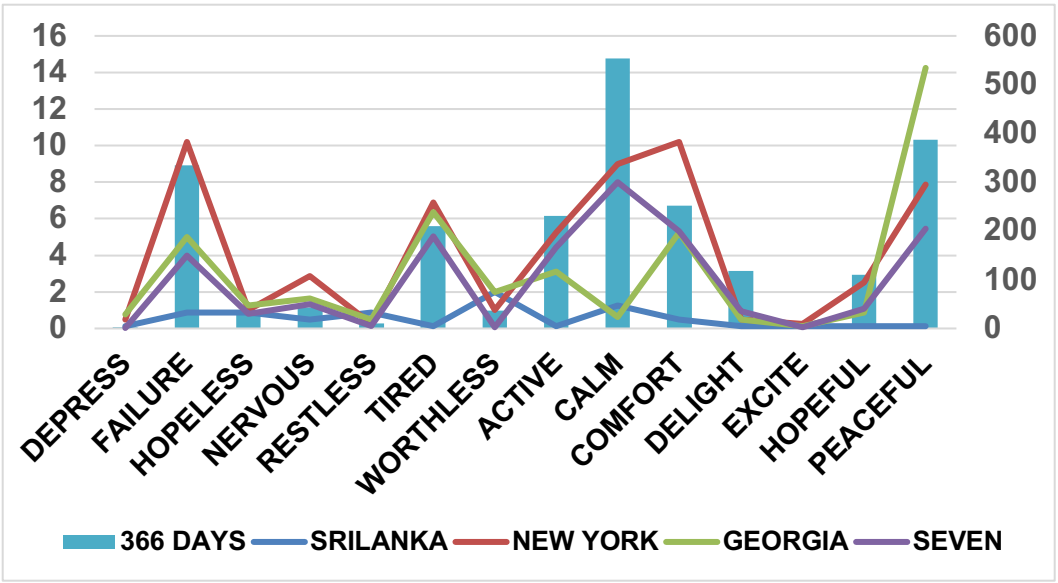

Figure 8.3 Happiness Index of Sri Lanka, New York, and Georgia along with the average lines

Our computations are based on 'outcome' with a hypothesis that happier people will tweet with happy keywords and unhappy people will tweet with depressive keywords. World Happiness Report (WHR-2020) is a significant scale of happiness ranking of 156 countries [191]. Gallup World Poll happiness scores are used in ranking happiness. Table 8.1 shows the Happiness Index

computed by our method in the two weeks and the Gallup score used in the WHR-2020 report [192]. We observed a decrease in Gallup data of WHR-2020 in Georgia to New York and Sri Lanka. Our results in HI-1 and HI-2 show a similar decline (Table 1). Analyzing the 2.3 million tweets collected over 01-April-2020 to 01-April-2021 (366 days) resulted in an average value of impact on the Happiness Index of each tweet keyword.

Table 8.1 Happiness values and Gallup data

	Georgia	New York	Sri Lanka
HI-1	0.808	0.737	0.454
HI-2	0.896	0.689	0.414
Gallup Score WHR-2020	7031	6964	4381

Researchers [191] classified the HI in five levels (a) unhappy (<20), (b) Less happy ($20 \leq HI < 40$), (c) Quite happy ($40 \leq HI < 60$), (d) Happy ($60 \leq HI < 80$) and (e) Very happy ($80 \leq HI < 100$) and our data resulted in the Happiness Index computed from the 366 days tweets data is 0.4039. The happiness Index for the week 4-10 July is 0.44, and for the week 25-31 July, it is 0.45. These values coincide with the scales classified by previous researchers [193].

It is observed that during these two weeks of observations (HI-1 and HI-2), Sri Lanka has less than the yearly average happiness reported by WHR-2020. New York showed balanced happiness during 4-10 July, but more happiness during 25-31 July. Georgia recorded a better happy situation in both weeks of observation while compared with the WHR-2020 report. Our two-week data HI-1 and HI-2 delivered the keywords that impacted the Happiness Index. The results from our observations indicate that Happiness Index in New York state has more impact from failure, hopeless, nervous, tired and comfort in both weeks. Georgia State's Happiness Index has more influence from nervous and peaceful in the HI-1 week and Hopeless, nervous, tired, and peaceful

in the HI-2 week. Sri Lanka Happiness Index did not show any specific impact (Figures 8.2 and 8.3).

During these two weeks of observation, People in Georgia are happier in the HI-2 observation period. People in New York are happier in both weeks. However, they are happier in the HI-1. Contrary to these observations, the Sri Lanka residents are more unhappy than average in both weeks of observation. Nervous, Hopeless, and Tired are the depressive states, and comfort and peaceful are the anti-depressive states that contributed substantially to HI computations [194].. Our database consists of tweet data date wise and keyword wise. In addition, we are collecting tweets with geographical locations. An interactive framework was developed that takes geographical location data and duration as input, collects the tweets from our database. Our system will collect the tweets related to the respective geographical area, and desired duration of study. It will process to compute the happiness index of that duration and delivers.

8.4 Mental Health Happiness and Feel-Good-Factors with Bad words

8.4.1 Methods

We identified seven keywords from Kessler's works [92] and obtained keyword related tweets from twitter.com using API on daily basis for the hashtags #active, #calm, #comfort, #delight, #excite, #failure, #hopeful, #hopeless, #nervous, #peaceful, #restless, #tired, #worthless, #depress and #corona. These keywords include the antonym of the seven Kessler words and 'Corona.' We have been collecting tweets with these keywords for the past few years. 2.3 million tweets data and tweets, in particular, belonging to May and June 2020 were used in this study. Special characters, hypertext links, numbers, non-English characters, punctuation marks, and stop words were removed. We selected 450 bad words that are used in our data set [195], and the

frequency was computed in each of the keyword tweets each day. Bad-word frequency ratios are calculated using the formulae:

$BW_i = b_i / \sum_{n=1}^{14}(b_i)$; $TW_i = b_i / \sum_{n=1}^{14}(t_i)$,

where b_i is the frequency of bad words in the keyword and t_i is the total number of words in each keyword tweet. This ratio gives the impact of the keyword within the tweet set concerning bad words and total words. We categorized #failure, #hopeless, #nervous, #restless, #tired, #worthless, #depress as one set that represents depressive tweet keyword set and #active, #calm, #comfort, #delight, #excite, #hopeful, #peaceful as another set represent anti-depressive keyword set. For each day (j), the depressive keyword weights are computed using the formulae. DW_j = BW_i/BW_k (i=1 to 7 depressive set; k = 1 to 7 anti-depressive set) and $EW_j = TW_i/TW_k$ (i=1 to 7 depressive set; k = 1 to 7 anti-depressive set).

The ratio of the number of bad words in the keyword tweet to the total number of bad words in all the depressive keywords is found, and a similar ratio is computed for anti-depressive keywords. DW shows these two values' ratios, representing the impact of bad words in depressive keywords and anti-depressive keywords (DW). Similarly, we calculated the impact for the bad words in respect to total words in the depressive and anti-depressive keywords (EW) and shown in Figures 8.4 and 8.5.

The Happiness Index is computed as the ratio of the above and by rationalizing.

$HI_j = DW_j / EW_j$ for each day j.

We computed the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values from the contributions of these depressive and anti-depressive keywords. The confusing matrix and depression index are shown in Table 8.2

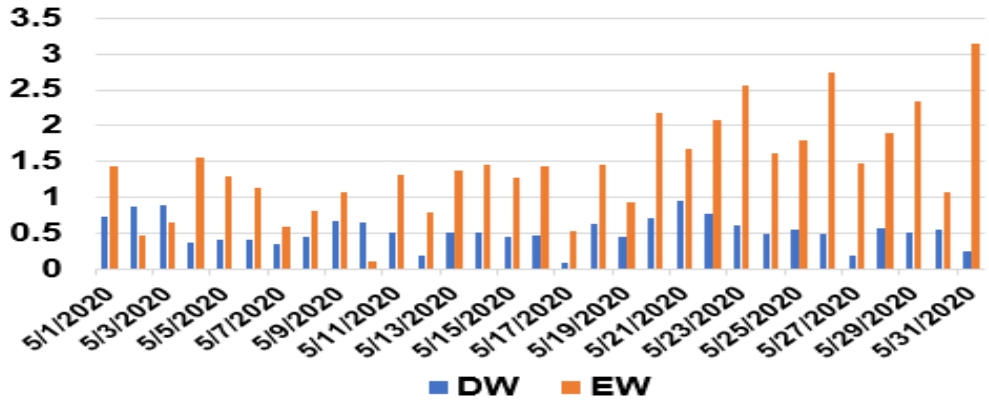

Figure 8.4 The ratios of depressive and anti-depressive tweet sets (May 2020)

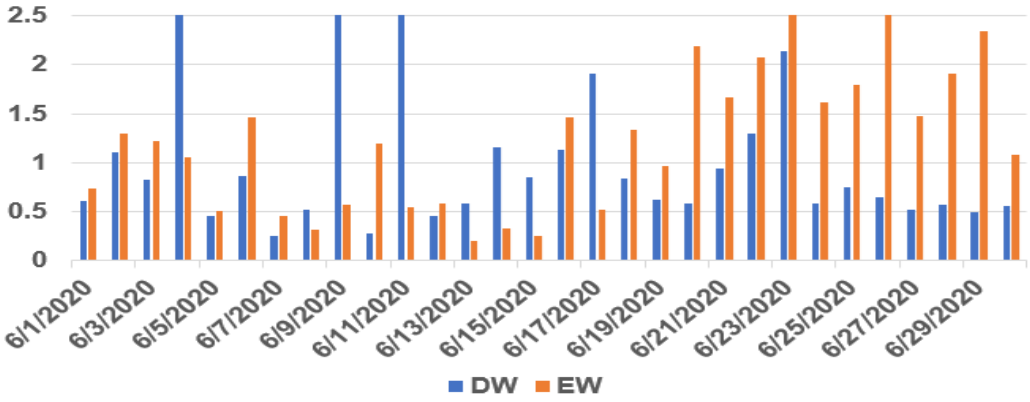

Figure 8.5 The ratios of depressive and anti-depressive tweet sets (June 2020)

F-1Score is computed using the formula:

F-1 = 2TP / (2TP + FP + FN).

Mathews Correlation Co-efficient (MCC) is computed as

*(MCC) = (TP*TN - FP*FN) /Sqrt((TP+FP) *(TP+FN) *(TN+FP) * (TN+FN)).*

Depression Index [DI] is computed using the formula:

DI = F1/(1-MCC) [31].

World Health Organization (WHO) publishes the number of COVID confirmed cases online, and we have collected the information for May and June 2020 [196]. The Happiness Index,

Depression Index [45], and the conformed covid instances are shown in Figures 8.6 and 8.7 for May 2020 and June 2020. The number of depressive and anti-depressive keywords contributed to compute the Happiness Index are shown in Figures 8.8 and 8.9 for May and June 2020.

8.4.2 Feel-Good-Factors

We computed the bad-words contribution to the Happiness Index. The average contribution of each depressive and anti-depressive keyword is processed for the months of May and June 2020. The keyword contribution is counted as '1' if it is more than the average and '0' otherwise.

Table 8.2 Confusion Matrix and Depression Index (May and June 2020)

May 2020					
DATE	***TP***	***FN***	***FP***	***TN***	***DI***
5/1/2020	11753	1703	2458	16287	1.88
5/2/2020	2984	18740	8595	3421	0.09
5/3/2020	2170	16762	10813	2466	0.06
5/4/2020	7064	1575	2462	18932	1.56
5/5/2020	7518	1733	2256	18921	1.63
5/6/2020	7826	2065	2337	19240	1.58
5/7/2020	8818	4982	2920	25912	1.22
5/8/2020	8988	2973	2429	20072	1.52
5/9/2020	8998	3957	4222	13301	1.08
5/10/2020	7968	19363	2030	12467	0.53
5/11/2020	10848	2115	2764	21803	1.74
5/12/2020	5503	3090	2431	30277	1.17
5/13/2020	4312	1831	2497	8590	1.05
5/14/2020	8748	1573	2276	17162	1.75
5/15/2020	3542	1751	2221	7851	0.98
5/16/2020	1826	1564	2241	3921	0.57
5/17/2020	2864	5594	2896	36897	0.53
5/18/2020	13863	2045	2963	21757	1.89
5/19/2020	8561	66341	62020	18922	0.06
5/20/2020	12609	3038	6642	17659	1.20
5/21/2020	8083	2071	3450	8499	1.24
5/22/2020	7876	1814	3758	10354	1.26
5/23/2020	11379	1883	4818	18644	1.45

5/24/2020	9016	1977	3207	18175	1.52
5/25/2020	9980	2195	3933	18022	1.44
5/26/2020	8650	2054	5639	17958	1.15
5/27/2020	5423	42313	62338	28118	0.06
5/28/2020	12220	1673	3190	21602	1.81
5/29/2020	10323	1824	4286	20689	1.49
5/30/2020	8997	1620	1751	16329	1.89
5/31/2020	5002	1461	4599	19854	0.99
June 2020					
DATE	***TP***	***FN***	***FP***	***TN***	***DI***
6/1/2020	11753	1703	2458	16287	1.88
6/2/2020	2984	18740	8595	3421	0.09
6/3/2020	2170	16762	10813	2466	0.06
6/4/2020	7064	1575	2462	18932	1.56
6/5/2020	7518	1733	2256	18921	1.63
6/6/2020	7826	2065	2337	19240	1.58
6/7/2020	8818	4982	2920	25912	1.22
6/8/2020	8988	2973	2429	20072	1.52
6/9/2020	8998	3957	4222	13301	1.08
6/10/2020	7968	19363	2030	12467	0.53
6/11/2020	10848	2115	2764	21803	1.74
6/12/2020	5503	3090	2431	30277	1.17
6/13/2020	4312	1831	2497	8590	1.05
6/14/2020	8748	1573	2276	17162	1.75
6/15/2020	3542	1751	2221	7851	0.98
6/16/2020	1826	1564	2241	3921	0.57
6/17/2020	2864	5594	2896	36897	0.53
6/18/2020	13863	2045	2963	21757	1.89
6/19/2020	8561	66341	62020	18922	0.06
6/20/2020	12609	3038	6642	17659	1.20
6/21/2020	8083	2071	3450	8499	1.24
6/22/2020	7876	1814	3758	10354	1.26
6/23/2020	11379	1883	4818	18644	1.45
6/24/2020	9016	1977	3207	18175	1.52
6/25/2020	9980	2195	3933	18022	1.44
6/26/2020	8650	2054	5639	17958	1.15
6/27/2020	5423	42313	62338	28118	0.06
6/28/2020	12220	1673	3190	21602	1.81
6/29/2020	10323	1824	4286	20689	1.49
6/30/2020	8997	1620	1751	16329	1.89

8.4.3 Happiness and Depressive Index

The moving average method was applied to find the mean square error in Happiness Index and shown in Table 8.3. Depression Index (Figures 8.10 and 8.11) and Happiness Index (Figures 8.12 and 8.13) show the actual and forecasted DI and HI for May and June 2020.

ARIMA model forecasting is carried out for the Depressive Index and Happiness Index, and the ACF, PACF plots are shown in Figures 8.14 to 8.17. Table 8.4 shows the number of days that each keyword is contributed beyond it norm.

Depressive and Anti-depressive tweet keyword data sets are used to compute the accuracies using the Latent Dirichlet Allocation method [197], [198] for both May and June 2020 data sets, and the results are shown in Table 8.5.

Table 8.3 Mean Square Error

	May 2020		**June 2020**	
	Happiness Index	***Depressive Index***	***Happiness Index***	***Depressive Index***
MSE	0.502	0.418	0.716	0.614

Table 8.4 Keyword factors and number of days of contribution

Depressive			**Anti-depressive**		
Keyword	***Contribution***		Keyword	***Contribution***	
	May 2020	***June 2020***		***May 2020***	***June 2020***
Failure	19	15	Active	20	15
Nervous	6	5	Calm	12	10
Restless	11	12	Comfort	14	10
Tired	18	13	Delight	5	4
Worthless	11	5	Excite	3	4
Depress	4	13	Hopeful	14	8
Hopeless	9	11	Peaceful	17	18

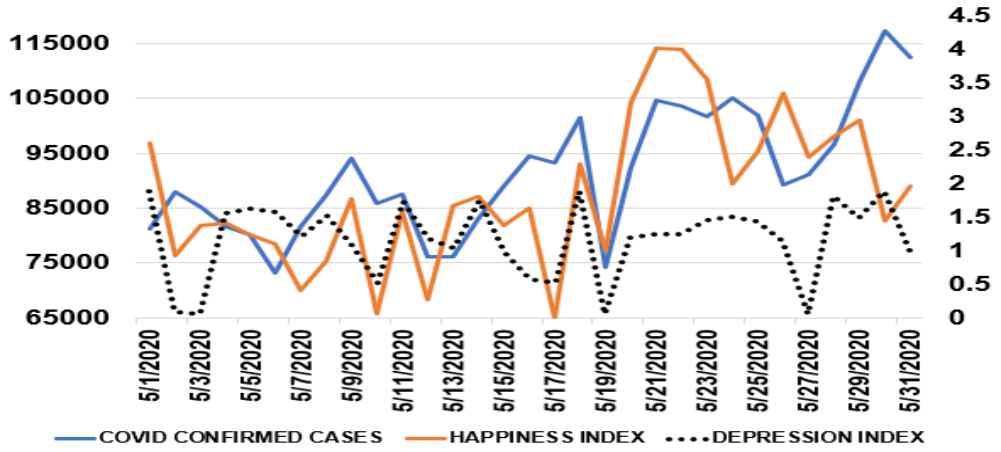

Figure 8.6 Happiness Index, Depression Index, and Covid confirmed cases [30] (May 2020)

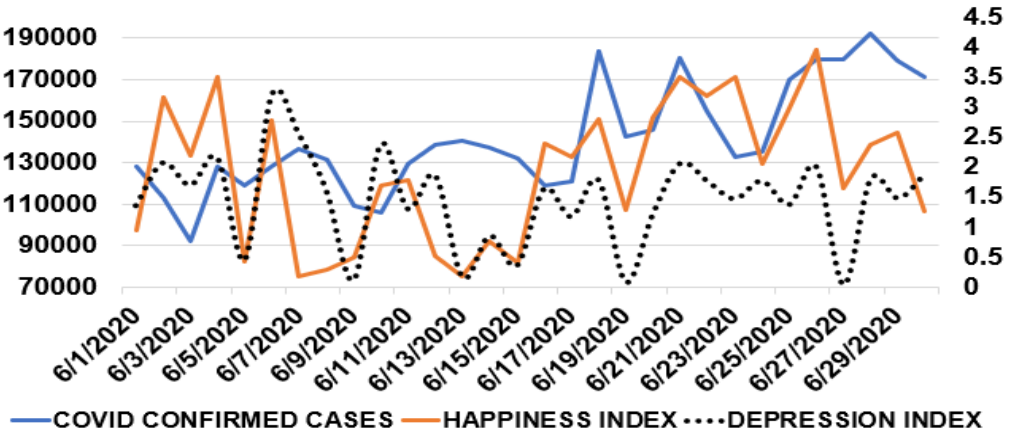

Figure 8.7 Happiness Index, Depression Index, and Covid confirmed cases [30] (June 2020)

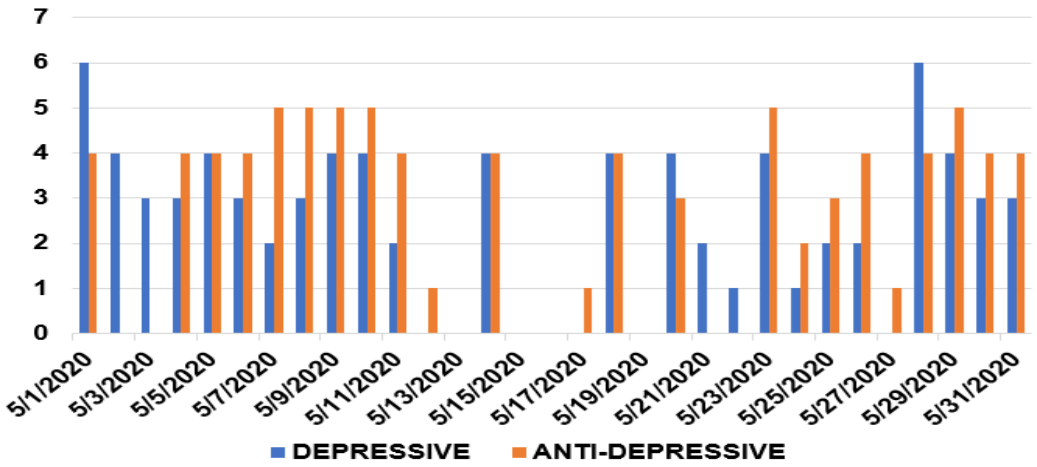

Figure 8.8 No of Depression and anti-depressive keywords contributed to HI (May 2020)

Table 8.5 Latent Dirichlet Allocation Results

Accuracy	**May 2020**		**June 2020**	
	HI	***DI***	***HI***	***DI***
Depressive set	89	78	88	75
Anti-Depressive	83	76	85	81

Table 8.6 Feel-Good-Factors contribution in Rank order

Rank	**May 2020**		**June 2020**	
	Dep	***Anti-Dep***	***Dep***	***Anti-Dep***
1	Failure	Active	Failure	Peaceful
2	Tired	Peaceful	Tired	Active
3	Restless	Comfort	Depress	Comfort
4	Worthless	Hopeful	Restless	Calm
5	Hopeless	Calm	Hopeless	Hopeful
6	Nervous	Delight	Nervous	Delight
7	Depress	Excite	Worthless	Excite

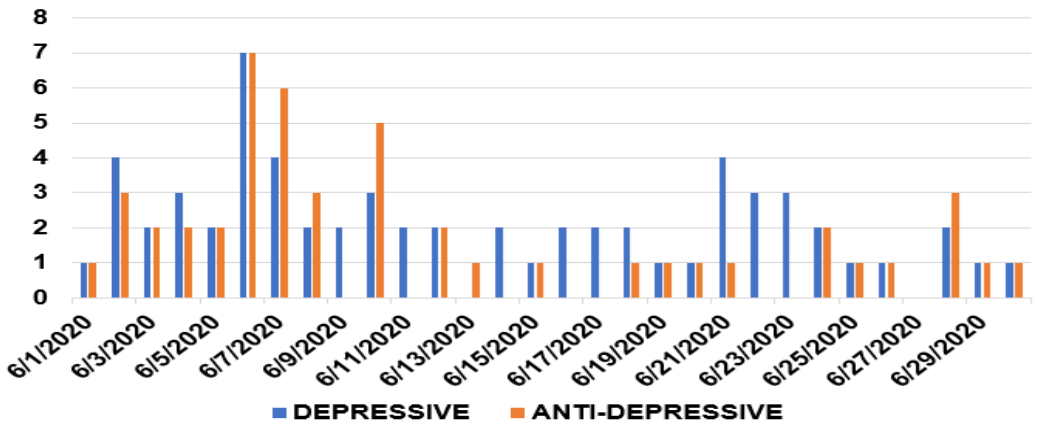

Figure 8.9 No of Depression and anti-depressive keywords contributed to HI (June 2020)

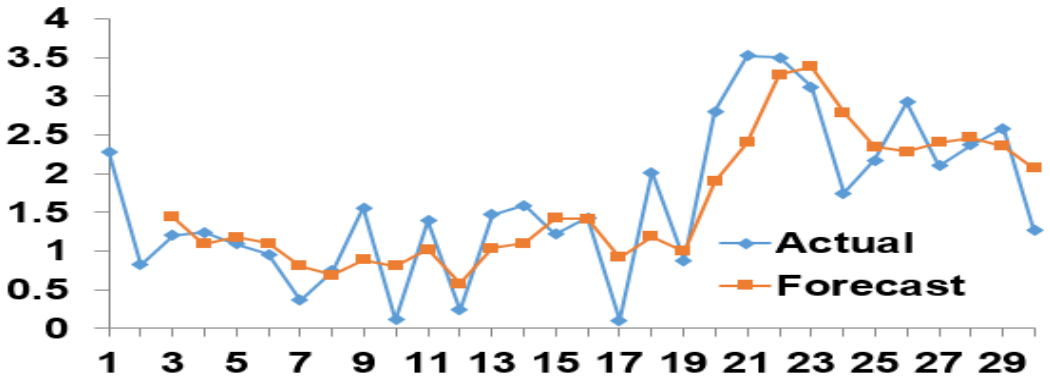

Figure 8.10 Happiness Index May 2020

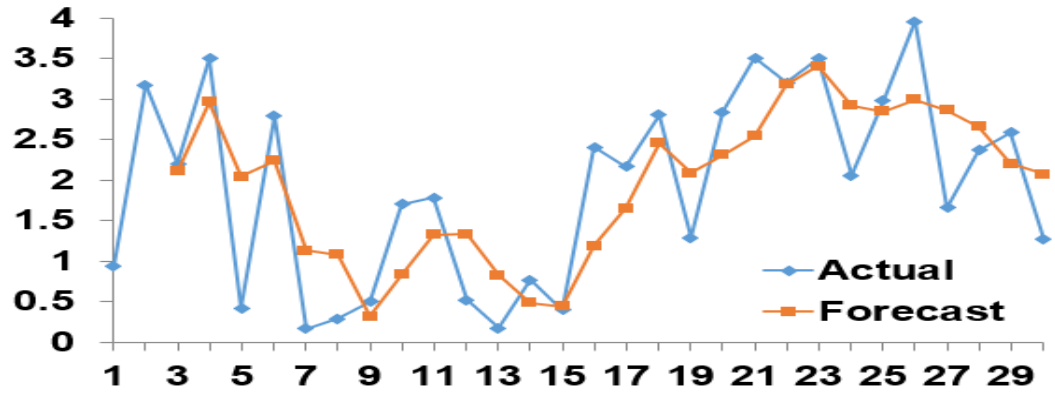

Figure 8.11 Happiness Index June 2020

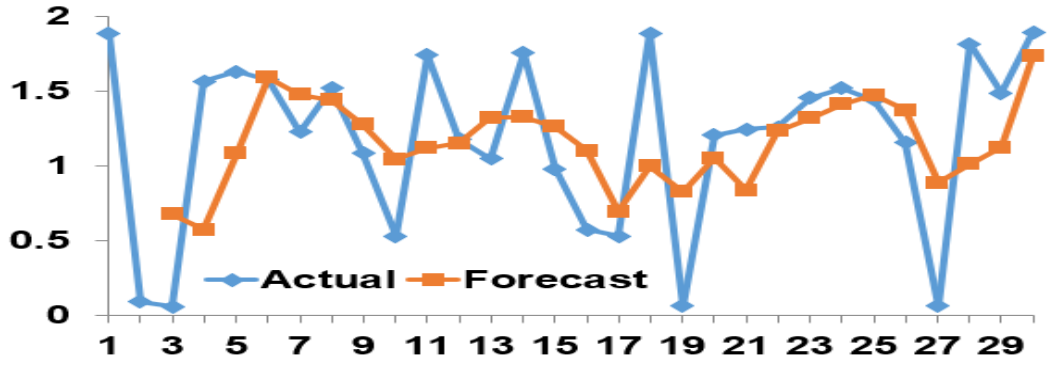

Figure 8.12 Depress Index May 2020

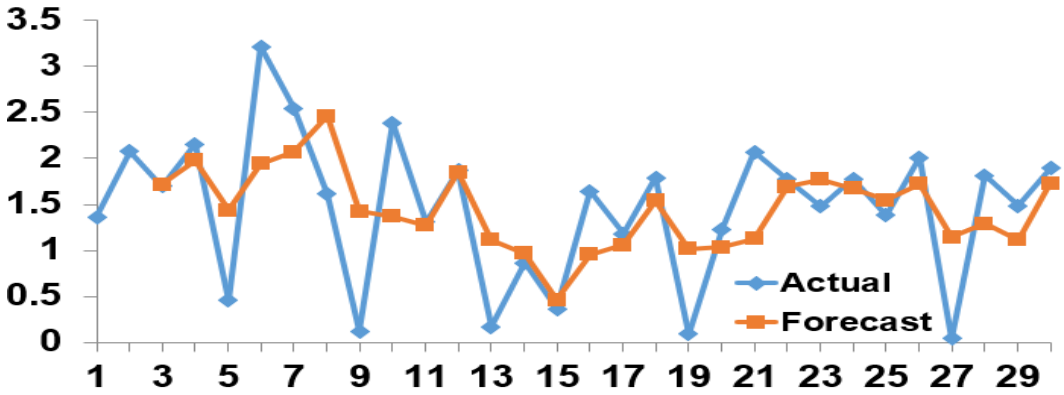

Figure 8.13 Depress Index June 2020

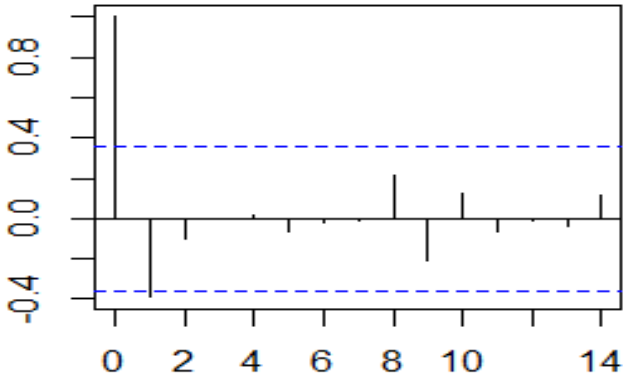

Figure 8.14 ACF plot in Happiness Index

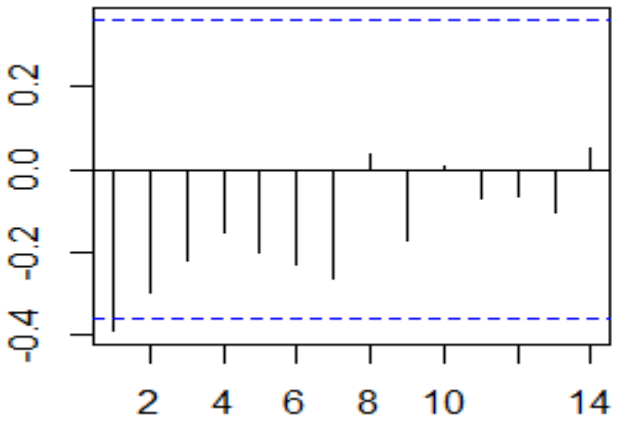

Figure 8.15 PACF plot in Happiness Index

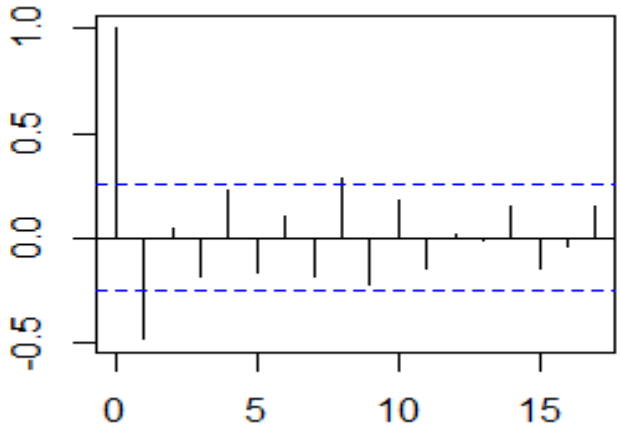

Figure 8.16 ACF plot in Depression Index

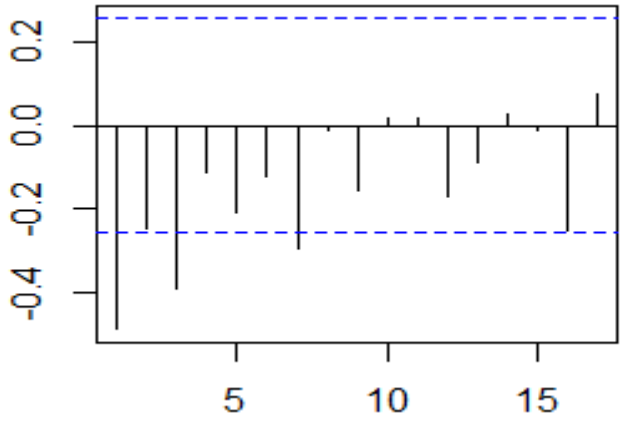

Figure 8.17 PACF plot in Depression Index

8.5 Happiness Index Map

WalletHub [199] studied three key dimensions (a) emotional and physical well-being, (b) Work environment, and (c) Community and the environment with 31 relevant metrics. The happiness index scaled to 0-1, with 1 being the happiest state, and 0 has the minimum happiness index. The value in blue indicates for the year 2021 (Fig. 8.18).

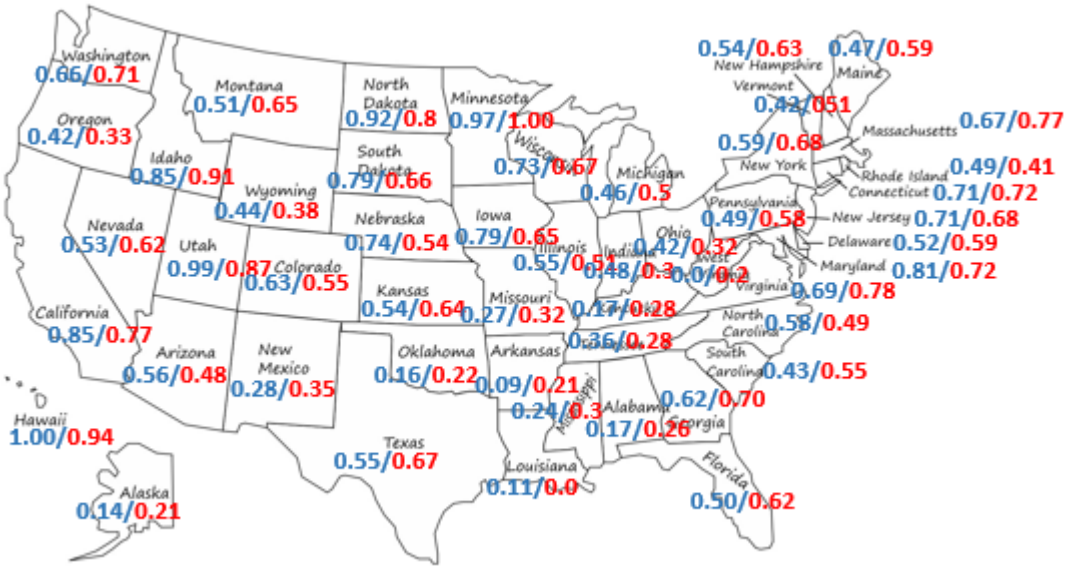

Figure 8.18 Happiness Index (2021) WalletHub data (blue), our method (red)

Depressive and anti-depressive tweets from each state are analyzed, and a happiness index map is prepared for one day (5 September 2021) using our method mentioned at 8.3.1. These values are shown in red in Fig 8.18. The tweets will be added to the database daily with the depressive and anti-depressive tweets data belonging to all the fifty states of the USA. This amounts to seven hundred files representing the seven depressive and seven anti-depressive keyword data of fifty states each day. We have developed a framework that collects this data and saves it. Our system takes input dates from the user interactively and generates an interactive map to show the happiness index from the WalletHub and our database for the desired dates.

8.6 Discussion

Ratios of depressive keyword and anti-depressive keyword to the depressive keyword set and anti-depressive keyword set and the total words in the tweet keyword for each day are presented in Fig. 8.4. Happiness Index, Depression Index, and Covid confirmed cases are shown in Figures 8.6 and 8.7. There exists a greater possibility that a covid confirmed individuals post depressive/anti-depressive tweets during this anxiety. The Happiness Index, Depression Index, and Covid confirmed cases reported by World Health Organization revealed a similar pattern (Figures 8.6 and 8.7).

Feel-Good-Factors offer an understanding of the depressive and anti-depressive keywords. In the set of depressive keywords, ‘failure’ showed the highest contribution followed by ‘tired,’ ‘restless,’ ‘worthless,’ ‘hopeless,’ ‘nervous,’ and ‘depress.’ Similar computations are made with anti-depressive keywords and observed that the keyword ‘active’ has contributed the maximum, followed by ‘peaceful,’ ‘comfort,’ ‘hopeful,’ ‘calm,’ ‘delight,’ and ‘excite’ (Table 8.6).

It is observed that there is no substantial contribution in seven days by depressive tweets and twelve days by the anti-depressive keywords during May 2020 (Fig. 9.8). Similar calculations are made for June 2020 and observed that there are two days with less contribution with a good match in the order in their contribution. The impact of depressive and anti-depressive keywords on Feel-Good-Factors was computed using the Keyword Contributions' average influence and daily values. Mean Square Error (MSE) in Happiness Index and Depressive Index resulted in low error rates.

The Happiness Index quantifies and compares with other days of observation. It thus facilitates understanding the hedonometric conditions. Collecting the tweets from different geographical regions, the Happiness Index can be correlated. ‘Failure’ and ‘Tired’ affect 47% in

May and 37% in June 2020 the depressive set and ‘Active’, ‘Peaceful’ and ‘Comfort’ affect 60% in May and 62% in June 2020, anti-depressive status. Feel-Good-Factors provided an insight into the mental status of the people. During May 2020, depressive keywords contributed 54%, and anti-depressive keywords 61% for the Feel-Good-Factors. In June 2020, we observed 43% influence from the depressive set and 40% from anti-depressive keywords. The bad words set is utilized in computing the Depressive Index and Happiness Index. These indices are forecasted with the moving average method and notice lesser Mean Square Errors. Accuracies obtained using the Latent Dirichlet Allocation (LDA) method are above 75% in all these cases. ACF and PACF plots from ARIMA signify, our model considered is appropriate [200].

The happiness index evaluations provide a broader measure of well-being inequalities. We have prepared a happiness map of all the states of the USA. It showed a better similarity. The happiness index will be helpful as an input to the Governmental, Non-Governmental agencies, and stakeholders in development. It can reinforce the mental health policies and Return On Investment (ROI) to scale up the prioritized sectors with effective interventions.

9 FORECASTING METHODS

9.1 Introduction

Forecasting is one of the useful statistical concepts that help to predict the future with the earlier data. Such methods help find the results of elections, website traffic, movie ratings, pricing of goods, and many more applications. The main issue in such a forecast is the accuracy of the prediction. The accuracy depends on many factors. These forecasting methods are classified as Qualitative and 'Quantitative methods. The qualitative methods are further split as (a) Executive opinion, where a group of intellectuals will collectively develop a forecast, (b) market survey in which surveys decide the forecast, (c) Salesforce, where an individual will submit the data belonging to his area and collectively a general forecast is estimated and (d) Consensus agreement is reached among a group of experts. Quantitative methods can further decompose as 'time-series methods' and 'associative models.

Computational models to predict the rise of depression and forecasting the mental illness with Twitter data was reported by several researchers [201] [8] [9] [32]. Prediction of Time Series data is carried out to comprehend the forecast and be ready with the demand/requirement for an activity, event, or an occurrence [202]. In the qualitative methods, we derive opinions, emotions, and personal experiences, which are subjective and do not depend on mathematical computations. In contrast, in quantitative methods, forecasting depends on mathematical models with calculations. The quantitative methodologies are further classified into time-series models and associative models. Time-series models assume that the past will repeat, whereas the associative models depend on the relationship between the response variables. [203] [204]. Moving average methods [205] were used in many applications.

In the time series model forecast, Naïve, Simple mean, simple moving average, weighted moving average, exponential smoothing, trend projecting, and seasonal indexes are few models that many researchers will apply to their data for the forecast.

9.2 Moving Average Models

9.2.1 Simple moving average model

The next value(s) in a time series is based on the previous values' average fixed finite number m. Thus, for all $i > m$

$$\hat{y}_i = \frac{1}{m} \sum_{j=i-m}^{i-1} y_i = (y_{i-m} + \cdots + y_{i-1})/m$$

9.2.2 Weighed moving average model

In this model, we assign m weights $w_1, \ldots, w_m$, where $w_1 + \ldots + w_m = 1$, and define the forecasted values as follows

$$\hat{y}_i = w_m y_{i-m} + \cdots + w_1 y_{i-1}$$

In the simple moving average method, all the weights are equal to $1/m$.

Tweets from the Twitter website are the source of data for our study. Media has become a source to share opinions with known and unknown people at large. The cost, technology, and speed at which the posts/tweets are generated have taken a lead role compared to other media, sources, and communication methods. Though some sections of social media serve a particular section, most social media slowly diverges to cater to all sections' needs. Another hypothesis is that the data available in social media is collective wisdom and certainly will have a vision in predicting the real-world outcomes.

9.2.3 Our moving average model

Tweet data related to failure, hopeless, active, tired, restless, worthless, calm, comfort, delight, hopeful, and corona collected from 01-April-2020 to 01-April-2021 (366 days) was used for this forecast.

We made an updated moving average model described below

If the y is a series with $y_1, y_2, y_3,\ldots y_n$ for n consecutive observations, the prediction for one day is computed as

$y_i = a_i+b_i+c_i$, where

$a_i = (y_{i-4}+ y_{i-3}+ y_{i-2}+ y_{i-1})/4$, $bi = (y_{i-1} - y_{i-4})/4$ and $ci = (b_{i-4}+ b_{i-3}+ b_{i-2}+ b_{i-1})/4$

The prediction values for two days are computed with

$y_i = a_i+b_i+c_i$, where

$a_i = (y_{i-5}+ y_{i-4}+ y_{i-3}+ y_{i-2})/4$, $b_i = (y_{i-2} - y_{i-5})/4$ and $c_i = (b_{i-5}+ b_{i-4}+ b_{i-3}+ b_{i-2})/4$

Similarly for two days the prediction values are computed with

$y_i = a_i+b_i+c_i$, where

$a_i = (y_{i-6}+ y_{i-5}+ y_{i-4}+ y_{i-3})/4$, $b_i = (y_{i-3} - y_{i-6})/4$ and $c_i = (b_{i-6}+ b_{i-5}+ b_{i-4}+ b_{i-3})/4$

One day, two days, and five days forecast values are plotted and shown in Figures 9.1 to 9.3 for 'failure,' 9.4 to 9.6 for 'hopeless,' and Figures 9.7 to 9.9 for 'corona' keywords. Table 9.1 shows the error percentage in the predicted values using our method for the keywords failure, hopeless, active, tired, restless, worthless, calm, comfort, delight, hopeful, and corona.

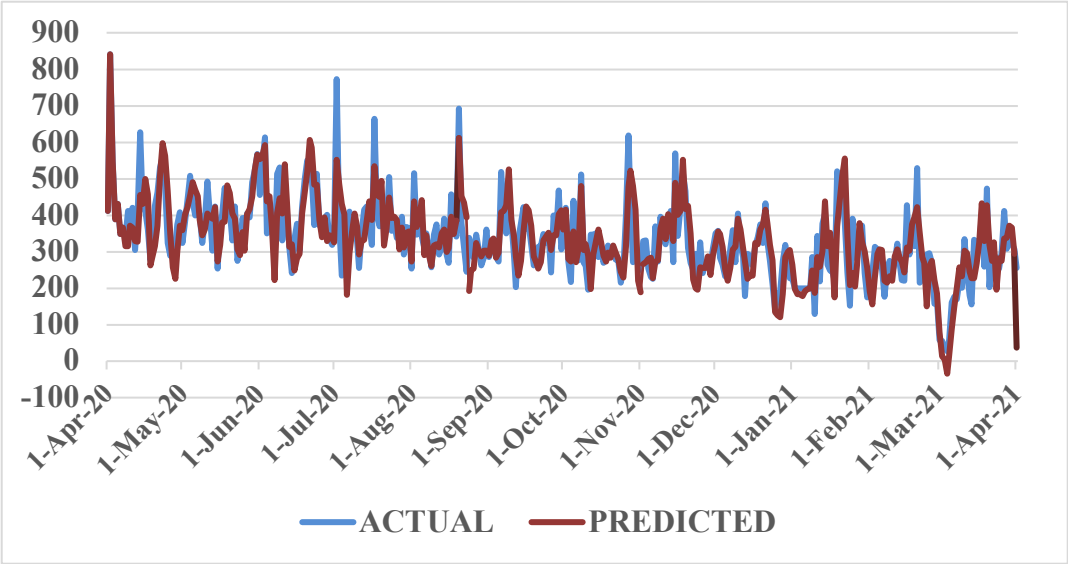

Figure 9.1 'Failure' tweet data from 01-April-2020 to 01-April-2021 (One-day prediction)

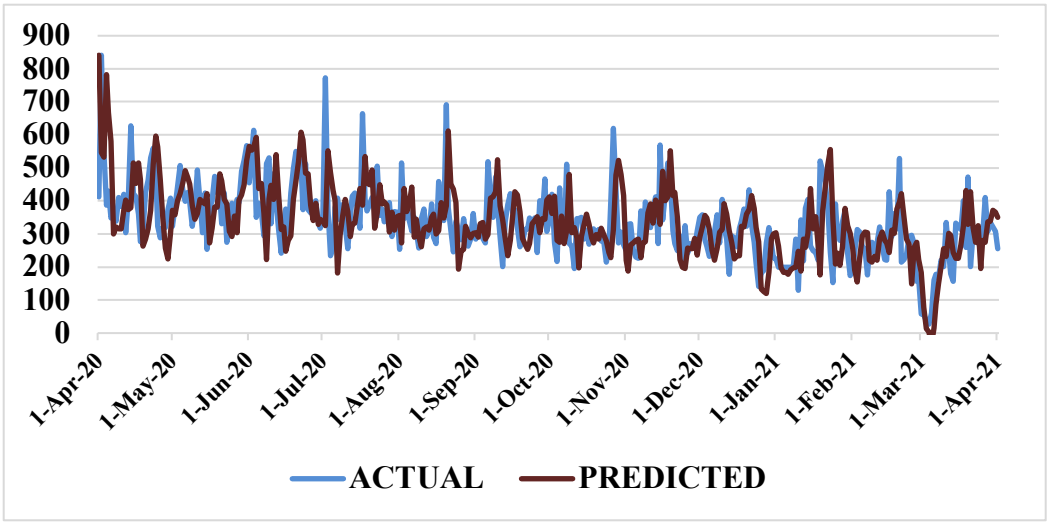

Figure 9.2 'Failure' tweet data from 01-April-2020 to 01-April-2021 (Two-day prediction)

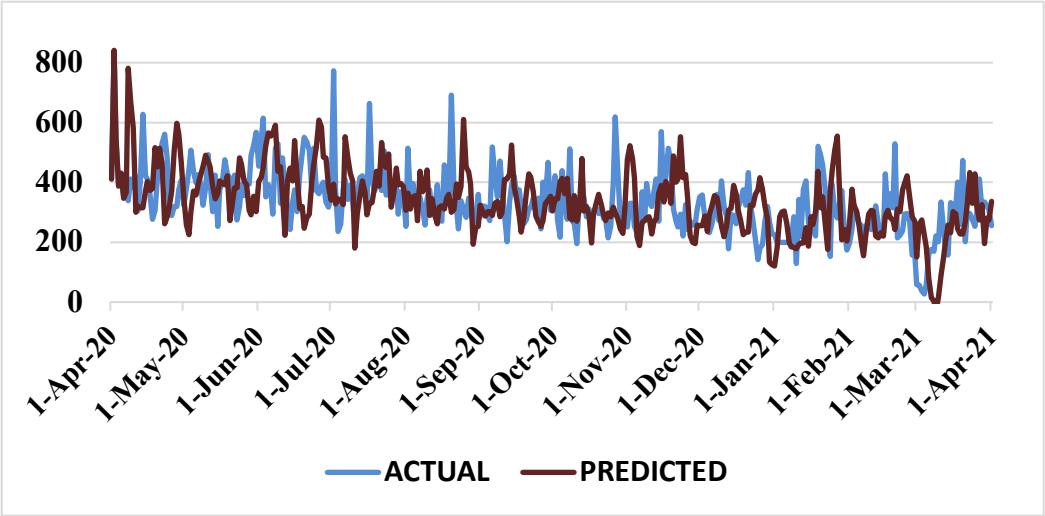

Figure 9.3 'Failure' tweet data from 01-April-2020 to 01-April-2021 (Five-day prediction)

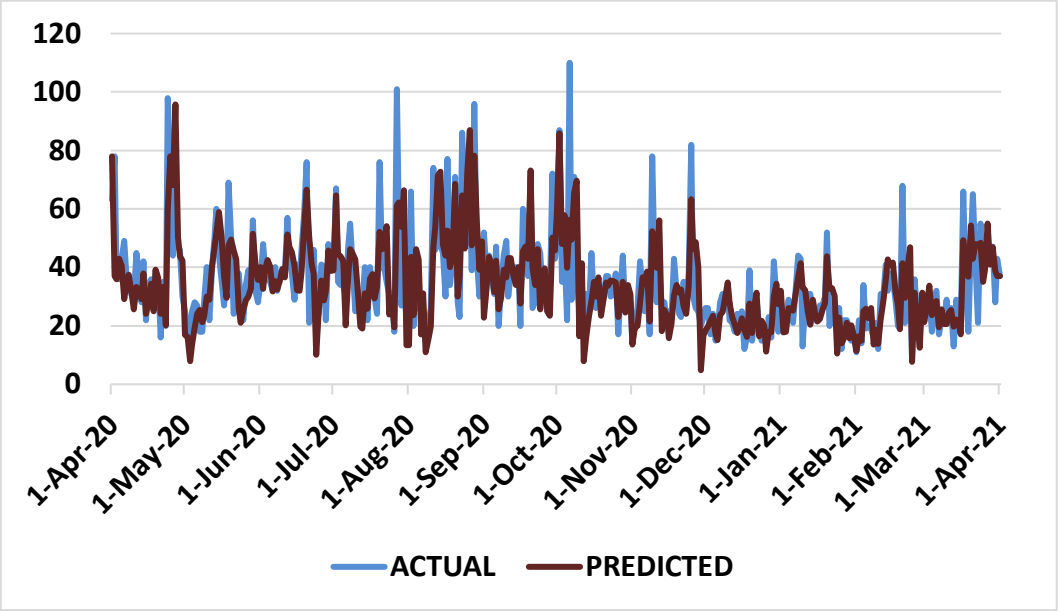

Figure 9.4 'Hopeless' tweet data from 01-April-2020 to 01-April-2021 (One-day prediction)

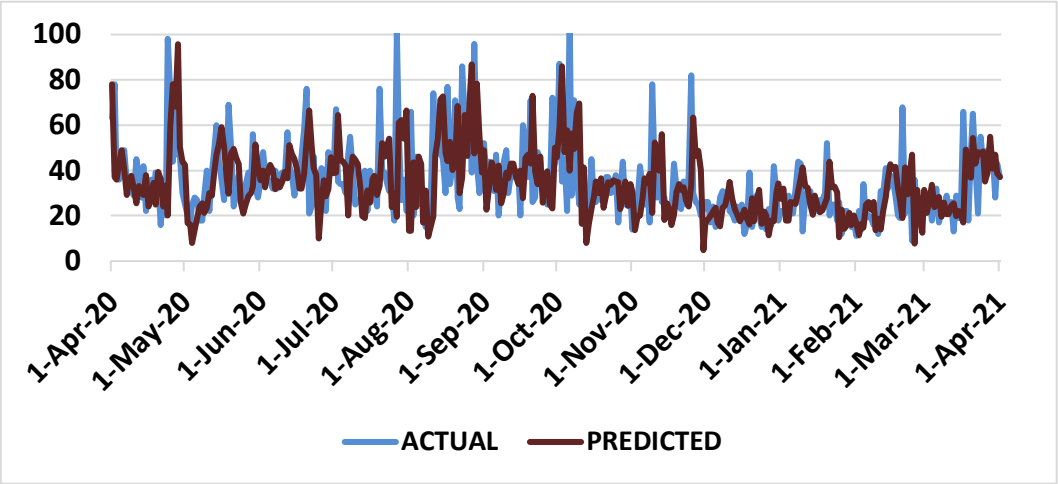

Figure 9.5 'Hopeless' tweet data from 01-April-2020 to 01-April-2021 (Two-day prediction)

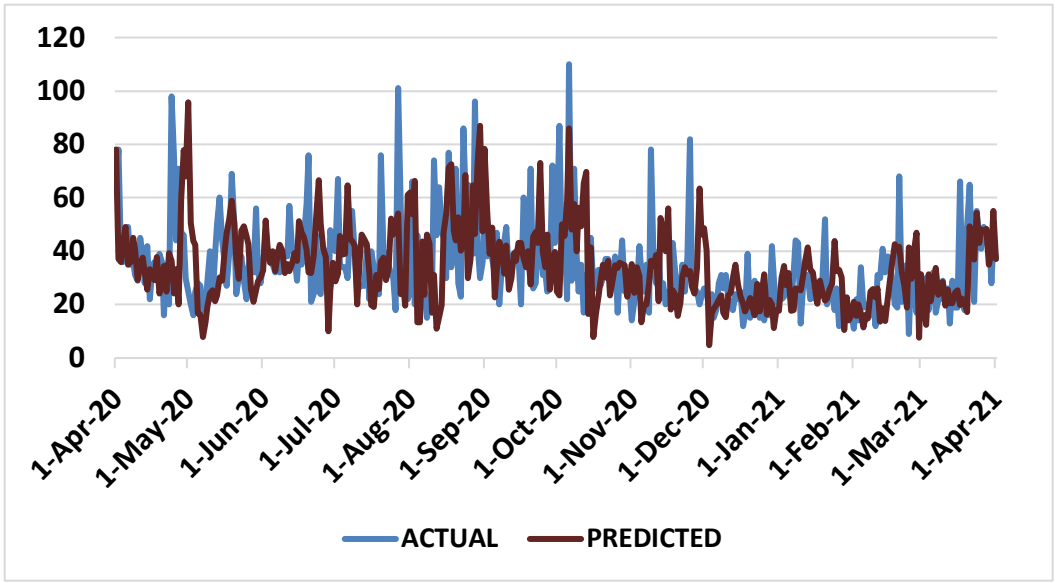

Figure 9.6 'Hopeless' tweet data from 01-April-2020 to 01-April-2021 (Five-day prediction)

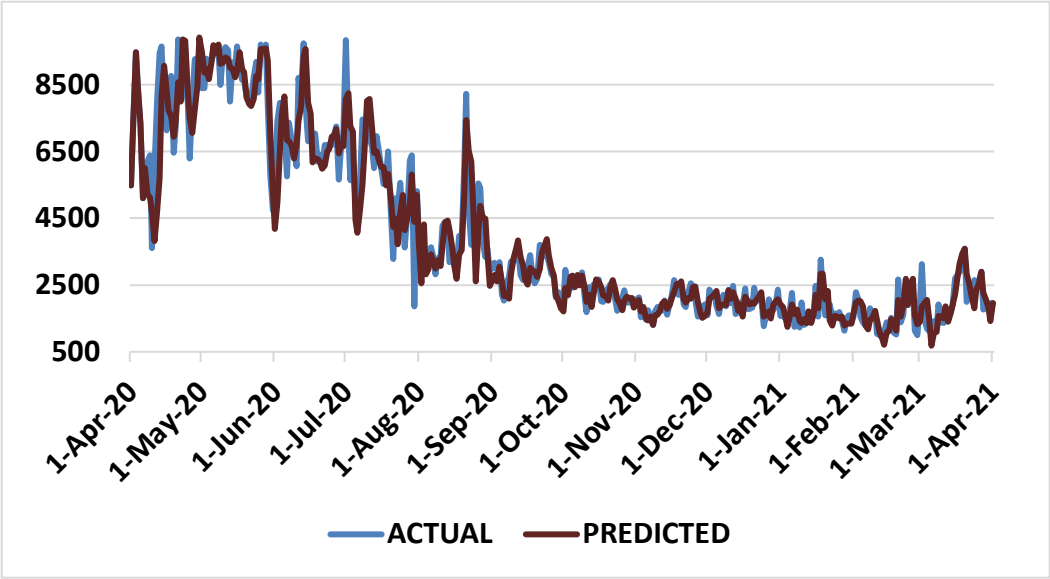

Figure 9.7 'Corona' tweet data from 01-April-2020 to 01-April-2021 (One-day prediction)

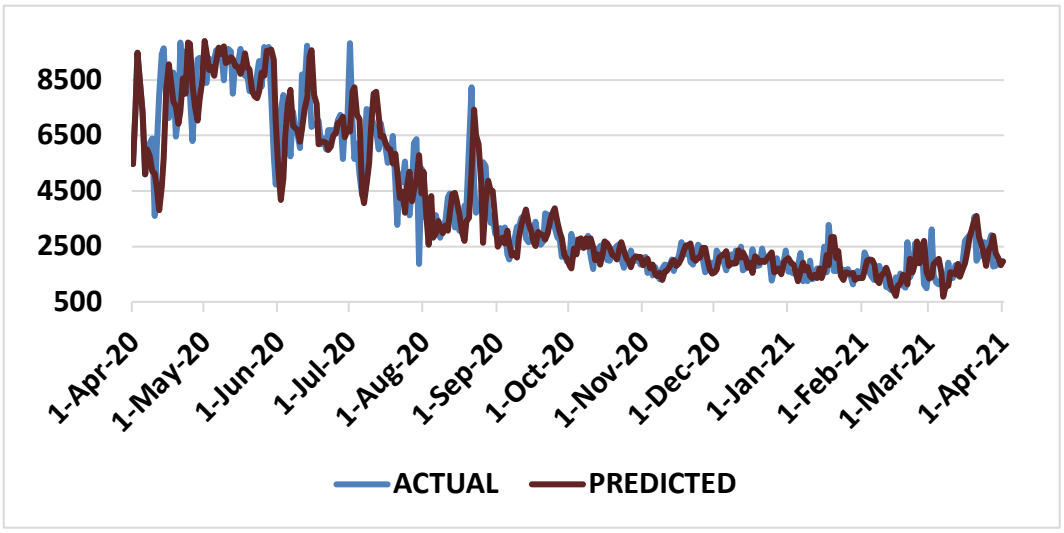

Figure 9.8 'Corona' tweet data from 01-April-2020 to 01-April-2021 (Two-day prediction)

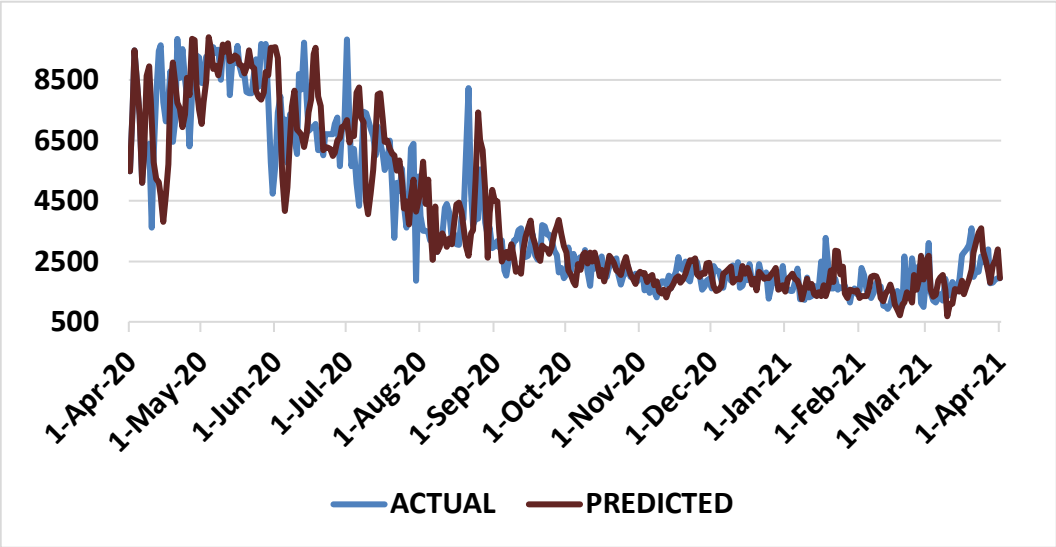

Figure 9.9 'Corona' tweet data from 01-April-2020 to 01-April-2021 (Five-day prediction)

Table 9.1 Percentage of error in predicting the depressive states in one, two and five days

	Keyword	Error percentage in predicting with our model			Simple Moving Average Model
		One day	Two days	Five days	One day
1	Failure	18	29	36	25
2	Hopeless	23	41	44	35
3	Active	22	34	35	27
4	Tired	15	25	26	21
5	Restless	50	86	96	75
6	Worthless	28	45	49	39
7	Calm	20	31	37	26
8	Comfort	23	35	38	29
9	Delight	37	58	73	50
10	Hopeful	25	40	49	33
11	Corona	12	20	23	21

9.3 ARIMA Model using COVID-19 epidemic dataset

9.3.1 ARIMA Model

ARIMA models provide another approach to time series forecasting. Exponential smoothing and ARIMA models are the two most widely used approaches to time series

forecasting. While exponential smoothing models are based on a description of the trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data [206].

9.3.2 Data and Methods

We collected 318,847 tweets related to depressive, anti-depressive, and corona hashtag keywords from Twitter during COVID-19. The tweets during COVID-19 of depressive and non-depressive hashtag keywords with their tweet count from 6th to 31 March 2020 are shown in Fig 9.11.

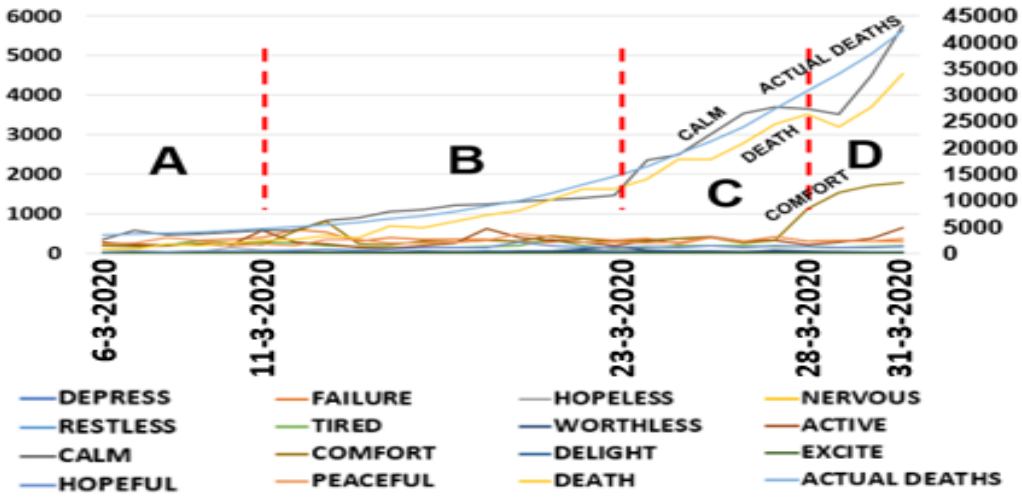

Figure 9.10 No. of Tweets Vs. the Hashtag keywords

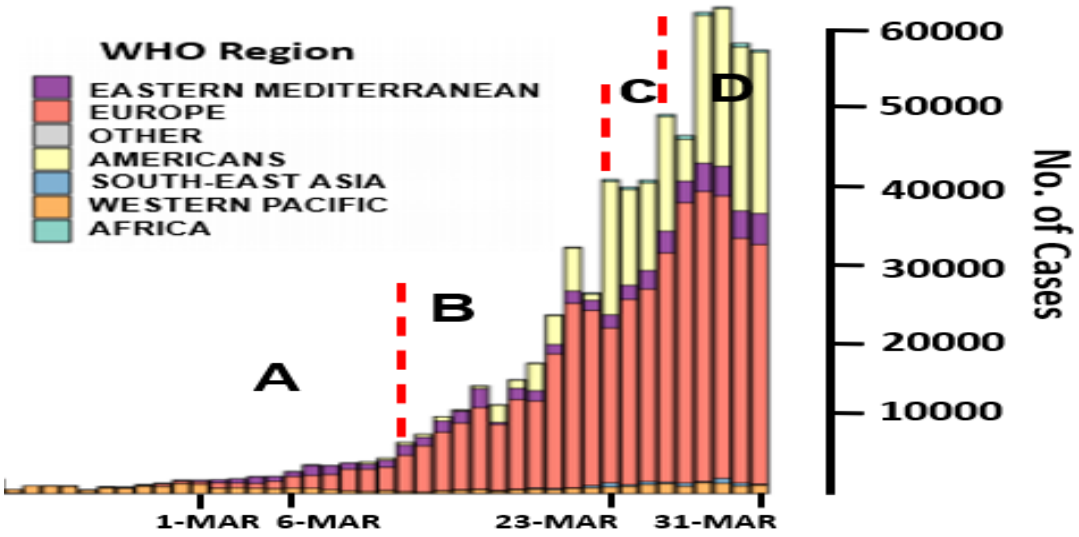

Figure 9.11 Epidemic curve of confirmed COVID-19 by date of the report and WHO region [207]

We grouped the COVID-19 tweet pattern into four regions A, B, C, and D from the gradient of the curve of the tweet graph of 'calm,' 'death,' and 'comfort,' as shown in Fig. 10.11. WHO is releasing situation reports during COVID-19 daily, and we have taken the data up to 31 March 2020 (Report-71) [208]. We marked A, B, C, and D regions on the epidemic curve of confirmed COVID-19 cases in Fig. 9.12

9.4 Discussion

9.4.1 Our Moving Average Model

We developed a new moving average model where we have taken the simple moving average value as a primary part. Two other small values are added to this to smooth the curve. The difference of the n-1 to n-4 value divided by four is an additional factor in our moving average model. The average of the last four such values is considered and explained in 9.2.3 above to minimize the error factor.

The results obtained from our model gave better results than simple moving average model. The table 9.1 depicts the error percentage with our model and simple moving average model. In all the keywords failure, hopeless, active, tired, restless, worthless, calm, comfort, delight, hopeful and corona, shown an improvement.

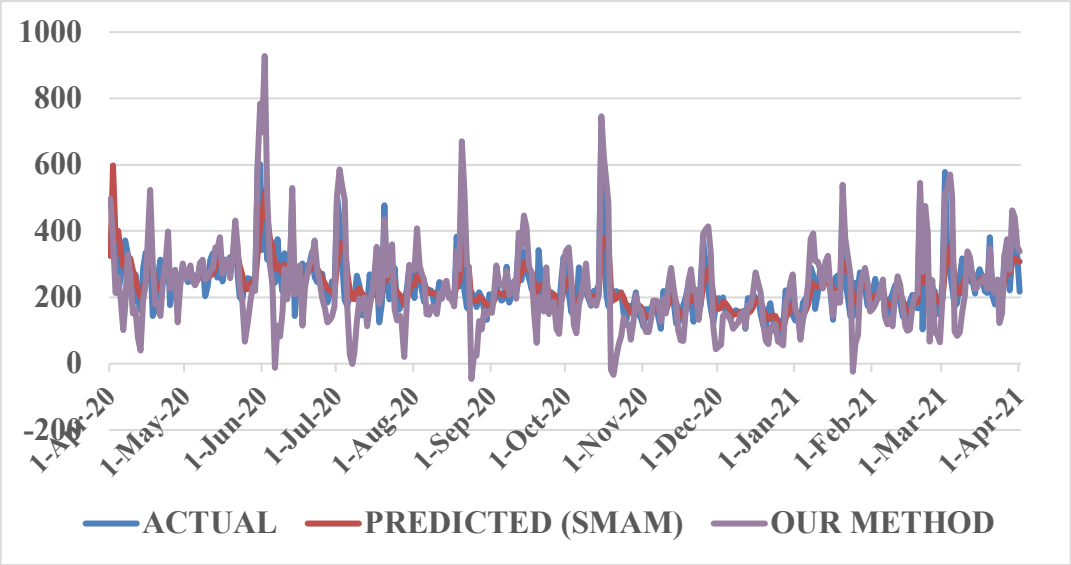

Figure 9.12 Actual, Predicted from SMAM and Our method

9.4.2 Arima Model

The growth rate in the number of tweets in depression and anti-depression is similar till 11 March in all the depressive and anti-depressive tweets. The tweet 'Calm' is shown a similar gradient with all other depressive and anti-depressive tweets, and this gradient is followed by the number of deaths after 11 March 2020. (Fig. 9.11). The 'Calm,' 'Death' graph showed a positive gradient increase on 11 and 23 March, whereas a negative gradient on 28th March 2020. WHO announced COVID-19 as a pandemic on 11 March, the deaths doubled in the USA with an increase of active COVID-19 cases on 23 March 2020. The first time there was a decrease in death count after COVID-19 effects, even though there was an increase in the number of active COVID-19 cases [63].

Table 9.2 Demarcation by our method, observations and WHO reports

Region	Our Observation	COVID-19 Epidemic Curve & reports
A-B 11-March-2020	Positive gradient noticed in 'Calm'	WHO announces COVID-19 outbreak a pandemic [209]
B-C 23-March-2020	Positive gradient noticed in 'Calm' and 'death'	First time change in deaths doubled. ~30% increase in the number of active Corona affected cases found [210]
C-D 28-March-2020	The graph showed a negative gradient in 'Calm' and 'death.'	Deaths decreased, while the number of active corona cases increased [64]

COVID-19 pandemic is a worldwide outbreak that affects many people. COVID-19 disaster crossed a significant phase around 11 March, as we observed that (a) the number of tweets is a similar ratio till that date. At around 11 March, there was an abnormality and (b) a high positive correlation with 'Death' tweets, as shown in Fig. 9.12. Our observations supported by WHO reports that COVID-19 developed into a pandemic [210] from 11 March 2020. We classified the COVID-19 shifted to other stages, as shown in Fig 10.11. WHO supports this change over to this stage, as shown in Fig. 9.12 and Table 9.2 Depressive and Anti-depression tweet data during COVID-19 were used to create the ARIMA model for the period of the next five days (1-Apr-2020 to 5-Apr-2020). The red lines in Fig. 9.13 show the predicted ranges. The blue line is the proposed best fit. The values of the failure and depress hashtag tweet count obtained from twitter.com and shown as dash lines within our proposed range. The number of tweets predicted using the ARIMA model for the next five days, and the real numbers are in the same range, as shown in Fig. 9.13.

We found a good association with the COVID-19 pandemic pattern using the tweets and with the WHO reports. The time series prediction system showed promising results for the next few days with our model. The proposed model will be beneficial for remedial non-clinical applications for helping the affected people.

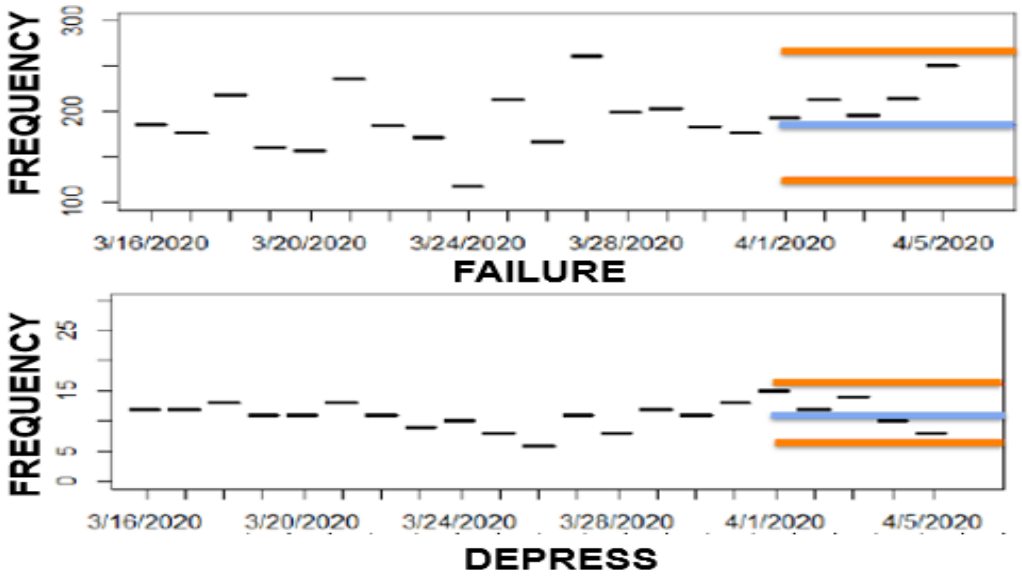

Figure 9.13 ARIMA forecast results for 'failure' (top) and 'depress' (bottom) hashtags

10 CONCLUSIONS

Social networks became an important part of human life today. People use one or more social network sites for communication and sharing information. The tweet data consisting of 2.3 million collected from 01-April-2019 to 01-April-2020 are the base data for our conclusions. We have analyzed the tweets with the Word frequency method, Singular Value Decomposition Method (SVD), Time series method, Time widow method, and Time stamp method and classified the tweet keys words.

The word frequency method resulted in the word usage pattern that an individual in a depressed stage may soon fall into another depressed stage. The singular Value Decomposition (SVD) method resulted in a compliment of 63% contribution to the first three pivotal values. The SVD method also supported our results, and we observe that there is always a chance of individual tweeting at least three of these keywords if he/she is mentally depressed. Everitt and Dunn [108] proposed an alternative approach based on comparing the component contribution of this diagonal element to almost 64%, which is in line with our results. We could identify the need for urgent attention, caution and monitor it with the Time series method. Such categorization of the tweet time will improve prediction for the time-series data, where an abnormality is observed. Our accuracy in the Time window method resulted in 20-40, 16-40, and 8-40 in Fine-tree algorithm, SVM methods, and KNN methods. We observed a similar trend by analyzing the depressive, anti-depressive, and COVID-19 to study the tweeting patterns, timings, and days of tweets by using the date and timestamp information. Classification of tweets followed that the depressive, anti-depressive tweets and corona confirmed cases (WHO data) show an increase from Wednesday. The results will be an indicator to identify mental illness people from social websites. Continuous monitoring of tweets of an individual who tweeted with one keyword will determine whether they

tweet with other depressive keywords to conclude that the individual is prone to Mental Health Illness soon. We also found that frequently used words depression and anti-depression tweets are posted at an interval of more than 10 minutes. The similarity is also seen in these tweet patterns within 10 minutes and beyond 10 minutes. The tweet pattern of 'tired' and 'restless' was different from other depressive keywords. The anti-depressive tweets followed the pattern of 'failure,' 'hopeless,' 'nervous,' and 'worthless' depressive tweets.

Our Studies also demonstrated a significant contribution through a new parameter that is computed using F1 and Matthew's Correlation Coefficient values. Our studies also suggest that this new parameter has an association with the tweeted data. We considered the positive for depressed tweets and the negative for the anti-depressed tweets in the confusion matrix and computed the new parameter. We concluded that the larger this value is the sign of higher depression on the tweets' day.

The clustering of depressed and anti-depressed keywords based on an event (Sri Lanka Bomb blast) using the text mining methods was applied to a set of data. The confusion matrix is computed from the Term-Document Matrix. We observed that the hashtags with the keywords Failure, Nervous, Comfort, Delight, Peaceful, Depress, Hopeless, Restless, Tired, Worthless, Active, Calm, Excite, and Hopeful are associated with the 'Bomb' hashtag keyword. Pearson Correlation coefficient was calculated within the set of keywords with the 'Bomb' keyword, and we grouped them into two clusters. Our assumption of grouping, Failure, Nervous, Comfort, Delight, and Peaceful into one category and Depress, Hopeless, Restless, Tired, Worthless, Active, Calm, Excite, and Hopeful into another group was supported with our results. The accuracy obtained from the clustering from the confusion matrix is 94% indicated that there is a good pattern match between the Similarity groupset and the Dissimilarity groupset. The above results support

the study of the clustering of tweet keywords into two groups (similarity and dissimilarity). Our work considered two sets of keywords belonging to depression and anti-depression and clustered into two balanced sets as similar and dissimilar in association with the 'event.'

We have collected the tweets during COVID-19, and using the tweets, we found a good association with the COVID-19 pandemic pattern and with the WHO reports. The time series prediction system showed good results for the next few days with our model. The proposed model will be beneficial for remedial non-clinical applications for helping the affected people.

Normalized tweet data of depressive, anti-depressive, and event tweets along with date wise used to compute the AUC with 'Total Period' under regular conditions, 'before the event,' 'during the event' and 'after the event.' It is noticed that the Tauktae cyclone created more destruction than Burevi Cyclone. These results correlate with the effect of events reported by other researchers. The Pearson Correlation validates that there is a decent correlation between depressive tweets and events. Collection of tweets online and computation of the impact will help governmental, non-governmental bodies plan and support the disaster preparedness and emergency team management in rescue operations during cyclones, forest fires of a particular geographical area.

Our study shows that the Tauktae cyclone impact more than the Burevi Cyclone, and the Burevi cyclone affect more than Srilanka bomb blasts incident in peoples' mental health. Similar results are noticed from these three events. The impacts '*before,' 'during'* and *'after'* the events are higher in the Tauktae cyclone than in the Burevi cyclone and Bomb Blasts incidents. The works done by other researchers also support our findings.

So far, no work has been reported earlier to scale Happiness Index using the mental health-oriented tweet data. Our computations are based on 'outcome' with a hypothesis that happier

people will tweet with happy keywords and unhappy people will tweet with depressive keywords. World Happiness Report (WHR-2020) is a significant scale of happiness ranking of 156 countries. Gallup World Poll happiness scores are used in ranking happiness. We observed a decrease in Gallup data of WHR-2020 in Georgia to New York and Sri Lanka. Our results in HI-1 and HI-2 show a similar decline.

It is observed that during these two weeks of observations (HI-1 and HI-2), Sri Lanka has less than the yearly average happiness reported by WHR-2020. New York showed balanced happiness during 4-10 July, but more happiness during 25-31 July. Georgia recorded a better happy situation in both weeks of observation while compared with the WHR-2020 report. Our two-week data HI-1 and HI-2 delivered the keywords that impacted the Happiness Index. The results from our observations indicate that Happiness Index in New York state has more impact from failure, hopeless, nervous, tired and comfort in both weeks. Georgia State's Happiness Index has more influence from nervous and peaceful in the HI-1 week and Hopeless, nervous, tired, and peaceful in the HI-2 week. Sri Lanka Happiness Index did not show any specific impact. We analyzed space tourism tweets with 'unit22', 'blue origin', and 'new shepherd keywords, which exhibited an effect on the depressive and anti-depressive moods during July 2021.

In another study, the happiness is computed during two weeks of observation and found that people in Georgia are happier in the HI-2 (Second week of observation) observation period. People in New York are happier in both weeks. However, they are happier in the HI-1 (first week of observation). Contrary to these observations, the Sri Lanka residents are more unhappy than average in both weeks of observation. Nervous, Hopeless, and Tired are the depressive states, and comfort and peaceful are the anti-depressive states that contributed substantially to HI computations.

A new moving average model was developed where we took major values from simple moving average computations. Two other small values are added to this to refine the approximations. Our results were better than the simple moving average model. The error percentages with our model and simple moving average model are compared. Failure, hopeless, active, tired, restless, worthless, calm, comfort, delight, hopeful, and corona showed improvement in all the keywords.

COVID-19 pandemic is a worldwide outbreak that affects many people. Our observations supported by WHO reports that COVID-19 developed into a pandemic from 11 March 2020. Our classification of COVID-19 shifted to other stages, as shown in Fig. 10.11. Depressive and Anti-depression tweet data during COVID-19 to predict for the next five days (1-Apr-2020 to 5-Apr-2020). We found a good association with the COVID-19 pandemic pattern using the tweets and with the WHO reports. The time series prediction system showed promising results for the next few days with our model. The proposed model will be beneficial for remedial non-clinical applications for helping the affected people.

11 FUTURE WORKS

11.1 Tweets during notable events

Our tweets data is increasing every day with depressive and anti-depressive tweets in general and location-specific of all the fifty states of the USA. We also add specific event tweet data to study the event significance and interpret them. We compare their influence with regular days. Such impact analysis will be helpful for future governmental and non-governmental schemes for people around that area.

11.2 World Health Organization Action Plan

WHO has initiated an action plan for the period 2013-2030, and the objectives mentioned are (a) to strengthen effective leadership and governance for mental health, (b) to provide comprehensive, integrated, and responsive mental health and social care services in community-based settings, (c) to implement strategies for promotion and prevention in mental health and (d) to strengthen the information systems, evidence, and research for mental health. WHO opined options to strengthen the information systems, evidence, and research for mental health using the indicators within the information systems. We collect and analyze territory-wise data and develop a framework that delivers results that opens new collaborations in interdisciplinary research in national, international research centers working in mental health, aligning the vision of WHO.

11.3 Real-time atlas of Hedonometric data

At present, we have about four million tweets related to depressive and anti-depressive tweets. In our works, we computed the Happiness Index for Sri Lanka, Georgia, and the New York States for a particular period. Our findings showed similar results with other similar data available. We propose to collect the depressive and anti-depressive tweets data state-wise of the USA daily and prepare an atlas of Happiness Index in real-time for a day, week, month, and year.

We propose this framework with automation will be available in the public domain. An interactive USA map will be generated and will show the WalletHub data and our data simultaneously. The change of the happiness index of a particular state will influence the societal schemes for better living.

11.4 Forecasting mental health states

Covid-19 has generated huge data, and we intend to develop a novel intelligent time series prediction system using both COVID-19 data and tweets to predict the future mental health status of people at a location. We also collect the tweets from the desired areas and forecast using time series prediction mental health status system and then improve it. Our studies will quantify the Depression of the location at that time and date and will be helpful in the computation of Hedonometric parameters.

We collected the Covid-19 data and predicted for one day using our moving average method. Our results showed a good similarity with WHO data. We propose to identify forthcoming events and collect specific 'event' data daily. We attempt to forecast the event parameters for a day using our moving average method for a given date and event. Cyclones, Space tourism, Covid-19 are the events presently available events in our database.

11.5 Natural Language Processing studies

The global artificial intelligence software market would witness massive growth. Applications like NLP, Robotic Process Automation (RPA), and Machine Learning primarily amount to the AI market. Researchers are using the data derived from web search, advertising, emails, customer service, language translation, virtual agents, medical reports, etc., in the efficient estimation of word representations and thereby deduce AI results. We attempt to use the tweet texts to estimate the distributed representations of words and phrases and their compositionality.

11.6 Development of a Website for real-world applications

We collect the tweets daily and add them to our database. We develop an interactive website for public use to get (a) Top three depressive keywords of a given date and location, (b) Dependable parameters of these depressive keywords of the day, which are the prominent events that occurred during that day, (c) Depressive keywords classification for a day and location, (d) Happiness Index of a day or period of a location, (e) Depressive Index of a day or a period of a location, (f) generate a happiness map of a period and geographical location, and (g) forecast the depressive keywords in a particular location such as a city, and a county.

REFERENCES

[1] "https://www.psychiatry.org/patients-families/what-is-mental-illness," [Online]. [Accessed 20 August 2018].

[2] "http://www.uniteforsight.org/mental-health/," [Online]. [Accessed 10 September 2021].

[3] "https://www.nap.edu/resource/other/dbasse/wellbeing-tools/interactive/resources.html," [Online]. [Accessed 19 March 2020].

[4] C. J. Murray and A. D. Lopez, "The Global burden of disease : a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020," Harvard School of Public Health on behalf of the World Health Organization and the World Bank, Harvard, 1996.

[5] C. Murray and A. D. Lopez, "Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study," Lancet, 1997.

[6] A. Roy, Comprehensive Textbook of Psychiatry, B. Sadock and V. Sadock, Eds., Philadelphia: Lippincott Williams & Wilkins, 2000, pp. 2031-2040.

[7] M. De Choudhury, M. Gamon, S. Counts and E. & Horvitz, "Predicting depression via social media.," in *In Seventh International AAAI Conference on Weblogs and social media.*, 2013.

[8] M. Park, C. Cha and M. Cha, "Depressive moods of users portrayed in Twitter," in *In Proceedings of ACM SIGKDD Workshop on hearlth care informatics (HI-KDD)*, 2012.

[9] M. Nadeem, M. Horn and G. & Coppersmith, "Identifying depression on Twitter," 2016.

[10] R. Gwynn, L. M. Hunter, H. Katharine, McVeigh, K. Renu, R. Thomas, Frieden and E. T. Lorna, "Prevalence, dianosis and treatment of depression and generalized anxiety disorder in a diverse urban community," *Psychiary Serv,* vol. 59, no. 6, pp. 641-647, 2008.

[11] L. J. , J. Du, H. Tao and Y. Zhang, "Exploring Temporal Patterns of Suicidal Behavior on Twitter Patterns," in *IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*, New York, 2018.

[12] B. O'Connor, R. Balasubbramanyan, B. Routledge and N. Smith, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington DC, 2010.

[13] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *In Proceedings of the 19th international conference on World wide web (WWW '10)*, New York, 2010.

[14] "http://www.euro.who.int/en/health-emergencies/corono-covid-19/news/news/2020/3/who-annouces-covid-19-outbreak-a-pandemic," 2020. [Online]. [Accessed 18 April 2020].

[15] D. Scanfeld, V. Scanfeld and E. L. Larson, "Dissemination of health information through social networks: twitter and antibiotics," *American journal of infection control,* vol. 38, no. 3, pp. 182-188, 2010.

[16] N. Seeman, "Use data to challenge mental-health stigma: web surveys of attitudes towards mental illness reveal the size of the problem--and offer a way to find fixes," *Nature,* vol. 582, no. 7582, 2015.

[17] A. Wongkoblap, M. Vadillo and V. Curcin, "Detecting and Treating mental Illness on Social Networks," in *5th IEEE International Conference on Healthcare Informatics, ICHI 2017*, Park City USA.

[18] S. Petrovic, M. Osborne and V. Lavrenko, "Streaming first story detection with application to Twitter," in *In Human Language TEchnologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*, Stroudsburg PA, 2010.

[19] D. Wenwen, W. Xiaoyu, R. William and Z. Michelle, "Event detection in Social Media data," in *IEEE Vis Week Workshop on Interactive Visual Text Analytics - Task driven Analytics of Social Media Content*, Willow AB, 2012.

[20] Z. Lingxue and L. Nikolay, "Deep and Confident Prediction for Time Series at Uber," in *IEEE International Conference on Data Mining Workshops*, 2017.

[21] K. Sato and J. Wang, "Detecting real-time events using tweets," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016.

[22] M. Radoslaw, K. Przemyslaw and K. Dawid , "Predicting Social Network Measures Using Machine Learning Approach," in *In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) (ASONAM '12)*, 2012.

[23] S. Tushara and Y. Zhang, "Using Gradient Methods to Predict Twitter Users' Mental Health with Both COVID-19 Growth Patterns and Tweets," in *2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI)*, 2020.

[24] D. D. Nikolaos , A. D. Doulamis, P. Kokkinos and E. Varvarigos, "Event Detection in Twitter Microblogging," *IEEE Transactions on Cybernetics,* vol. 46, no. 12, pp. 2810-2824, December 2016.

[25] E. J. Adam, S. Mahrotra and N. Venkatasubramanian, "Social Media Alert and Response to Threats to Citizens (SMARTC)," in *Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing*, Pittsburgh, 2012.

[26] H. Achrekar, A. Gandhe, R. Lazarus, Y. Ssu-Hsin and B. Liu, "Predicting Flu Trends using Twitter data," in *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2011.

[27] S. Ishikawa, Y. Arakawa, S. Arakawa and A. Fukuda, "Hot topic detection in local areas using Twitter and Wikipedia," in *ARCS Workshops, ARCS 2012 [6222198] (ARCS Workshops, ARCS 2012)*, 2012.

[28] E. Kalampokis, E. Tambouris and K. Tarabanis, "Understanding the predictive power of Social media," *Internet Research,* vol. 23, no. 5, pp. 544-559, 2013.

[29] H. Schoen, D. Gayo-Avello, P. Metaxas, E. Mustafaraj and M. Strohmaier, "The power of prediction with social media," *Internet Research,* vol. 23, no. 5, pp. 528-543, 2013.

[30] H.-H. Shuai, C.-Y. Shen, D.-N. Yang, Y.-F. Lan, W.-C. Lee, P. Yu and M.-S. Chen, "Mining online social data for detecting social network mental disorders," in *In Proceedings of the 25th International Conference on World Wide Web*, 2016.

[31] P. Resnik, A. Garron and R. Resnik, "Using topic modeling to improve prediction of neuroticism and depression," in *In Proceedings of the 2013 Conference on Empirical Methods in Natural*, 2013.

[32] M. G. M. C. S. H. E. De Choudary, "Predicting depression via Social Media," in *Seventh AAAI Confrence on Weblogs and Social Media*, 2013.

[33] S. Tushara and Y. Zhang, "Analyzing Tweets to Discover Twitter Users' Mental Health Status by a Word-Frequency Method," in *2019 IEEE International Conference on Intelligent Systems and Green Technology (ICISGT)*, Visakhapatnam, 2019.

[34] W. Chou, Y. Hunt, E. Beckjord, R. Moser and B. Hesse, "Social media use in the United States: implications for health communication," *J Med Internet Res.,* vol. 11, no. 4, p. E48, 27 November 2009.

[35] M. Park, "Depressive moods of users portrayed in Twitter," in *ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, 2012.

[36] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh and H. Ohsaki, "Recognizing Depression from Twitter Activity," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[37] M. Keracher, "Young Health Movement," 2017. [Online]. Available: https://www.rsph.org.uk/static/uploaded/d125b27c-0b62-41c5-a2c0155a8887cd01.pdf. [Accessed 12 September 2021].

[38] L. Lin, J. Sidani, A. Shensa, A. Radovic, E. Miller , J. Colditz, B. Hoffman, L. Giles and B. Primack, "Association between Social Media use and depression among US Young adults," *Depression and anxiety,* vol. 33, no. 4, pp. 323-331, 2016.

[39] I. Passos, B. Mwangi, B. Cao, J. Hamilton, M. Wu, X. Zhang, Z.-S. Giovana B, Joao Queveodo, Marcia Kauer-Sant'Anna and Flávio Kapczinski, "Identifying a clinical signature of suicidality among patients with mood isorders: A pilot study using machine learning approach," *J Affect Disord,* vol. 193, pp. 109-116, 2016.

[40] S. Tushara and Y. Zhang, "Finding a Depressive Twitter User by Analyzing Time Series Tweets," in *2020 IEEE India Council International Subsections Conference (INDISCON)*, Visakhapatnam, 2020.

[41] S. Tushara and Y. Zhang, "Finding a Depressive Twitter User by Analyzing Depress and Anti-depressant Tweets," in *2020 IEEE India Council International Subsections' Conference (INDISCON)*, Visakhapatnam.

[42] Y. Kelly, A. Zilanawala, C. Booker and A. Sacker, "Social media use and adolescent mental health: Findings from the UK Millennium Cohort Study," *EClinical Medicine,* vol. 6, pp. 59-68, 2018.

[43] J. Twenge and W. Campbell, "Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population based study," in *Preventative Medicine Reports*, 2018.

[44] K. Mogg, B. Bradley, R. Williams and A. Mathews, "Subliminal processing of emotional information in anxiety and depression," *J Abnorm Psychol,* vol. 102, pp. 304-311, 1993.

[45] S. Tushara and Y. Zhang, "A New Method for Discovering Daily Depression from Tweets to Monitor Peoples Depression Status," in *Second IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI 2020),*, Irvine CA, 2020.

[46] N. Lane, M. Mohammod, M. Lin, X. Yang, H. Lu, S. Ali, A. Doryab, E. Berke, T. Choudhury and A. Campbell, "BeWell: A Smartphone Application to Monitor ,Model and Promote Wellbeing," *Pervasive Health Convergence,* 2011.

[47] M. Rabbi, S. Ali, T. Choudhury and E. Berke, "Passive and in-situ assessment of mental and physical well-being using mobile sensors," in *In: Proceedings of the 13th International conference on Ubiquitous computing. UbiComp '11*, 2011.

[48] M. Mendoza, B. Poblete and C. Castillo, "Twitter Under Crisis: Can we trust what we RT?," in *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*, 2010.

[49] M. Gupta, R. Li and K. Chang, "Towards a social media analytics platform: event detection and user profiling for twitter," in *Proceedings of the 23rd International Conference on World Wide Web*, Seoul Korea ACM New York NY USA, 2014.

[50] A. Hossny and L. Mitchell, "Event Detection in Twitter: A Keyword Volume Approach," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, Singapore.

[51] Z. Liu, Y. Huang and R. Joshua, "LEDS: local event discovery and summarization from tweets," in *In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPACIAL '16)*, New York USA, 2016.

[52] K. Florian and V. Antal, "Event detection in Twitter: A machine-learning approach based on term pivoting," in *Proceedings of the 26th Benelux Conference on Artificial Intelligence*, 2014.

[53] Y. Ohta, M. Mine, M. Wakasugi, E. Yoshimine, Y. Himuro, M. Yoneda, S. Yamaguchi, A. Mikita and T. Morikawa, "Psychological effect of the Nagasaki atomic bombing on survivors after half a century," *Psychiatry and Clinical Neurosciences,* vol. 54, no. 1, pp. 97-103, 2000.

[54] F. Yasmin and H. Maria, "Gender differences in anxiety, depression and stress among survivors of suicide bombing," *Pakistan Journal of Social and Clinical Psychology,* vol. 8, no. 2, pp. 145-153, 2010.

[55] S. Tushara and Y. Zhang, "Clustering Depressed and Anti-Depressed keywords Based on a Twitter Event of Srilanka Bomb Blasts using text mining methods," in *Second IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI 2020)*, Irvine CA USA, 2020.

[56] W. Stephen and C. Paris, "Understanding Public Emotional Reactions on Twitter," in *In Proceedings of the Ninth International AAAI Conference on Web and Social Media*, England, 2015.

[57] M. Sykora, T. Jackson, A. O'Brien, S. Elayan and A. Lunen, "Twitter based analysis of public, fine-grained emotional reactions to significant events," in *Proceedings of the European Conference on Social Media ECSM2014*, University of Brignton UK, 2014.

[58] K. Haewoon, C. Lee, H. Park and S. Moon, "What is twitter, a social network or a news media?," in *In proceedings of the 19-international conference on World wide web, WWW '10*, New York NY USA, 2010.

[59] P. Nirmata, K. Rabah, O. Kendal, C. Cynthia, G. Rachel , H. Liz, M. Cailey and C. Priya, "The Implications of COVID-19 for Mental Health and Substance Use," [Online]. Available: https://www.kff.org/health-reform/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/, April 21, 2020. [Accessed 25 April 2020].

[60] L. Hawryluck, W. Gold, S. Robinson, S. Pogorski, S. Galea and R. Styra, "SARS control and psychological effects of quarantine, Toronto, Canada," *Emerging infectious diseases,* vol. 10, no. 7, pp. 1206-1212, 2004.

[61] E. Robertson, K. Hershenfield, S. Grace and D. Stewart, "The psychosocial effects of being quarantined following exposure to SARS: a qualitative study of Toronto health care workers," *Can J Psychiatry,* vol. 49, pp. 403-07, 2004.

[62] U. Pellecchia, R. Crestani, T. Decroo, R. Van den Bergh and Y. Al-Kourdi, "Social Consequences of Ebola Containment Measures in Liberia," *PloS one,* vol. 10, no. 12, 2015.

[63] "https://news.un.org/en/story/2020/03/1059542," [Online]. [Accessed 6 April 2020].

[64] "https://www.worldometers.info/coronavirus/coronavirus-death-toll/," [Online]. [Accessed 17 April 2020].

[65] O. F. Wahl, "Stigma as a barrier to recovery from mental illness," *Trends Cogn Sci,* vol. 16, no. 1, pp. 9-10, January 2012.

[66] R. Kessler , P. Berglund, M. Bruce, J. Koch, E. Laska, P. Leaf, R. Manderscheid, R. Rosenheck, E. Walters and P. Wang, "The prevalence and correlates of untreated serious mental illness," *HSR: Health Services Research,* vol. 36, no. 6, pp. 987-1007, 2001.

[67] S. Saxena, G. Thornicroft, M. Knapp and H. Whiteford, "Resources for mental health: scarcity, inequity, and inefficiency," *Lancet,* vol. 370, pp. 878-889, 2007.

[68] B. Saraceno, O. M. van, R. Batniji, A. Cohen, O. Gureje, J. Mahoney, D. Sridhar and C. Underhill, "Barriers to improvement of mental health services in low-income and middle-income countries," *Lancet,* vol. 370, pp. 1164-1174, 2007.

[69] D. Chisholm, A. Flisher, C. Lund, V. Patel, S. Saxena, G. Thornicroft and M. Tomlinson, "Scale up services for mental disorders: a call for action," *Lancet,* vol. 370, pp. 1241-1252, 2007.

[70] "https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200331-sitrep-71-covid-19.pdf?sfvrsn=4360e92b_8," [Online]. [Accessed 17 April 2020].

[71] L. Silver, "The scope of the problem in children and adolescents," in *Chronic mental illness in children and adolescents*, J. Looney, Ed., Washington, American Psychiatric Press, 1988, pp. 39-51.

[72] G. Klerman and M. Weissman, "Increasing rates of depression," *JAMA,* vol. 261, no. 15, pp. 2229-2235, 1989.

[73] R. Mudgal, R. Niyogi, A. Milani and V. Franzoni, "Analysis of tweets to find the basis of popularity based on events semantic similarity," *International Journal of Web Information Systems,* vol. 14, no. 4, pp. 438-452, 2018.

[74] V. Letica, A. Gaston, R. Sara and Jaeger, "Use of emoticon and emoji in tweets for food-related emotional expression," *Food Quality and Preference,* vol. 49, pp. 119-128, 2016.

[75] A. Gopanarayan and S. Deshpande, "Tweets Analysis for Disaster Management: Preparedness, Emergency Response, Impact, and Recovery," *Innovative Data Communication Technologies and Application. ICIDCA 2019. Lecture Notes on Data Engineering and Communications Technologies,* vol. 46.

[76] R. Bakshi, N. Kaur, R. Kaur and G. Kaur, "Opinion mining and sentiment analysis," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016.

[77] M. Florian and E. David, "Tweets I've seen: analysing factors influencing re-finding frustration on Twitter," in *In Proceedings of the 5th Information Interaction in Context Symposium (IIiX '14). Association for Computing Machinery*, New York, 2014.

[78] A. Tonon, P. Cudré-Mauroux, A. Blarer, V. Lenders and B. Motik, "ArmaTweet: Detecting Events by Semantic Tweet Analysis," in *The Semantic Web. ESWC 2017 Lecture notes in Computer Science*.

[79] J.-Y. Antoine , S.-C. Alejandra , T. Karen, R.-d. Jimena , T.-A. Alessandra, D. Daniela and C. Yhuri , "Association between social media use (Twitter, Instagram, Facebook) and depressive symptoms: Are Twitter users at higher risk?," *International Journal of Social Psychiatry,* vol. 65, no. 1, 2019.

[80] Tutaj, Karolina, Reijmersdal and Eva, "Effects of online advertising format and persuasion knowledge on audience reactions," *Journal of Marketing Communications,* vol. 18, pp. 5-18, 2012.

[81] E. Zerubavel, Seven Day Circle, Chicago: University of Chicago Press, 1989.

[82] A. Orsama, E. Mattila, M. Ermes, M. van Gils, B. Wansink and I. Korhonen, "Weight rhythms: weight increases during weekends and decreases during weekdays," *Obesity facts,* vol. 7, no. 1, pp. 36-47, 2014.

[83] "https://blog.hootsuite.com/twitter-statistics/," [Online]. Available: https://blog.hootsuite.com/twitter-statistics/. [Accessed 15 July 2021].

[84] T. Sakaki, M. Okazaki and Y. Matsuo, "Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development," *IEEE Transactions on Knowledge and Data Engineering,* vol. 25, no. 4, pp. 919-931, April 2013.

[85] L. Laranjo, A. Arguel, A. Neves, A. Gallagher, R. Kaplan, N. Mortimer, G. Mendes and A. Lau, "The influence of social networking sites on health behavior change: a

systematic review and meta-analysis," *Journal of the American Medical Informatics Association: JAMIA,* vol. 22, no. 1, pp. 243-256, 2015.

[86] B. Hasler, J. Allen, D. Sbarra, R. Bootzin and R. Bernert, "Morningness-eveningness and depression: preliminary evidence for the role of the behavioral activation system and positive affect," *Psychiatry research,* vol. 176, no. 2-3, pp. 166-173, 2010.

[87] H. Xu, G. Xingyu and P. Shuai, "Analysis of Tweet Form's effect on users' engagement on Twitter," *Cogent Business & Management,* vol. 6, no. 1, p. 1564168.

[88] G. Shahi, A. Dirkson and T. Majchrzak, "An exploratory study of COVID-19 misinformation on twitter," 2020.

[89] Juntunen, Mari, Ismagilova, Elvira, Oikarinen and Eeva-liisa, "B2B brands on Twitter: Engaging users with a varying combination of social media content objectives, strategies, and tactics," *Industrial Marketing Management,* vol. 89, 2019.

[90] S. Kunal, P. B. Brian and R. M. George, "Should tweets differ for B2B and B2C? An analysis of Fortune 500 companies' Twitter communications," *Industrial Marketing Management,* vol. 43, no. 5, pp. 873-881, 2014.

[91] P. Maria, "Twitter-Based Dissemination of Corporate Disclosure and the Intervening Effects of Firms' Visibility: Evidence from Australian-Listed Companies," *Journal of Information Systems,* vol. 29, no. 2, pp. 107-136, 1 August 2015.

[92] R. Kessler, G. Andrews, L. Colpe, E. Hiripi, D. Mroczek, S. Normand, E. Walters and A. Zaslavsky, "Short screening scales to monitor population prevalence's and trends in non-specific psychological distress," *Psychological Medicine,* vol. 32, no. 6, pp. 959-976, 2002.

[93] M. Ferro, "The Psychometric Properties of the Kessler Psychological Distress Scale (K6) in an Epidemiological Sample of Canadian," *The Canadian Journal of Psychiatry,* vol. 64, no. 9, pp. 647-657, 2019.

[94] M. Kourosh, "Unsupervised Feature Extraction Using Singular Value Decomposition," in *Procedia Computer Science, In ICCS 2015 International Conference On Computational Science Selection*.

[95] C.-H. ChengCheng and H.-H. Chen, "Sentimental text mining based on an additional features method for text classification," in *PLoS ONE*, 2019.

[96] L. S. Jeffrey, Text Data Mining: Theory and Methods, Statistics Surveys, vol. 2, 2008, pp. 94-112.

[97] R. Wright, B. Richmond, A. Odlyzko and B. McKay, "Constant Time Generation of Free Trees," *SIAM J. Computing,* vol. 15, pp. 540-548, 1986.

[98] P. Shili, H. Qinghua, C. Yinli and D. Jianwu, "Improved support vector machine algorithm for heterogeneous data," *Patten Recognition,* vol. 48, no. 6, pp. 2072-2083, 2015.

[99] P. Soucy and G. Mineau, "A simple KNN algorithm for text categorization," in *Proceedings 2001 IEEE International Conference on Data Mining*, San Jose CA USA, 2001.

[100] E. Alwagait and B. Shahzad, "Maximization of Tweet's viewership with respect to time," in *2014 World Symposium on Computer Applications & Research (WSCAR), IEEE 2014*.

[101] S. Basit, "Best and the worst times to Tweet: An Experimental Study," in *8th International Conference on Management, Marketting and Finances (MMF-14)*, Boston MA, 2014.

[102] J. Harada, D. Darmon, M. Girvan and W. Rand, "Forecasting high tide: Predicting times of elevated activity in online social media," in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2015.

[103] N. Grinberg, B. Naaman, B. Shaw and G. Lotan, "Extracting Diurnal Patterns of Real-World Activity from Social Media," in *ICWSM*, 2013.

[104] "https://blog.hubspot.com/marketing/best-times-post-pin-tweet-social-media-infographic," [Online]. Available: https://blog.hubspot.com/marketing/best-times-post-pin-tweet-social-media-infographic. [Accessed 15 July 2021].

[105] "https://www.contentcal.io/blog/best-time-to-post-twitter-2021/," 2021. [Online]. [Accessed 15 July 2021].

[106] "https://buffer.com/resources/best-time-to-tweet-research/," 15 July 2021. [Online].

[107] "https://covid19.who.int," [Online]. [Accessed 15 July 2021].

[108] B. Everitt and G. Dunn, Applied Multivariate Data Analysis, London, 2001.

[109] S. Cramer and I. Becky, "https://www.rsph.org.uk/our-ork/policy/social-media-andyoung-people-s-mental-healthand-wellbeing.html," [Online]. [Accessed 24 March 2019].

[110] C. Davide, "Ten quick tips for machine learning in computational biology," *BioData Mining,* 2017.

[111] J. Zhu, X. Zhao, H. Li, H. Chen and G. Wu, "An Effective Machine Learning Approach for Identifying the Glyphosate Poisoning Status in Rats Using Blood Routine Test," *IEEE Access,* vol. 6, pp. 15653-15662.

[112] A. Schlemmer, H. Zwirnmann, M. Zabel, U. Parlitz and S. Luther, "Evaluation of machine learning methods for the long-term prediction of cardiac diseases," in *8th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO)*, Trento, 2014.

[113] "Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines," *Knowledge Inf Syst,* vol. 58, no. 139, 2019.

[114] R. Dela, . &. S. Kevin, &. L. Rushin , . &. G. Bo, F. Anatole and Robert, "Topical Clustering of Tweets," in *3rd Workshop on Social Web Search and Mining.* , 2011.

[115] E. Bralis, T. Cerquitelli, S. Chiusano, L. Grimaudo and X. Xiao, "Analysis of Twitter Data Using a Multiple-level Clustering Strategy," in *In: Third International Conference on Model and Data Engineering (MEDI 2013)*, Amantea Italy, 2013.

[116] G. Dagmar and D. Thierry, "Hashtag processing for enhanced clustering of tweets," in *Proceedings of Recent Advances in Natuaral Language Processing* , Varna Bulgaria, 2017.

[117] T. Jan, "Clustering of Tweets: A Novel Approach to Label the Unlabelled Tweets," in *Proceedings of ICRIC 2019. Lecture Notes in Electrical Engineering*, 2019.

[118] Q. Wei and R. J. Dunbrack, "The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics," *PLoS ONE,* vol. 8, no. 7, 2013.

[119] H. Whiteford, A. Ferrari, L. Degenhardt, V. Feigin and T. Vos, "The global burden of mental, neurological and substance use disorders: An analysis from the global burden of disease study," *PLoS ONE,* vol. 10, no. 2, 2015.

[120] A. Ferrai, F. Charlson, R. Norman, S. Pattern and C. Murray, "Burden of depressive disorders by country, sex, age and year: Findings from the Global Burden of Disease Study," *PLOS Medicine 10: e1001547,* 2013.

[121] J.-Y. Antoine, S.-C. Alejandra , T. Karen , R.-d. Jimena , T.-A. Alessandra, D. Daniela and C. Yhuri , "Association between social media use (Twitter, Instagram, Facebook) and depressive symptoms: Are Twitter users at higher risk?," *International Journal of Social Psychiatry,* vol. 65, no. 1, pp. 14-19, 2019.

[122] I. Pantic, A. Damjanovic, J. Todorovic, D. Topalovic, D. BojovicJovic, S. Ristic and S. Pantic, "Association between online social networking and depression in high school students," *Behavioral physiology viewpoint. Psychiatria Danubina,* vol. 24, pp. 90-93, 2012.

[123] M. Merono, L. Jelenchick, R. Koff and J. Eickhoff, "Depression and Internet Use among Older Adolescents: An Experience Sampling Approach," *Psychology,* vol. 3, pp. 743-748, 2012.

[124] N. Reavley and P. Pilkington, "Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study.," *PeerJ,* vol. 2, no. e647, 2014.

[125] A. Shepard, S. Caroline, D. Michael and S. Jenny, "Using social media for support and feedback by mental health service users: thematic analysis of a twitter conversation," *BMC Psychiatry,* vol. 15, 2015.

[126] P. Cavazos-Rehg, M. Krauss, S. Sowles, S. Connolly, C. Rosas, M. Bharadwaj and L. Bierut, "A content analysis of depression-related Tweets," *Comput Human Behav,* vol. 54, pp. 351-357, 1 Jan 2016.

[127] J. Singh, D. K. Yogesh, P. R. Nripendra, K. Abhinav and K. K. Kawaljeet, "Event classification and location prediction from tweets during disasters," *Annals of Operations Research,* vol. 283, pp. 737-757, 2019.

[128] L. Ryong and S. Kazutoshi, "Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection," in *In Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks (LBSN '10)*, New York NY, 2010.

[129] S. N. Arjun, P. K. Anathu, C. Naveen and R. Balasubramani, "Survey on Pre-Processing Techniques for Text Mining," *International Journal of Engineering and Computer Science,* vol. 5, no. 6, pp. 148-157, 2018.

[130] M. Samanesh and E. Martin, "On the design of LDA models for aspect-based opinion mining," in *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12)*, New York NY USA, 2012.

[131] A. Weiler, M. Grossniklaus and M. Scholl, "Run-time and task-based performance of event detection techniques for twitter," in *Advanced Information Systems Engineering : 27th International Conference, CAiSE 2015*, Stockholm, Sweden, 2015.

[132] S. Choudary and H. Alani, "Personal life event detection from social media," 2014.

[133] L. Cui, X. Zhang, X. Zhou and F. Salim, "Topical event detection on Twitter," in *Australasian Database Conference*, 2016.

[134] "https://www.ft.lk/front-page/Tourist-hotels-fear----1-5-b-revenue-loss-from-terror-attacks/44-676993," [Online]. Available: https://www.ft.lk/front-page/Tourist-hotels-fear----1-5-b-revenue-loss-from-terror-attacks/44-676993. [Accessed 16 June 2021].

[135] "https://www.cnbc.com/2019/09/11/sri-lankas-economy-tourism-sector-after-easter-sunday-bombings.html," [Online]. [Accessed 16 June 2021].

[136] "http://www.rsmcnewdelhi.imd.gov.in/uploads/report/ 34/34_230b1d_nburevi.pdf," [Online]. Available: http://www.rsmcnewdelhi.imd.gov.in/uploads/report/ 34/34_230b1d_nburevi.pdf. [Accessed 12 June 2021].

[137] "https://earthobservatory.nasa.gov/images/148325/cyclone-tauktae-strikes-india," [Online]. Available: https://earthobservatory.nasa.gov/images/148325/cyclone-tauktae-strikes-india. [Accessed 12 June 2021].

[138] O. Edo-Osagie, G. Smith, I. Lake, O. Edeghere and B. De La Iglesia, "Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance," *Plos ONE,* vol. 14, no. 7, p. e0210689, 2019.

[139] S. De, Lalindra, Riloff and Ellen, "User Type Classification of Tweets with Implications for Event Recognition," in *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, Baltimore, Maryland, 2014.

[140] A. Farzindar and W. Khreich, "A Survey of Techniques for Event Detection in Twitter," *Computational Intelligence,* vol. 31, pp. 132-164, 2015.

[141] "https://covid19.who.int/," [Online]. [Accessed 10 August 2021].

[142] [Online]. Available: https://en.wikipedia.org/wiki/Cyclone_Tauktae. [Accessed 13 June 2021].

[143] "https://en.wikipedia.org/wiki/Cyclone_Burevi," [Online]. Available: https://en.wikipedia.org/wiki/Cyclone_Burevi. [Accessed 13 June 2021].

[144] H. J. Barry and W. Bousfield, "A quantitative determination of euphoria and its relation to sleep," *The Journal of Abnormal and Social Psychology,* vol. 29, no. 4, pp. 385-389, 1935.

[145] B. Neugarten, R. Havighurst and S. Tobin, "The measurement of life satisfaction," *Journal of gerontology,* vol. 16, pp. 134-143, 1961.

[146] M. Rosenberg, Society and the adolescent self-image, Princeton, NJ: Princeton University Press, 1965, p. 340.

[147] M. Lawton, "The Philadelphia Geriatric Center Morale Scale: a revision," *Journal of gerontology,* vol. 30, no. 1, pp. 85-89, 1975.

[148] R. Ryan and E. Deci, "On happiness and human potentials : A review of research on hedonic and eudaimonic wellbeing," *Annual Review of Psychology,* vol. 52, pp. 141-166, 2001.

[149] Shin, Doh and D. Johnson, "Avowed happiness as an overall assessment of the quality of life," *Social Indicators Research,* vol. 5, pp. 475-492, 1978.

[150] D. Kahneman, Objective Happiness, Well-Being: The Foundation of Hedonic Psychology, D. Kahneman, E. Diener and N. Schwarz, Eds., New York NY: Russell Sage Foundation, 1999, pp. 3-25.

[151] A. Brooke and A. Monica, "Social Media Use in 2021," Pew Research Center, 2021.

[152] F. Karim, A. Oyewande, L. Abdalla, E. R. Chaudhry and S. Khan, "Social Media Use and Its Connection to Mental Health: A Systematic Review," *Cureus,* vol. 12, no. 6, p. e8627, 2020.

[153] J. Sachs, R. Layard and J. Helliwell, "World Happiness Report. Technical report,," 2012.

[154] B. Carrol, M. Feinberg, P. Smouse and S. J. Greden, "The Carroll Rating Scale for Depression. I. Development, reliability and validation," *British Journal of Psychiatry,* vol. 138, pp. 194-200, 1981.

[155] L. Radloff, "The CED-D scale: A self-Report depression scale for research in the general population," *Applied Psychological Measurement,* vol. 1, pp. 385-401, 1977.

[156] G. J. Parkerson, W. Broadhead and C.K.J., "Anxiety and depressive symptom identification using the Duke Health Profile," *J. Clin. Epidemiol,* vol. 49, no. 1, pp. 85-93, 1996.

[157] K. Kobak, J. Lipsitz and A. Feiger, "Development of a standardized training program for the Hamilton Depression Scale using internet-based technologies: results from a pilot study," *JPsychiatric Res,* vol. 37, pp. 509-515, 2003.

[158] W. W. Zung, "A Self-Rating Depression Scale," *Arch Gen Psychiatry,* vol. 12, pp. 63-70, 1965.

[159] T. Furukawa, R. Kessler, T. Slade and G. Andrews, "The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being," *Psychol Med,* vol. 33, no. 2, pp. 357-362, February 2003.

[160] J. Davitz, A Dictionary and Grammar of Emotions in Arnold, 1970, pp. 251-258.

[161] N. Bradburn and D. Caplovitz, Reports on Happiness, N. O. R. Center, Ed., Chicago USA: ALDINE Publishing Company, 1965.

[162] R. Irwin, R. Kammann and G. Dixon, "If You Want to Know How Happy I am You'll Have to Ask Me," *New Zealand Psychologist,* vol. 8, pp. 10-12, 1979.

[163] R. Kammann, D. Christie, R. Irwin and G. Dixon, "Properties of an Inventory to Measure Happiness (and Psychological Health)," *New Zealand Psychologist,* vol. 8, pp. 1-9, 1979.

[164] B. Underwood and W. Froming, "The Mood Survey: A Personality Measure of Happy and Sad Moods," *Journal of Personality Assessment,* vol. 4, pp. 404-414, 1980.

[165] A. Oswald and S. Wu, "Objective Confirmation of Subjective Measures of Human Well-Being: Evidence from the U.S.A," *Science,* vol. 327, pp. 576-579, Jan 2010.

[166] G. Touburg and R. Veenhoven, "Mental Health Care and Average Happiness: Strong Effect in Developed Nations," *Adm Policy Ment Health,* vol. 42, pp. 394-404, 2015.

[167] D. Galati, M. Manzano and I. Sotgiu, "The subjective components of Happiness and their attainment: a cross-cultural comparison between Italy and Cuba," *Social Science Information,* vol. 45, no. 4, pp. 601-630, 2006.

[168] T. Fryers, D. Melzer, R. Jenkins and T. Brugha, "The distribution of the common mental disorders: social inequalities in Europe," *Clin Pract Epidemiol Ment Health : CP & EMH,* pp. 1-14, 2005.

[169] E. Dienner, Suh, Eunkook, Oishi and Shigehiro, "Recent findings on subjective well-being," *Indian Journal of Clinical Psychology,* vol. 24, pp. 25-41, 1997.

[170] P. Dodds, K. Harris, I. Kloumann, C. Bliss and C. Danforth, "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter," *Plos ONE,* vol. 6, no. 12, p. e26752, 2011.

[171] J. Fowler and N. Christakis, "Dynamic spread of Happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study," *BMJ,* p. 337:a2338, 4 December 2008.

[172] K. Kircanski, M. Lieberman and M. Craske, "Feelings Into Words: Contributions of Language to Exposure Therapy," *Psychological Science,* vol. 23, no. 10, pp. 1086-1091, 2012.

[173] L. Barrett, "Solving the emotion paradox: categorization and the experience of emotion," *Pers Soc Psychol Rev,* vol. 10, no. 1, pp. 20-46, 2006.

[174] A. Wright, A. Jorm and A. Mackinnon, "Labels used by young people to describe mental disorders: which ones predict effective help-seeking choices?," *Soc Psychiatry Psychiatr Epidemiol,* vol. 47, pp. 917-926, 2012.

[175] D. Rose, G. Thornicroft, V. Pinfold and A. Kassam, "250 labels used to stigmatise people with mental illness," *BMC Health Serv Res,* vol. 97, pp. 1-7, 2007.

[176] T. Jay, Why we curse: A neuro-psycho-social theory of speech, John Benjamins Publishing, 2000.

[177] S. Lyubomirrsky and H. Lepper, "A measure of subjective happiness: Preliminary reliability and construct validation," *Social Indicators Research,* vol. 46, no. 2, pp. 137-155, 1999.

[178] M. Seligman, T. Steen, N. Park and C. Peterson, "Positive psychology progress: empirical validation of interventions," *The American psychologist,* vol. 60, no. 5, pp. 410-421, 2005.

[179] A. Beck, C. Ward, M. Mendelson, J. Mock and J. Erbaugh, "An inventory for measuring depression," *Archives of general psychiatry,* vol. 4, pp. 561-571, 1961.

[180] S. Lyubomirsky and L. Ross, "Hedonic consequences of social comparison: a contrast of happy and unhappy people," *Journal of personality and social psychology,* vol. 73, no. 6, pp. 1141-1157, 1997.

[181] W. Glatzer and J. Gulyas, Cantril Self-Anchoring Striving Scale In: Encyclopedia of Quality of Life and Well-Being Research, A. Michalos, Ed., Springer, Dordrecht, 2014.

[182] N. Bradburn and C. Noll, The structure of psychological well-being, Chicago IL : Aldine, 1969.

[183] G. Gurin, J. Veroff and S. Feld, Americans view their mental health: A nationwide interview survey, Basic Books, 1960.

[184] L. Singh, "Accuracy of Web Survey Data: The State of Research on Factual Questions in Surveys," *Information Management and Business Review,* vol. 3, pp. 48-56, 2011.

[185] C. Andrade, "The Limitations of Online Surveys," *Indian Journal of Psychological Medicine,* vol. 42, no. 6, pp. 575-576, 2020.

[186] J. Evans and A. Mathur, "The value of online surveys," *Internet Research,* vol. 15, no. 2, pp. 195-219, 2005.

[187] C. Cornesse and A. Blom, "Response Quality in Nonprobability and Probability-based Online Panels," *Sociological Methods & Research,* 2020.

[188] K. Kate, C. Belinda, C. Elinda, B. Vivienne and S. John, "Good practice in the conduct and reporting of survey research," *International Journal for Quality in Health Care,* vol. 15, no. 3, pp. 261-266, May 2003.

[189] W. Coster, "Making the best match: Selecting outcome measures for clinical trials and outcome studies," *Am. J. Occup. Ther,* vol. 67, pp. 162-170, 2013.

[190] D. Dfarhud, M. Malmir and M. Khanahamadi, "Happiness & Health: The Biological Factors- Systematic Review Article," *Iranian journal of public health,* vol. 43, no. 11, pp. 1468-1477, 2014.

[191] Helliwell, F. John, L. Richard, S. Jeffrey and D. N. Jan-Emmanuel, Eds., in *World Happiness Report 2020. New York: Sustainable Development Solutions Network*, New York , 2020.

[192] V. Jaswal, K. Kishore, M. Muniraju, N. Jaswal and R. Kapoor, "Understanding the determinants of happiness through Gallup World Poll," *Journal of family medicine and primary care,* vol. 9, no. 9, pp. 4826-4832, 2020.

[193] D. A. Ismu Rini, H. Septiana and S. W. Budi, "Understanding Three Dimensions of Happiness Index in Kedungkandang District, Malang City, Indonesia," *IOP Conference Series: Earth and Environmental Science,* vol. 799, no. 1, 2021.

[194] S. S. Tushara and Y. Zhang, "An outcome-based method for computing happiness index from mental health related tweets of Twitter users," in *2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT – 2021)*, Visakhapatnam India, 2021.

[195] L. V. Ahn, "https://www.cs.cmu.edu/~biglou/resources/bad-words.txt," [Online]. [Accessed 31 May 2021].

[196] "http://covid19.who.int," [Online]. [Accessed 30 May 2021].

[197] M. Madeja and J. Porubän, "Accuracy of Unit Under Test Identification Using Latent Semantic Analysis and Latent Dirichlet Allocation," in *2019 IEEE 15th International Scientific Conference on Informatics*, 2019.

[198] E. Negara, D. Triadi and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," in *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 2019.

[199] "https://wallethub.com/edu/happiest-states/6959," [Online]. [Accessed 16 August 2021].

[200] S. S. Tushara and Z. Yanqing, "Analyzing the Bad-Words in tweets of Twitter users to discover the Mental Health Happiness Index and Feel-Good-Factors," in *2021 In: IEEE International Conference on Data Mining : Social Data Mining in the Post-Pandemic Era (ICDM SDM-2021)* , Auckland, 2021.

[201] G. R. Andrew, J. R. Andrew, L. L. Katharina, S. D. Peter, M. D. Chistopher and J. L. Ellen, "Forecasting the onset and course of mental illness with Twitter data," *Sci Rep,* Vols. 7, 13006, pp. 12961-9, 2017.

[202] Asur, Sitaram, Huberman and Bernardo, "Predicting the Future with Social Media," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010*, 2010.

[203] D. S. Pollock, R. C. Green and T. Nguyen, Handbook of Time Series Analysis, Signal Processing, and Dynamics, Elsevier, 1999.

[204] I. López-Yáñez, L. Sheremetov and C. Yáñez-Márquez, "Associative Model for the Forecasting of Time Series Based on the Gamma Classifier," in *Pattern Recognition*, J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, J. F. Martínez-Trinidad and G. S. di Baja, Eds., Berlin, Heidelberg, Springer Berlin Heidelberg, 2013, pp. 304-313.

[205] F. R. Johnston, J. E. Boyland, M. Meadows and E. Shale, "Some properties of a simple moving average when applied to forecasting a time series," *Journal of the Operational Research Society,* vol. 50, no. 12, pp. 1267-1271, 1999.

[206] R. Hyndman and Y. Khandakar, "Automatic Time Series Forecasting: The forecast Package for R," *Journal of Statistical Software,* vol. 27, no. 3, pp. 1-22, 2008.

[207] "http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic," [Online]. [Accessed 18 April 2020].

[208] "https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200331-sitrep-71-covid-19.pdf?sfvrsn=4360e92b_8," [Online]. [Accessed 17 April 2020].

[209] "https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200331-sitrep-71-covid-19.pdf?sfvrsn=4360e92b_8," [Online]. [Accessed 17 April 2020].

[210] "https://www.worldometers.info/coronavirus/coronavirus-death-toll/," [Online]. Available: https://www.worldometers.info/coronavirus/coronavirus-death-toll/. [Accessed 17 April 2020].

[211] S. Choudhury and H. Alani, "Personal life event detection from social media," in *Workshop Proceedings CEUR*, 2014.

[212] L. Laranjo, A. Arguel, A. Neves, A. Gallagher, R. Kaplan, N. Mortimer, G. Mendes and A. Lau, "The influence of social networking sites on health behavior change: a systematic review and meta-analysis," *Journal of the American Medical Informatics Association: JAMIA,* vol. 22, no. 1, pp. 243-256, 2015.

[213] "https://www.real-statistics.com/time-series-analysis/," [Online]. [Accessed 10 August 2021].

[214] "https://www.worldometers.info/coronavirus/coronavirus-death-toll/," [Online]. [Accessed 17 April 2020].

[215] A. Yossi, F. Amos, K. Anna, M. Frank and S. Jared, "Spectral analysis of data," in *In Proceedings of the thirty-third annual ACM symposium on Theory of computing (STOC '01)*, New York, 2001.

APPENDICES

Appendix A: Tweet data collected from twitter.com

Sample Tweet data

	text	favorited	favoriteCo	replyToSN	created	truncated	replyToSII	id	replyToUII	statusSource	screenNa	retweetC	isRetweet	retweete	longitude	latitude
1	RT @PC98	FALSE	0	NA	9/23/2021 1:24	FALSE	NA	1440849533563370000	NA	<a href="https://mobi	IcarusFW	14	TRUE	FALSE	NA	NA
2	Complete	FALSE	0	NA	9/23/2021 1:10	TRUE	NA	1440846065796720000	NA	<a href="http://twitte	j_t_91	0	FALSE	FALSE	NA	NA
3	RT @PC98	FALSE	0	NA	9/23/2021 1:04	FALSE	NA	1440844454005330000	NA	<a href="http://twitte	MaybeIm	14	TRUE	FALSE	NA	NA
4	Lockedin I	FALSE	0	NA	9/23/2021 1:00	TRUE	NA	1440843399020560000	NA	<a href="https://www	LockedINI	0	FALSE	FALSE	NA	NA
5	Check out	FALSE	0	NA	9/23/2021 0:47	TRUE	NA	1440840193666210000	NA	<a href="http://twitte	PacificRpr	0	FALSE	FALSE	NA	NA
6	Need	FALSE	0	NA	9/23/2021 0:30	TRUE	NA	1440835861826830000	NA	<a href="https://ifttt.(	radoncnot	0	FALSE	FALSE	NA	NA
7	RT @PC98	FALSE	0	NA	9/23/2021 0:22	FALSE	NA	1440833816340220000	NA	<a href="http://twitte	L0l1c0mm	14	TRUE	FALSE	NA	NA
8	RT @PC98	FALSE	0	NA	9/23/2021 0:12	FALSE	NA	1440831425800920000	NA	<a href="https://mobi	AnimeBus	14	TRUE	FALSE	NA	NA
9	RT	FALSE	0	NA	9/23/2021 0:10	FALSE	NA	1440830961403400000	NA	<a href="https://mobi	AgustinOc	5	TRUE	FALSE	NA	NA
10	RT @PC98	FALSE	0	NA	9/23/2021 0:07	FALSE	NA	1440830180226830000	NA	<a href="http://twitte	arimahyo	14	TRUE	FALSE	NA	NA
11	<f0><U+0	FALSE	0	NA	9/23/2021 0:06	TRUE	NA	1440829996168210000	NA	<a href="http://twitte	Nancijean	0	FALSE	FALSE	NA	NA
12	RT @PC98	FALSE	0	NA	9/23/2021 0:05	FALSE	NA	1440829714415810000	NA	<a href="https://mobi	mxngmet	14	TRUE	FALSE	NA	NA
13	RT @PC98	FALSE	0	NA	9/23/2021 0:04	FALSE	NA	1440829325683550000	NA	<a href="http://twitte	InsaneNa	14	TRUE	FALSE	NA	NA
14	RT	FALSE	0	NA	9/23/2021 0:03	FALSE	NA	1440829254447470000	NA	<a href="http://weigh	diethealth	1	TRUE	FALSE	NA	NA

S.No, Tweet Text, favored, favorite count, Reply, Created date and time, Truncated or not, Reply_to, id, reply_to_UID, statusSource, screenname, retweet_count, Is_Retweeted, Longitude, and Latitude.

Sample tweet data collected from twitter.com on 20-July-2021 (unity22)

	text	favorited	favoriteCo	reply	created	truncated	replyToSII	id	replyToUI	statusSou	screenNar	retweetCo	isRetweet	retweeted	longitude	latitude
3539	RT	FALSE	0	NA	7/15/2021 14:19	FALSE	NA	1.42E+18	NA	<a href="h	Marcin015	5475	TRUE	FALSE	NA	NA
3540	RT @richardbr	FALSE	0	NA	7/15/2021 14:19	FALSE	NA	1.42E+18	NA	<a href="h	saliamma	4848	TRUE	FALSE	NA	NA
3541	RT @richardbr	FALSE	0	NA	7/15/2021 14:16	FALSE	NA	1.42E+18	NA	<a href="h	schnodde	4848	TRUE	FALSE	NA	NA
3542	RT	FALSE	0	NA	7/15/2021 14:14	FALSE	NA	1.42E+18	NA	<a href="h	karess_ca	6666	TRUE	FALSE	NA	NA
3543	RT	FALSE	0	NA	7/15/2021 14:11	FALSE	NA	1.42E+18	NA	<a href="h	johnnycas	6666	TRUE	FALSE	NA	NA
3544	RT	FALSE	0	NA	7/15/2021 14:10	FALSE	NA	1.42E+18	NA	<a href="h	xsteguin	645	TRUE	FALSE	NA	NA
3545	RT @Branston	FALSE	0	NA	7/15/2021 14:10	FALSE	NA	1.42E+18	NA	<a href="h	getcarter4	849	TRUE	FALSE	NA	NA
3546	RT @richardbr	FALSE	0	NA	7/15/2021 14:09	FALSE	NA	1.42E+18	NA	<a href="h	Robert_23	3533	TRUE	FALSE	NA	NA
3547	RT @richardbr	FALSE	0	NA	7/15/2021 14:08	FALSE	NA	1.42E+18	NA	<a href="h	nabale96	4848	TRUE	FALSE	NA	NA
3548	RT	FALSE	0	NA	7/15/2021 14:05	FALSE	NA	1.42E+18	NA	<a href="h	np_grwl	5475	TRUE	FALSE	NA	NA
3549	RT @richardbr	FALSE	0	NA	7/15/2021 14:04	FALSE	NA	1.42E+18	NA	<a href="h	ake333333	6364	TRUE	FALSE	NA	NA
3550	RT @prattandt	FALSE	0	NA	7/15/2021 14:00	FALSE	NA	1.42E+18	NA	<a href="h	dancarnel	7	TRUE	FALSE	NA	NA
3551	RT @lorenzap	FALSE	0	NA	7/15/2021 14:00	FALSE	NA	1.42E+18	NA	<a href="h	TommyAr	3	TRUE	FALSE	NA	NA
3552	RT	FALSE	0	NA	7/15/2021 14:00	FALSE	NA	1.42E+18	NA	<a href="h	np_grwl	4	TRUE	FALSE	NA	NA
3553	RT	FALSE	0	NA	7/15/2021 13:58	FALSE	NA	1.42E+18	NA	<a href="h	Ana_An_M	28	TRUE	FALSE	NA	NA
3554	RT @virgingal	FALSE	0	NA	7/15/2021 13:53	FALSE	NA	1.42E+18	NA	<a href="h	GILBERTde	3544	TRUE	FALSE	NA	NA
3555	RT @richardbr	FALSE	0	NA	7/15/2021 13:53	FALSE	NA	1.42E+18	NA	<a href="h	aleifdania	4848	TRUE	FALSE	NA	NA
3556	RT @richardbr	FALSE	0	NA	7/15/2021 13:51	FALSE	NA	1.42E+18	NA	<a href="h	FeisalSala	15809	TRUE	FALSE	NA	NA
3557	RT @richardbr	FALSE	0	NA	7/15/2021 13:49	FALSE	NA	1.42E+18	NA	<a href="h	kong_mar	6947	TRUE	FALSE	NA	NA
3558	RT @virgingal	FALSE	0	NA	7/15/2021 13:47	FALSE	NA	1.42E+18	NA	<a href="h	kong_mar	820	TRUE	FALSE	NA	NA
3559	RT @richardbr	FALSE	0	NA	7/15/2021 13:46	FALSE	NA	1.42E+18	NA	<a href="h	shaorabds	4848	TRUE	FALSE	NA	NA