A·P·U

ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT098-3-2 RMCT

Research Methodology in Computing and Technology (RMCT)

UC2F2008CS(DA)

HAND OUT DATE: 22 MARCH 2021

HAND IN DATE : 7 MAY 2021

WEIGHTAGE : 50%

STUDENT NAME : RUBAN RAJ MURUGAN

TP NUMBER : TP051462

LECTURER NAME : DR. SHUBASHINI RATHINA VELU

Table of Contents

Sentiment Analysis: Lexicon-Based Sentiment Analysis to Identify Depression Indicative Tweets using Twitter as Data Source

Ruban Raj Murugan
Asia Pacific University of Technology & Innovation (APU)
tp051462@mail.apu.edu.my

Abstract

Mental health is as important as physical health. At present, while the whole world encounters the Covid-19 pandemic, mental health problems and cases escalated far more than anyone imagined. Lockdowns and Movement Control Order that has been implemented have contributed to the rise in mental health cases most depression. Depression is known as one of the biggest and well known mental health problems that have affected people during this pandemic era. In midst of all this, people nowadays don’t have many things to do as they were forced to stay at home because of lockdowns and most of them started working from home. This made people spend much more time in their social media applications rather than doing physical activities and social media became a platform for people to share their everyday life happenings. This proposal’s purpose is to propose an efficient and easy method to identify and monitor people mental health especially depressed people through Twitter data. By utilizing the Syuzhet package in R programming, Bi-gram and Tri-gram models and finally matching the tweets with SentiWordNert and AFFIN lexicon dictionaries, potential depression users can be identified by totalling the sentiment score at the end.

1. Introduction

Mental health is a hot topic that has been widely talked about these days. Mental health is a necessary and fundamental part of wellbeing. Based on the WHO official website it is defined as a condition of prosperity where an individual understands their capacities, can adapt to the ordinary anxieties of life, can work gainfully, and can make a commitment to their local area. Mental health is key to our group and individual capacity as people might suspect, act out, cooperate, earn and appreciate life. On this premise, the advancement, insurance and rebuilding of psychological well-being can be viewed as a fundamental worry of people, networks and social orders all through the world (World Health Organization, 2018).

Among the mental health conditions, depression is the most common medical disorder as up to 264 million individuals are influenced around the world, making it one of the major causes for suicide and disability where the suicide count reaches up to

800 000 individuals each year (World Health Organization, 2020). In the Pacific region, the number of individuals who fall under depression is between 1.1% and 19.9%, an average percentage of 3.7% (Eusof Izzudin, Al–Bedri, Subramaniam, Matthews & Ai Theng, 2018). As for in Malaysia, depression is the most well-known mental disease detailed with roughly 2.3 million individuals being

influenced sooner or later in their lives (Mukhtar & P. S. Oei, 2011). Furthermore, inside the Malaysian essential consideration patient populace, the predominance of depression is assessed to be within 6.7 to 14.4% (ZamZam, Thambu, Midin, Omar & Kaur, 2009).

At present, diverse social communication applications have empowered everybody to without any problem express and offer their sentiments with individuals throughout the planet. Plenty of individuals use text for conveying, which should be possible through various social platforms informing stages accessible today such as Facebook, Instagram and Twitter and so forth, as they think that it's simpler to communicate their sentiments through text rather than talking them out. By using social media data, it is easier to collect users sentiment.

Even though the Malaysian Health Organization doing their utmost to help people regarding depression issues the problem is yet not solved completely. Therefore, To help ease the tracking and identifying individuals who are going through depression, social media platform like Twitter is used as a potential data source to do sentiment analysis as most individuals these days share their emotions, thoughts and happenings of their everyday life in social media platform. A model that allows doing a sentiment analysis using Twitter as a data source would assist in identifying individuals who are most likely going through depression, mental health issues.

2. Literature Review

Depression, the most well known mental health condition, is rapidly growing as we can see the number of cases has significantly increased. Researchers have done some research in the past to find factors that cause depression, affecting mental health and impacts of depression among people through several research methods. This section of the paper will investigate and analyse the previous researches done about depression and the methods used to track or find the relevant information.

There were several past research done about depression and mental health among people. As a feature of a public scale exertion to control depression and mental health cases, Centers for Disease Control and Prevention researches depression by directing Behavioral Risk Factor Surveillance Framework survey or review through phone to assess the percentage of depression between grown-ups In the United States (Centers of Disease Control and Prevention, 2020). Anyway, the huge fleeting holes across which these estimations are made and the small number of respondents and very time consuming as it makes it hard for an organization to follow and recognize hazard factors that might be related to mental health and depression, or to grow viably mediation programs. By using social media as the data source these limitations can be overcome as the number of respondents we can collect will not be limited because we are analyzing and detecting potential depressed people using the comments posted on their Twitter page.

Kawachi et al (2001), investigated the job of social ties in the upkeep of mental prosperity and treatment of social wellbeing concerns. In line with this, Rosenquist, Fowler & Christakis (2011) have discovered that degrees of depression indicated dissemination up to three levels of partition in a huge informal social web, recommending a network impact segment to depression. These past researches give solid proof that people's social media contain fundamental data which will be helpful for understanding and interceding on psychological well-being. Even Though the past research about social media relating to depression and mental health have been done this research had limitations of using a basic algorithm which can be overcome by using a more efficient modal to do sentiment analysis using Twitter. The usage of Twitter as the data source for analyzing users mental well-being will be an efficient method as it contains more than enough data to fulfil the requirements.

Other than that, Rude, Gortner & Pennebaker (2004), showed that automated analysis knew as a computerized review of texts has additionally been known to uncover signals about depression and other related mental health illness and also Low, Bentley & Ghosh (2020), exhibited that analyzing speech through linguistic analysis could characterize patients into indicative groups like those who going through depression. These prior studies show that using such methods, especially using online media given their solid association with individuals' social life can help in overcoming the constraints of studies for comprehension of the mental health and mental well-being of people.

Furthermore, J. Paul & Dredze (2011) built an illness explicit point model dependent on Twitter comments and posts all together on model conduct around numerous illnesses of significance in general wellbeing. Cavazos-Rehg et al. (2017) discovered introductory proof that individuals do post about their moods, about their depression and surprisingly they also do share the treatments they went to overcome that depression on Twitter and even recommended that depression and mental health of individuals can be identified using the Twitter data using efficient and new algorithms. Numerous kinds of research have been done to examine the scope of challenges around general wellbeing, research on tackling web-based media for understanding depression and mental health issues is as yet in its early stages but this early work hence focuses on the capability of social media as a sign to use in the investigation about depression. With the current work, we can expand the research of the social network especially Twitter in investigations by analyzing depression and mental health from a new perspective and a bigger scale.

3. Problem Statement

Spike up in mental health cases during the Covid-19 pandemic

As the world encountered the covid-19 pandemic the mental health-related problems started spiking up. As the Malaysian Government imposed Lockdown and Movement Control Order, the mental health cases increased significantly higher. The health ministry of Malaysia has recorded 465 attempted suicide cases between January till July in the year 2020 and most of the suicide cases are because of depression (Bernama, 2020). Mental health has always been an issue in Malaysia, Nearly 500,000 people experience symptoms or mental health-related problems like depression, based on the 2019 national Health and Morbidity Survey (NHMS 2019). Befrienders, a non-governmental organization that aid people in improving the emotional health and well-being stated that they had received around 5059 calls nationwide from April till June 2020 and an average of 770 calls a month in Johor Bahru alone which is 41% of the numbers before that (Muthiah, 2016). During the Covid-19 pandemic, mental health-related problems became the second major and common health condition after heart disease (Ucanews, 2021). This shows that the mental health cases were skyrocketed during the Covid-19 pandemic.

The mental health cases were already increasing in Malaysia, the covid-19 pandemic made it even worse. Government measurements show paces of mental wellness issues have significantly increased in the course of recent many years and the monetary effect on the nation is required to flood to US$6 trillion by 2030. As indicated by the 2019 National Health and Morbidity Survey, generally a large portion of 1,000,000 individuals in Malaysia detailed manifestations of depression (Jamal, 2021). The prevention methods carried out by the government to flatten the curve made the mental health problems in Malaysia even worse. The lockdown implementation made people lose jobs, students were told to stay at home and learning was done using online communication platforms, people were forced to stay at home and not leave the house. All these were the causes that made depression and mental health problems increase significantly (Sundarasen et al., 2020).

The increase in mental health problems shows that even though the government and related organizations are doing their best to help curb depression and mental health cases. This shows that the method that has been used to do the tracking and predicting is not fruitful. To make things worse, the covid-19 pandemic is one of the causes that contributed to this significant increase in mental health cases. To make things easier the sentiment analysis using Twitter as a data source to identify individuals with depression mental health symptoms at the initial stage will be an efficient way for the government and the related organization to manage and resolve the mental health problems which will gradually reduce the number of cases in Malaysia.

4. Aims and Objectives of the Research

This research expects to propose a viable and effective way to answer the problem of identifying and keeping track of depression moods of people using Twitter Data by focusing on the accompanying explicit objectives:

i . To develop a model for pinpointing depression indicative tweets using data mining techniques and sentiment analysis.
ii . To enable the extraction of expressions and comments related to emotions
iii . To evaluate the implementation of sentiment analysis using tweets assists in reducing depression among Malaysians.

5. Research Questions

i . How to recognize the element of depression in Tweets using data mining techniques and sentiment analysis?
ii . How to predict the emotions behind the expression and comments using data mining technique and sentiment analysis?
iii . How sentiment analysis using tweets assists in reducing depression cases among Malaysians?

6. Significance of the Research

The discoveries of the examination will contribute to the advantage to give a sense of comprehension with respect to depression magnitude in various users and the sentiment scores can be related to the fundamental information. The fruit of this research also will give an advantage for the government and mental health-related organization to identify the potential depressed users at an early stage which can significantly decrease the amount of self-harm, suicide and depression-related problems.

7. Methodology

Basically for this research qualitative methodology approach is going to be used as this research analyses twitter comments and tweets to find out depression indicative tweets. The overview of the methodology section can be categorized into 3 parts:

i) Data Extraction and Corpus building

Firstly, to get the data for this research, it has to be extracted from the Twitter database. To get potential depressed users and depression indicating tweets, depression indicating keywords as hashtags will be used to fetch a list of potential depressed users. Tags like #bullying, #depressed, #sad, #broken will be used to fetch the required data from Twitter Database using Twitter API. The data that has been extracted from the Twitter Database will focus on Malaysian Twitter users. Then the extracted data will be used to compile the tweets and comments of the same users and make a list of compiled users. For example, if 4000 tweets were collected from the Twitter database, then the 4000 tweets will be categorized according to the users. Once the compilation of users and their tweets is complete, the number of users will be recorded.

ii) Data Cleaning and Preprocessing

The data extracted from the Twitter Database will be as a JSON file as we will be using API to extract data from the Twitter Database. The JSON file format will not be supported as it will be harder in the later stages to perform the analysis. The JSON file will be converted to a CSV file. Then the special symbols and Unicode symbols are removed. Then the variables will be parsed as tweets and users and then will be saved in a CSV file format. Once the file conversion is done, data cleaning will take place. Unwanted symbols like exclamation marks, arrows, numbers and unwanted characters will be removed. If null values are present the null values are removed and each word in tweets will be separated using a comma. Then the words will be converted to lowercase characters.

ii) Depression weight analysis

To do the sentiment analysis on tweets to identify potential depressed users and depression indicating tweets 3 main things will be incorporated with this system. First, tweets will be evaluated using the Syuzhet package using r programming to get the base emotions of the tweet. Then the tweets with base emotions that are relevant to depression will be collected. The collected tweets then will be broken (tokenized) into bigrams and trigrams. Then the tokenized tweets will then be compared with the lexicon dictionary like SentiWordNwet and AFINN to get the frequency of the particular words that have been used and the sentiment score for the words. By this, we can identify potential depressed users and tweets.

8. Overview of Proposed System

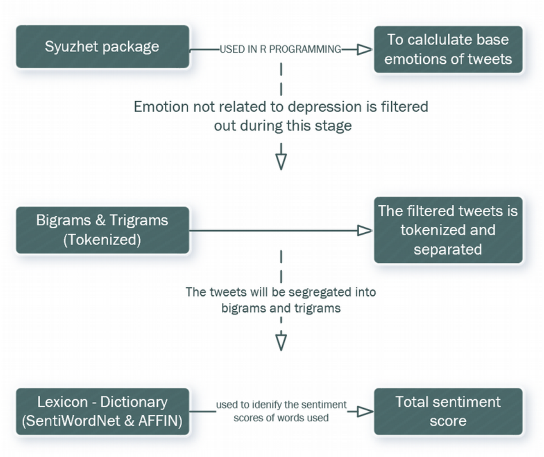

Syuzhet package
USED IN R PROGRAMMING
To calclulate base emotions of tweets
Emotion not related to depression is filtered out during this stage
Bigrams & Trigrams (Tokenized)
The filtered tweets is tokenized and separated
The tweets will be segregated into bigrams and trigrams
Lexicon - Dictionary (SentiWordNet & AFFIN)
used to idenify the sentiment scores of words used
Total sentiment score

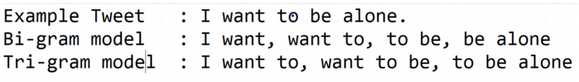
Figure 8.1 (Proposed System)

In this part of the proposal, the overview of the suggested lexicon-based sentiment analysis modal will be discussed. The above figure shows the overview of the lexicon-based sentiment analysis modal. According to figure 8.1, once the data pre-processing and data cleaning are done the CSV file consisting of tweets and the respective users are then uploaded in Rstudio to be evaluated using the Syuzhet package to determine the base emotion of the tweets. Using the Syuzhet package we can identify the 8 primary emotions which are anger, disgust, sadness, fear, anticipation, joy, surprise and trust within each tweet. The tweets base emotions that are relevant to depression is then stored and the others are filtered out.

	I want to be alone...	I'm glad that you came...	I was so pissed...
anger	0	0	1
sadness	1	0	0
disgust	0	0	1
fear	1	0	0
anticipation	0	1	0
joy	0	1	0
surprise	0	1	0
trust	0	1	0

Figure 8.2 (example output of Syuzhet package)

Figure 8.2 shows the example output of how the base emotion detection using the Syuzhet package will be. Once the base emotion is detected and the relevant tweets regarding depression are separated from the others. The depression indicative tweets will then broken (tokenized) using the Bigrams and Trigram word sequence. This process will help to identify words and phrases related to depression which then will be used to analyse. Bigrams are adjacent to 2 words sequence and trigrams are adjacent to 3 words sequence. With this, it would be easier to detect words and phrases relating to depression.



```
Example Tweet   : I want to be alone.
Bi-gram model   : I want, want to, to be, be alone
Tri-gram model  : I want to, want to be, to be alone
```

Figure 8.3 (example output of Bi-gram and Tri-gram model)

Figure 8.3 shows the example of a bi-gram and tri-gram model that will be implemented in the proposed system. With this Bi-gram and tri-gram model, evaluating the depression based words and phrases will be easier. Once the Bi-gram and Tri-gram models are built, the words and phrases obtained from the models are then compared with the SentiwordNet and AFFIN lexicons. SentiWordNet and AFFIN are lexicon databases that have been extracted and derived using the WordNet. SentiWordNet is related to mathematical scores showing the positive and negative value of the sentiment analysis while the AFFIN lexicon dictionary is associated with integer values between -5 till +5. The tweets are then compared using these 2 lexicons dictionaries. If the values are shown as negative or closer to negative those tweets are possible depression indicative. The Tweets that scored negative scores is that compiled with the respective user to calculate the total depression score for each user. Hence the people with depression can be identified using this method easily which can replace the actual manual process of consultation and survey to identify people with depression.

9. Conclusion

This research proposed a lexicon-based sentiment analysis using Twitter data as the data source to measure the sentiment of the users using the tweets sent by them on Twitter. By identifying the potential depressed user using Twitter, it can help them at the initial stage or help them at an early stage which can prevent the condition from being far worse. The proposed system can be implemented by the government or the mental health-related NGO’s to treat and monitor the patients or Malaysian population so that the mental health cases can be reduced in Malaysia.

References

Bernama. (2020). Almost 500,000 Msians depressed; nearly 500 suicide attempts this year. New Straight Times. Retrieved from https://www.nst.com.my/news/nation/2020/10/631154/almost-500000-msians-depressed-nearly-500-suicide-attempts-year

Cavazos-Rehg, P., Krauss, M., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., & Bierut, L. (2017). A content analysis of depression-related tweets. *Computers In Human Behavior*, *54*, 351-357. doi: 10.1016/j.chb.2015.08.023

Centers of Disease Control and Prevention. (2020). CDC - BRFSS - Survey Data & Documentation. Retrieved 9 March 2021, from https://www.cdc.gov/brfss/data_documentation/index.htm

Eusof Izzudin, M., Al–Bedri, A., Subramaniam, V., Matthews, P., & Ai Theng, C. (2018). Prevalence and Related Factors of Depression among Healthcare Personnel at Primary Healthcare Centers. *Malaysian Journal Of Medicine And Health Sciences*, *14*(SP2), 32-36.

Institute for Public Health 2020. National Health and Morbidity Survey. (2019) : Non-communicable diseases, healthcare demand, and health literacy—Key Findings

J. Paul, M., & Dredze, M. (2011). You Are What You Tweet: Analyzing Twitter for Public Health. *Proceedings Of The Fifth International AAAI Conference On Weblogs And Social Media*.

Jamal, U. (2021). COVID-19 exposes Malaysia's growing mental health challenges | ASEAN Today. Retrieved 10 March 2021, from https://www.aseantoday.com/2021/03/covid-19-exposes-malaysias-growing-mental-health-challenges/

Kawachi, I. (2001). Social Ties and Mental Health. *Journal Of Urban Health: Bulletin Of The New York Academy Of Medicine*, *78*(3), 458-467. doi: 10.1093/jurban/78.3.458

Low, D., Bentley, K., & Ghosh, S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96-116. doi: 10.1002/lio2.354

Mukhtar, F., & P. S. Oei, T. (2011). A Review on the Prevalence of Depression in Malaysia. *Current Psychiatry Reviews*, *7*(3), 234-238. doi: 10.2174/157340011797183201

Muthiah, W. (2016). Depression will be world No.1's disability by 2020, warns mental health group. *Thestar*. Retrieved from https://www.thestar.com.my/news/nation/2016/10/11/marked-increase-in-depression-itll-be-world-no1-disability-by-2020-warns-mental-health-group/

Rosenquist, J., Fowler, J., & Christakis, N. (2011). Social network determinants of depression. Molecular Psychiatry, 16(3), 273-281. doi: 10.1038/mp.2010.13

Sundarasen, S., Chinna, K., Kamaludin, K., Nurunnabi, M., Baloch, G., & Khoshaim, H. et al. (2020). Psychological Impact of COVID-19 and Lockdown among University Students in Malaysia: Implications and Policy Recommendations. *International Journal Of Environmental Research And Public Health*, *17*(17), 6206. doi: 10.3390/ijerph17176206

Ucanews. (2021). Malaysian archdiocese focuses on mental health during Lent - UCA News. Retrieved 9 March 2021, from https://www.ucanews.com/news/malaysian-archdiocese-focuses-on-mental-health-during-lent/91772

World Health Organization. (2018). Mental health: strengthening our response. Retrieved 10 March 2021, from https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

World Health Organization. (2020). Depression. Retrieved 2 March 2021, from https://www.who.int/news-room/fact-sheets/detail/depression

ZamZam, R., Thambu, M., Midin, M., Omar, K., & Kaur, P. (2009). Psychiatric morbidity among adult patients in a semi-urban primary care setting in Malaysia. Retrieved 1 March 2021, from https://pubmed.ncbi.nlm.nih.gov/19538711/

RMCT Marking Sheet Student Name: **Ruban Raj A/L Murugan** Student ID: **TP051462**

Criteria			**weight**	**First Marker** **(out of 100)**
C1	**Documentation (40%)**	Grammar, formatting, citation	5	
C2		Background, Justification & Scope of the Research	15	
C3		Problem Statement, Aim & Objectives	10	
C4		Research Methodology	5	/40
C5		Overview of System	5	= %
C6	**Presentation (10%)**	Presentation	4	
C7		Slides Quality	2	/10
C8		Questions and Answers	4	= %
Total Mark of this assignment *			**50**	

Comments:____

____

____

____