Early Prediction On Depression Based On Text Classification Method of Machine Learning

Norazlina Osman[1], Norshaliza Kamaruddin*[2], Ganthan Narayana Samy[3], Nurazean Maarop [4], Pritheega Magalingam[5], Fiza Abdul Rahim[6], Noor Hafizah Hassan [7]

Universiti Teknologi Malaysia
[1]norazlina @gmail.com, [2]norshaliza.k@utm.my, [3]ganthan.kl@utm.my, [4]nurazean.kl@utm.my, [5]pritheega.kl@utm.my, [6]fiza.abdulrahim@utm.my, [7]noorhafizah.kl@utm.my

Article history

Received:
1 Nov 2021

Received in revised form:
15 Nov 2021

Accepted:
1 Dec 2021

Published online:
21 Dec 2021

*Corresponding author
norshaliza.k@utm.my

Abstract

Mental illness includes emotional, psychological, and social well-being. Mental illness starts from a mild depression and it could lead to other serious types of mental illness. In 2017, a study estimates that 792 million people lived with mental illness which takes about 10.7% globally. Tools such as questionnaire is normally used by the clinician in diagnosing the early symptom of depression and there are various type of questionnaire that is used to predict various level of depression. In this study, the Diagnostic and Statistical Manual of Mental Disorders, version 5 (DSM-5) is used with several modification and were sent to 113 participants through the platform of media social. Findings show that, females suffered most from depression with 65% rather than males with only 35%. Besides, based on the age factors, adults that age between 36 to 45 years old suffered from mental illness more than other group of age with the percentage of 25%. Then, experiment also have been executed with three Machine Learning methods namely, Gauss Naïve Bayes, Random Forest and Decision Tree to observe the accuracy of these methods in doing the classification in predicting the depression into depress and not depress. The finding shows that Decision Tree shows the highest accuracy with 86.96% when compared to Gauss Naives Bayes and Random Forest with the accuracy of 82.6% and 78.26% respectively.

Keywords: *Mental Illness, Early Depression, Machine Learning Classification, Prediction on Depression, classification methods.*

1. Introduction

Mental illness or also known as mental disorder is an illness that could affect people's mood, thinking and also behavior. Based on research, mood disorders include anxiety disorder, bipolar disorder, postpartum disorder and many more. Major Depression which started with normal depression is also considered as mental illness [1]. Depression could affect a person time to time, but it may be severe that lead to Major Depression when ongoing signs and symptoms cause frequent stress and may affects the ability to functions [2].

Currently, mental illness is a global problem faced by the world. Some of the signs and symptoms of mental illness are feeling sad frequently, extreme feelings of guilt, extreme mood changes of highs and lows and many more. There are more than 264 million people, which is 3.37% of the world's population affected by

* *Corresponding author.* ***norshaliza.k****@utm.my*

depression, reported by WHO. In Malaysia, the prevalence of depression was estimated to be between 8 and 12% and the figures were higher among women of low socio-economic background [3].

Depression is a common illness faced by most people but depression may make a person feels miserable and can cause problems in the daily life. In most cases, symptoms can be managed with a combination of medications and talk therapy. Psychological treatments exist for moderate and severe depression but the cost for this treatment can be unaffordable for some peoples [3][4]. It is estimated that, 76–85% of people suffering from mental illness, lack access to the treatment they should get. People with depression should do a therapy which need them to meet someone through physical or even through telecommunication. Besides, medication which is considered expensive, changes in lifestyles could also help in reducing the daily depression [4][5]. This includes eating healthy, exercises, sleep well and others.

This paper explores and focuses on the early prediction of depression. The study involves the quantitative study by distributing a questionnaire to the respondents. The output gained from the respondents is studied and is categorized into positive and negative words. Those positive words will categorize the respondent as free from depression. Besides, outcome which are detected as negative words are categorized as respondents with depression. The next steps in the study, is to apply three Machine Learning which are Gaussian Naïve Bayes, Decision Tree (DT) and Random Forest (RF) to do the classification process whether the respondent are depressed or not depressed. Comparison among the machine learning methods is executed to see which method of Machine learning that contribute to the highest accuracy in dong the classification. In the next section, the problem background of the study is presented.

2. Literature Review

Previously, psychologist or psychiatrist predict a patient with mental illness using a standard questionnaire or tools, besides observing the patients' behavior by several session with the patient. Numerous tools currently, are available in accessing the early detection of depression. Some classic tools include the Zung Self-rating Depression Scale (SDS), Symptom Checklist (SCL-90) and Beck Depression Inventory (BDI), and many others [6]. However, not all of these tools can be used in predicting the early symptom of depression. Most of these tools are suitable in diagnosis the Major Depression Disorder (MDD).

On the other hand, Diagnostic and Statistical Manual of Mental Disorders (DSM–5) is an online questionnaire that is use to measures initial patient interview and to monitor treatment progress [6]. It serves to advance the use of initial symptomatic status and patient reported outcome information. In this paper, DSM-5 is being used with some modification to collect data in predicting the early symptom of depression. Besides, using tools to collect data of peoples regarding their mood disorder, this paper explores some machine learning methods that could be used to classify the input into depressed and not depressed.

Based on study, many researchers have worked on classifying depression with machine learning algorithms, such as Random Forest Tree (RFT), the Support Vector Machine (SVM) and the Convolution Neural Network (CNN). Work by Rida Zainab et.al (2020), applied the natural language processing and explainable artificial intelligence (AI) to analyze and rate depression related linguistic biomarkers. They investigate the English and Urdu text data that demonstrate the significance of semantic and cultural differences in language and individuals for the diagnosis of depression. Accuracy of depression diagnosis using Logistic Regression for Urdu text, BagofWords is 0.855. However, one limitation is the inability of the features to take word ordering into account.

In the study [7] have conducted experiments to identify text segments representing names, amounts and temporal data. They employ the extraction of information as a categorization issue while extracting information about natural catastrophes from newspaper stories in Spanish. By using a very limited training package, they achieved an average F-measure of 72 percent for the extraction task. The disadvantage is that extracting and relating data from documents with multiple intriguing cases can be difficult. This problem can be partially handled by using some level of linguistic analysis as a pre-processing step before using regular expression analysis. It makes use of a relatively limited training set. It is impossible to extract information that has been expressed implicitly. On the other hand, extracting and linking information from texts reporting several noteworthy events is difficult.

Study in [8] analyze the sentiment and characterize the sentiment expressed in blogs about specific brands and products. This Sentiment Analysis focuses automatically focuses on the identifying of whether a piece of text expresses a positive or negative opinion on the subject concerned. The major contributions of this study are developed effective framework for incorporating lexical knowledge in supervised learning for text categorization. Then, apply the approach to the task of sentiment classification extending the state-of-the-art in the field which has focused primarily on using either background knowledge or supervised learning in isolation. There are few kind of datasets, first is the Lotus blogs, a labelled sentiment related to the IBM Lotus software, political candidate blogs comes second which portray a continuous updated of 16,741 political blogs, containing over 2 million posts, third is the movie reviews consisting of 1000 positive and 1000 negative reviews from the Internet Movie Database [8][9]. Reviews that had rating above 3.5 starts are considered as positive labels, while the rest are considered as the negative labels in movie reviews. The accuracy for sentiment classification is compared using different approaches like Lexical Classifier, Feature Supervision, Naïve Bayes, Linear Pooling and Log Pooling. Statistically compared among all the methods, Linear Pooling performs the best as it improves holistically in terms of accuracy [11][12]. Accounting for the training and test sets being from different distribution is the most challenging. In the next section presents the methodology used in this paper.

3. Methodology

This paper focusses on the early prediction of depression using the Diagnostic and Statistical Manual of Mental Disorders (DSM) questionnaire. DSM is a handbook by health care in United State in diagnosing mental disorder. This handbook assists the clinician in diagnosing the mental illness which contains descriptions, symptoms and other criteria. DSM is first published in 1952 and until now the latest version is DSM-5. In this study, a questionnaire is design which is based on DSM-5. Data were collected from a total of 113 participants via Google forms and subsequently classified using three machine learning algorithms – namely Random Forest (RF), Gaussian Naïve Bayes (GNB) and Decision Tree (DT).

3.1. Participants

This study was conducted on a total of 113 participants, aged between 18 and 60 years, both males and females. The participants were also categorized into employed and unemployed and with a wide range of responsibilities from household chores to professional duties.

3.2. Questionnaires

The data for the study were collected through DSM-5, the Diagnostic and Statistical Manual of Mental Disorders (version 5) questionnaire. The questionnaire is modified and divided into 2 main section, demographic and mental disorders. The questionnaire is comprised into 24 questions with 10 questions allocated to the scales of Depression. The possible answers for each question – which could be given in text or numeric form – are as follows:

i. Never
ii. Occasionally
iii. Often

Among the three answers given above number 1 – 'never' is relate to a positive word which classified the respondent as no early symptom of depression, whereas answer number 3 depicted to the negative word that relate to the early detection of depression. The positive and negative words collected from this questionnaire is used to be processed and classified using three Machine Learning Methods.

The questions asked from individual are described in Table 1.

Table 1: Questions in the questionnaire to predict the early depression

No	Questions to respondents
1.	Think of hurting yourself
2.	Hear things other people couldn't hear

3.	Feel that someone could hear your thoughts
4.	Problems with sleep
5.	Problems with memory
6.	Problems finding a location that you wish to go
7.	Unpleasant thoughts repeatedly enter your mind
8.	Not sure about who you really are
9.	Know what is your life goals
10.	Feel not enjoying your relationships with other people

In overall, based on the category of positive and negative words, from 113 respondents that answered the questionnaire, it is found that most adult in the range age between 36 to 45 years suffered from anxiety and worried which contributes of 45% from the survey conducted. The next group that suffered from anxious and anxiety came from the group between the age of 45 to 56 years old, which is about 25%. This is shown in Figure 1 below which also shows in detail the distributions based on the symptom of depression on several category of age.

Based on the mood and feelings of the respondents that determine the early symptoms of depression, the distribution based on age factors shows several findings. For group between the age of 36 to 45 years old, this group is affected with the feeling of worried which took about 25%. About 15% from this group was affected with the feeling of anxious and followed by the feeling of nervous which contributes about 11%. The next group that shows some early symptom of depression is the group with the age between 45 to 56 years old. However, the teenager group with the age between 18 to 25 years old have less affected with the symptom of depression. Besides, the findings also show that, most females (65%) suffered from depression rather than males (35%). Figure 1 below.,

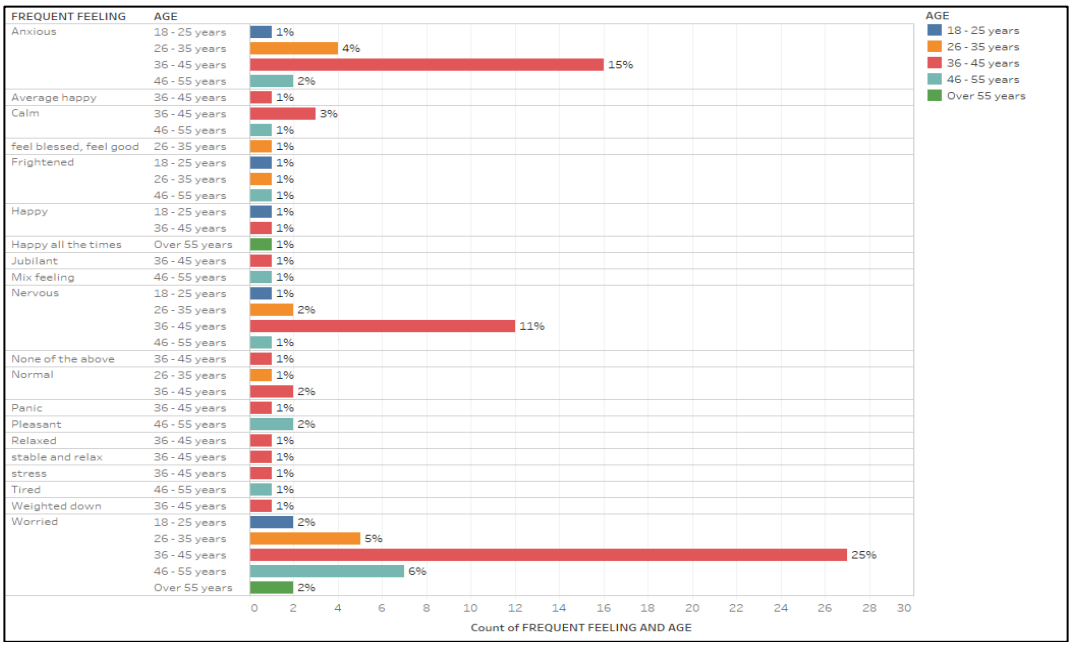

Figure 1: Horizontal bar chart distribution of participants (%) by frequent feeling and age

In the next section, experiment will be executed to observe which machine learning methods among the three methods chose which are Gauss Naïve Bayes, Random Forest and Decision Tree shows the most accuracy in doing the classification based on the positive and the negative words collected. It is collected that 91 negative words and only 22 positive words are determined from the survey.

In the next section, the paper will discuss on the finding of the study and the classification methods of Machine Learning apply in this study.

4. Findings

The results gained from the previous section which is the analysis made based on the questionnaire is used as an input for this section. This section describes the findings on observing the accuracy of the Three machine learning in performing the classification on the data obtained for the early prediction of depression. The results of the implementation and the performance evaluation of accuracy, precision, recall and F1-scores are summarized in Table 2. It is emphasized that the desired score of the performance evaluation is between 80% to 100% because the higher percentages of the accuracy, precision, recall and F1-scores, the better the algorithms in producing accurate classification output for early detection depression.

Table 2: Performance evaluation of the Supervised ML algorithms based on text classification

Algorithm		Precision	Recall	F1 scores
RF Accuracy = 0.782	Depression	0.78	1.00	0.88
	Not Depressed	0	0	0
GNB Accuracy = 0.826	Depression	0.85	0.94	0.89
	Not Depressed	0.67	0.40	0.50
DC Accuracy = 0.869	Depression	0.86	1.00	0.92
	Not Depressed	1.00	0.40	0.57

Table 3 Supervised ML algorithms of accuracy

Model	Accuracy
Random Forest Classifier	78.26 %
Gaussian Naïve Bayes Classifier	82.60 %
Decision Tree Classifier	86.96 %

Based on the findings, Decision Tree method shows the best Supervised ML algorithm which predicts accurately the depression status. Hence it is appropriate for the development of the early detection on depression model for psychiatric expert by applying classification methods. All the model give best range of accuracy from range 78 % to 87 %. The highest accuracy is Decision Tree Classifier which is 86.96%.

5. Conclusion

This section concludes the findings of the study. The first findings is to predict the depression on the factor of gender. Based on the experiments conducted using Phyton program, 65% of female gender shows the symptom of early depression to compare to male gender with 35%. Then the finding continues to look into the age factor which includes the findings of frequent feeling distribution of the participants in the dataset. It is explored by age and found that most affected is aged 36 – 45 years old who depressed which take around 25% followed by 11% by the second age of 45 to 56 years old. Prior to the implementation of the Supervised ML algorithms on the dataset, the results of the performance evaluation of the algorithms were 86.96% for Decision Tree. This shows better performance with accuracy more than 80%.

References

[1] Marcus, M. & Yasamy, Mohammad Taghi & Ommeren, M. & Chisholm, D. & Saxena, (2012). Depression: A global public health concern. World Health Organization Paper on Depression. 6-8.

[2] Akshay Bhavani Kumar Kulkarni. (2018) 'Early Detection of Depression', Faculty of the Department of Computer Science, University of Houston.

[3] Ng CG (2014). A Review of Depression Research in Malaysia. Med J Malaysia. 69 Suppl A:42-5. PMID: 25417950.

[4] Ellen W Freeman (2003) 'Premenstrual syndrome and premenstrual dysphoric disorder: definition and diagnosis', Elsevier

[5] Laura J.Miller, (2015) 'Postpartum Depression', Clinicians Corner

[6] Anu Priyaa, Shruti Garga, Neha Prerna Tiggaa, (2020), Predicting Anxiety, Depression and Stress in Modern Life Using Machine Learning Algorithm, International Conference on Computational Intelligence and Data Science, Science Direct.

[7] Alberto Tellez-Valero, Manuel Montes-y-Gomez and Luis Villasenor-Pineda. (2005), 'A Machine Learning Approach to Information Extraction', Conference Paper in Lecture Notes in Computer Science .

[8] Prem Melville, Wojciech Gryc, and Richard Lawrence, (2009), Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification, In Proceedings of the 15th Conference on Knowledge Discovery and Data Mining (KDD-09), Paris, France.

[9] Amritha S Nadarajan and Thamizharasi A. (2018) 'A Survey on Text Detection in Natural Images', IJEDR Volume 6, Issue 1 ISSN: 2321-9939

[10] A Hernandez-Castaneda and H.Calvo. (2017) 'Deceptive text detection using continuous semantic space models', Intell. Data Anal., vol. 21, no. 3, pp. 679-695.

[11] Bahzad Taha Jijo and Adnan Mohsin Abdulazeez. (2021) 'Classification Based on Decision Tree Algorithm for Machine Learning', Vol. 02, No. 01, ISSN:2708-0757, pp 20-28

[12] Elisabeth Schramm, Daniel N Klein, Moritz Elsaesser, Toshi A Furukawa and Katharina Domschke (2020) 'Review of dysthymia and persistent depressive disorder:history, correlates, and clinical implications', Lancet Psychiatry Vol 7 September 2020