Current Applied Science and Technology Vol. 23 No. 1 (January-February 2023)

Research article

Weighted Voting Ensemble for Depressive Disorder Analysis with Multi-objective Optimization

Wongpanya Nuankaew[1], Pratya Nuankaew[2], Damrongdet Doenribram[3] and Chatklaw Jareanpon[3]*

[1]*Polar Lab, Department of Information Technology, Faculty of Informatics, Mahasarakham University, Mahasarakham, Thailand*
[2]*School of Information and Communication Technology, University of Phayao, Phayao, Thailand*
[3]*Polar Lab, Department of Computer Science, Faculty of Informatics, Mahasarakham University, Mahasarakham, Thailand*

Received: 9 December 2021, Revised: 17 March 2022, Accepted: 18 July 2022

DOI: 10.55003/cast.2022.01.23.015

Abstract

The Twitter platform is a popular tool that is widely used by researchers to collect data on users' personal lives, feelings and emotions. These data sets can be further analyzed using text mining techniques to predict the disorder of depression. There are nine symptoms of depression that are classified by American Psychiatric Association using DSM-5 criteria. The symptoms can be difficult to identify effectively. The unweighted vote ensemble is not practical for multi-class data. Therefore, this research proposes the multi-objective optimization algorithms for depressive symptom prediction modeling (MOADSP) for the weighted voting ensemble, which can improve its effectiveness compared to the singer model. The objectives of this research were 1) to find the appropriate number of features; 2) to improve the weights of the prediction models based on the recall of the class for the ensemble; and 3) to compare the performance of the single, unweighted, and weighted voting ensemble models for depressive disorder. An information gain was used to select the features. The single classification techniques used in the experiment that had their frameworks tested were the Naïve Bayes, Random Forest, and K-Nearest techniques, while the vote ensemble models used were the unweighted and weighted models. MOADSP was applied to the weighted vote ensemble models. The results showed that the best recall classifier was KNN (98.60%), and the highest recall classifier was AVG TP weighted (98.43%) for the training model. The highest recall in the class depressive classifier was AVG TP weighted (80.00%) for the testing. This proposed method was beneficial for the prediction of depressive disorder.

Keywords

weighted voting ensemble;
multi-objective optimization;
major depressive disorder;
text classification;
depressive disorder analysis

*Corresponding author: Tel.: (+66) 985951653
E-mail: chatklaw.j@msu.ac.th

1. Introduction

Major depressive disorder (MDD) is a mental disorder characterized by people experiencing at least two continuous weeks of low mood, low self-esteem, dejection, sadness, desperation, and loss of interest in their activities [1]. An unbalanced level of neurotransmitters causes these symptoms of MDD to respond to cognitions, feelings, moods, behaviors, and depression. Like any other disorders, MDD can occur in people of all ages and genders. Risk factors for depression include stress, social rejection, adverse mental state, trauma, and specific behaviors. The symptoms vary among patients and can sometimes be severe as the patients may have low self-esteem, loss interest, insomnia or excessive sleepiness, unintentional weight loss or weight gain, fatigue, hallucinations, delusions, and even suicidal tendencies [2].

The World Health Organization (WHO) reported that 280 million people suffered from depression in 2021 [3]. Ukraine (6.3%), the United States (5.9%), Estonia (5.9%), Australia (5.9%), and Brazil (5.8%) were the countries with the most significant number of depression cases [4]. Moreover, depression can lead to suicide. More than 700,000 people per year committed suicide, and there were many more who failed in their attempt to commit suicide. Suicide was the fourth leading cause of death for young people aged 15-29 years [3] (after road accidents, tuberculosis, and interpersonal violence [5]). The cause of depression may be biological and social factors, such as lacking of proper life management, increasing emotional complexity, and changes in culture, values, or lifestyle. A significant number of adolescents need social support from friends and family to relieve stress and depression. Still, they are frequently perceived in a negative way by others around them. Often, appropriate treatments are needed to help them deal with the effects of social media [6]. People between the ages of 19-32 years with their frequent use of social media are highly susceptible to depression because most of the posts on social media display happiness and perception of a better lifestyle [6]. These posts can inadvertently make some people feel disappointment and resentment towards themselves. As of 2020, there were 396 million social media accounts worldwide [7]. The formats that appeared on social media include texts, images, animations, videos, emoticons, and so on. Since there is a massive amount of social media data generated by millions of users at a high rate, the data can be utilized can utilize for analysis to help and resolve the problems of depression.

Data mining, or the process of extracting significant data sets for relationships and patterns need to be understood and is precious to the data collector [8]. Nowadays, social media messages are analyzed via techniques such as text classification to find the relationships and patterns, the aim frequently being to use techniques to study human behavior [9]. Several researchers in psychology have extracted behavioral data from the social media platforms such as Twitter, Sina Weibo, Tumblr, Reddit in order to analyze the psychological states of social media users [10]. Twitter data were obtain to analyze the four types of mental disorders: 1) emotional disorder, 2) trauma, 3) seasonal depression, and 4) major depressive disorder [11], to identify and study the types of depression symptoms, to detect the depression from features associated with depression [12], and to classify the depression [13]. User behavior from Reddit data was used to identify depression [14] and to detect depression evident in the users [15]. Onan *et al.* [16] researched with social media data and text classification consisted of experiments involving message filtering, stop word elimination, feature selection (FS), categorization, and efficiency measuring. They applied an appropriate method to transform raw data to information of interest that combined data mining, ensemble methods, and weight adjustment for prediction outputs through multi-objective optimization (MOO). Another method was the combination of ensemble pruning based on clustering and randomized search and multi-objective evolutionary algorithm to categorize messages. Zul *et al.* [17] specified the number of labels needed to analyze confidence level on Facebook and Twitter, and Doenribram *et al*. [13] categorized depression by selecting features via the information gain (IG) method with Naive Bayes classification and vote ensemble.

Furthermore, several researchers proposed data mining techniques to diagnose mental disorders [18], disease classification [19], and other areas that emphasized the FS method. Moreover, there were several methods used for reducing the dimension of data, optimizing the processing time, and increasing the performance model, such as the utilization of FS in the prediction of depression in the elderly [20], analysis of FS when identifying a depression model [2], and the reduction of the size of features of the prediction model of depression among adolescents [21].

A weighted voting ensemble is an ensemble machine learning algorithm that improves the performance for the classification models. The unstable problem of the unweighted called hard voting ensemble came from the number of weak classifiers having more than the robust classifiers, making the wrong vote and wrong answer [22]. Thus, the difficult task of the weighted voting ensemble and multi-objective optimization problem (MOOP) is to find the method, calculate, and determine the best-performing method from the various classifiers.

The depressive disorder classification models have two problems, which are 1) finding the appropriate number of features by obtaining the accuracy and time effectiveness, and 2) determining the appropriate weight for the weighted voting ensemble.

This research proposes 1) to find the appropriate number of features, 2) to improve the weights for the prediction models performance based on the recall of the class for the weighted voting ensemble model, and 3) to compare three methods: a single model, an unweighted model, and weighted voting models. In this research, data were collected from Twitter posts and comments from Doenribram *et al*. [13], classified the label of the data, and utilized the symptoms of depression via the DSM-5 questionnaire [23].

2. Materials and Methods

The materials and methods studied are divided into five categories: related work, information gain, classification algorithms, multi-objective optimization, and methodology. The details are as follows:

2.1 Related work

Nowadays, text classification has become a helpful tool for mental health applications such as depression. In particular, the analysis of user behavior in social networks involves correlation of key parameters and individual psychological traits. This research proposes a body of work related to the major depressive disorder classification (MDDC).

Twitter data was used to classify the severity of major depression disorder by considering the attributes and frequency of key words [24]. Kang *et al*. [25] analyzed comments, emoticons, and images to identify people susceptible to depression and then achieved accurate predictions through the weights assigned to various emoticons. Doenribram *et al*. [13] proposed the classification of depression from nine symptoms designated by the American Psychiatric Association using hashtags to assess the efficiency of the model and processing time. In that research, FS was performed using IG with various numbers of the features, 2000, 4000, 6000, and all features. The Naive Bayes technique was also used to classify symptoms with the probability boundaries specified as 0, 10, 20, 30, 40, 50, 60, 70, 80, and 90 for filtering the message, then the vote ensemble utilized for the output predictions. Burdisso *et al*. [26] presented a text classification method called SS3 for depression detection over the social media streams. Alabdulkreem [27] studied posts from women during the COVID-19 pandemic, specifically in 50 or more posts, and compared the classification efficiency of traditional and deep learning. Finally, Aldarwish and Ahmad [28] presented a classification of the severity of depression from social media posts to aid in diagnosis and treatment that was appropriate to the severity level. A method to correct imbalance in the dataset by randomly selecting an equal number of relevant and irrelevant posts, and by using the Porter stemming algorithm in shortening the word length were also recommended.

From Doenribram *et al*. [13], citing the American Psychiatric Association, Table 1 shows an analysis of DSM-5 questionnaires for respondent who had a history of symptoms lasting for 10-14 days. The total sums of scores for all nine symptoms were used for the assessment of depression. They were divided into four levels, 1) 0-8 non-depression or minimal depression, 2) 9-14 mild, 3) 15-19 moderate, and 4) 20-27 server level. These questionnaires were one of the standard methods used to estimate the level of depression of the respondents.

Table 1. DSM-5 criteria from American Psychiatric Association

No	Symptoms	Non	Someday	Frequently	Symptoms
1	Depressed mood	0	1	2	3
2	Diminished interest	0	1	2	3
3	Change in appetite	0	1	2	3
4	Change in sleep	0	1	2	3
5	Slowed thinking	0	1	2	3
6	Worthlessness or guilt	0	1	2	3
7	Fatigue	0	1	2	3
8	Agitation or retardation	0	1	2	3
9	Suicidal ideation	0	1	2	3

2.2 Information gain

An information gain (IG) method in FS reduces the size of the original data without loss of essential data characteristics and can help to make a fast and efficient model. IG is a filter method that is fast, accurate and easy to interpret [29]. Among the filter methods, IG has often been seen to give better results than other filter methods [30]. It is widely used to reduce the number of features for text classification [31], and many researchers have applied it in their research [13, 30, 32].

IG considers the probability of each possible feature, then measures the entropies to select the important attributes or divide the data into subsets by calculating the gain value. If the gain value of the highest dimension is chosen, it will identify the tuple in set *D*. Entropy is the measurement of the impurity of the data. The information gain based on the information theory is calculated, computed the difference between entropy before and after the splitting process, and specified the impurity in the class elements from equations (1) (2) and (3). This can partition the classification [33].

$$\mathrm{Info(D)}= -\textstyle\sum_{\mathrm{i=1}}^{\mathrm{m}} P_i \log_2 P_i \tag{1}$$

$$\mathrm{Info_A(D)} = \sum_{\mathrm{j=1}}^{\mathrm{v}} \frac{|D_j|}{|D|} \times \mathrm{Info}_{D_j} \tag{2}$$

$$\mathrm{Gain(A)} = \mathrm{Info}(D) - \mathrm{Info_A}(D) \tag{3}$$

where P_i is the probability of record *i*, m is the number of members in set data *D*
D_j is the member of set data *D*.

2.3 Classification algorithms

Classification problems can be solved by various algorithms, but this research focuses on multi-class classification algorithms because it deals with the nine symptoms specified by the American Psychiatric Association.

Naive Bayes (NB) is a simple supervised learning method for classification by calculating the probability to infer the solution. A conditional probability model is used as the data training model [8]. It is appropriate for classifying a large sample set with independent features and is famous for text classification [34] as calculated by equation (4).

$$\mathrm{P(\,A\mid B\,)}=\frac{P(B|A)\ P(A)}{P(B)} \tag{4}$$

where P(A|B) is the probability of *A*, given *B*, denoting the incidence of *B* happening on condition that *A* happens. P(A), P(B) is the likelihood of event *A* or *B*.

K-nearest neighbors (KNN) is a simple classifying method that involves scrutinizing the closest distance between the data and the learning set, where k is the number of the nearest neighbors. K should be an odd number to avoid the same number of label neighbors problems [35]. The KNN method compares all the features between the unknown and the training dataset then considers the rows closest to the classified rows measured by a distance function such as Euclidean Manhattan and Minkowski. The simple Euclidean formula calculates the nearest distance as shown in the equation (5).

$$\mathrm{d}=\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}}\left(p_i-q_i\right)^2} \tag{5}$$

where *p* and *q* are two n-space coordinate points, *d* is the distance between p_i and q_i, and *i* is the number of set members.

Random forest (RF) is a technique of ensemble learning. The training data and distinct features are randomly selected into multiple sets. Then the models are constructed using a collection of decision trees [36], where the out-of-bag data are collected into the data test for the prediction. Finally, the model results are brought into voting, and the result with the most votes is the solution. Random forest has a better prediction efficiency than class tree classifiers.

Vote ensemble is a classification method of voting that is very simple, efficient, and prevalent in ensemble learning. This method applies the same training data to the models from various techniques. Similarly, the same set of test data is then applied to the models. Finally, the prediction results from the models are voted, and the one with the most votes is taken as output. Generally, there are two types of voting schemes: unweighted and weighted voting schemes. The unweighted voting scheme is simple and widely used, and all classifiers get the same weight [37]. A disadvantage of the unweighted vote technique is that it is not always possible to predict well because sometimes predictions may not find plurality voting. If weak classifiers are voted, the model's performance is affected. The weighted voting method can solve this problem. It chooses the label by the most weighted value instead of the majority votes. There are a number of variations of weighted voting and included are simple weighted voting, rescaled weighted voting, best-worst weighted voting, and quadratic best-worst weighted voting [22].

2.4 Multi-objective optimization

Optimization is a method used to find the best solution for a problem under specific conditions, and it may involve single-objective optimization problem (SOP) or MOOP. The difficulty of finding a solution depends on various factors such as the nature of the problem, search space, local optima, and the size of the search space. The SOP method finds out the output to serve only one objective and the single best output. Therefore, it is not suitable for multi-objective situations in which there is more than one conflicting objective function, and a suitable solution is required to assign the weights to each output for a different class. The determining solutions to the problem include the weighted sum approach and the Pareto-based approach [22]. This research focuses on the weighted sum approach and MOOP. The weights were obtained by the correct prediction rate of the classes from the true positive values. The basic classifier techniques [38] are combined as shown in equations (6) (7) and (8).

$$\text{TP-rate}=\frac{\textit{True positive class } n}{\textit{Total member of class } n} \tag{6}$$

The formula calculates the new weights (TP-weight class) as shown in the following equations.

$$\text{TP-weight class} = \textit{Probability x TP-rate} \tag{7}$$

$$\text{AVG TP-weight class} = \text{weight max}\,\frac{\sum_{n=c}^{class} \left(P\ x\ (T_r)\right)_w}{C} \tag{8}$$

where n is the number of the class, P is the probability of each member of class n, T_r is the weight of class from the prediction (TP-rate), w refers to the total number of instances, and C is the number of basic models used to create learning models.

This research uses the theory of information gain to feature selection, Naive Bayes, K-nearest neighbors, Random forest and compares the performance of the unweighted and weighted voting of the MOOP method using equations (6)(7) and (8) of the models.

2.5 Methodology

Text mining is currently a popular method of finding and extracting hidden information in a large amount of data. It is often used in text classification, and includes data selection, data preparation, model construction, and evaluation [9]. This research proposed a multi-objective optimization algorithm for depressive symptoms prediction modeling (MOADSP) to solve the problems of the unweighted vote method, which was then deployed to predict the depression from Twitter data. The MOADSP calculates the maximum average weight of the predicted results from the probability of each record and the calculated recall of each class. This research methodology is divided into four phases, i.e. 1) data collection, 2) feature selection, 3) MOADSP, and 4) evaluation, as shown in Figure 1.

2.5.1 Data collection

The data collection consisted of the following dataset from the depression dataset from Twitter [13]. The data set was divided into two parts, which were the training data and testing data. The training data was selected from hashtags that indicated the nine symptoms of depression and normal behavior, as shown in Table 2. It contained 30,000 messages indicative of 10 symptoms of depression: 1) depressed mood, 2) diminished interest, 3) changes in appetite, 4) changes in sleep, 5) slowed thinking, 6) feelings of worthlessness or guilt, 7) fatigue, 8) agitation or retardation, 9) suicidal ideation, and 10) normality. The testing data comes from the post of 30 persons who are celebrity patients: 15 persons with depression (34,435 messages) and 15 persons with non-depression samples (22,498 messages).

Table 2. The training data selected from the hashtags

No	Symptoms	Hashtag	Number of messages
1	Depressed mood	#sadness #depressive	3,000
2	Diminished interest	#loss of interest #lose interest	3,000
3	Change in appetite	#appetite #hunger	3,000
4	Change in sleep	#sleepless #lethargy	3,000
5	Slowed thinking	#unthinking #out thinking	3,000
6	Worthlessness or guilt	#guilt #disgrace #dishonor	3,000
7	Fatigue	#tired #bored #fatigued	3,000
8	Agitation or retardation	#lackadaisical #lazy #loafing #phlegmatic	3,000
9	Suicidal ideation	#suicidal #dangerous #destructive	3,000
10	Normal	#happy	3,000
	Total		**30,000**

2.5.2 Feature selection

The feature selection process of the training data was divided into two sections, and the details are as follows:

The text preprocessing section managed the training data for modeling. The information obtained from Twitter had to be modified or have information removed. Unnecessary data and insignificant words such as the stop words, links, punctuation, symbols, emails, retweets, and usernames were removed. This section was divided into five steps: Step I: Regular expression management, Step II: Word segmentation management, Step III: Transformation management, Step IV: Filter stop words, and Step V: Binary term weighting.

Step I: Filtering the irrelevant data, such as retweeted messages, links, and usernames.

Step II: Dividing the words from each sentence and constructing the bags of words (BOW) to compute the frequency of each word.

Step III: Transforming the words in the bags of words to the lowercase.

Step IV: Filtering out the stop words such as any, before, cannot, into, most, shan't, which, and yourself.

Step V: Performing the weighting designates for each word in the bag of words using the binary term occurrence.

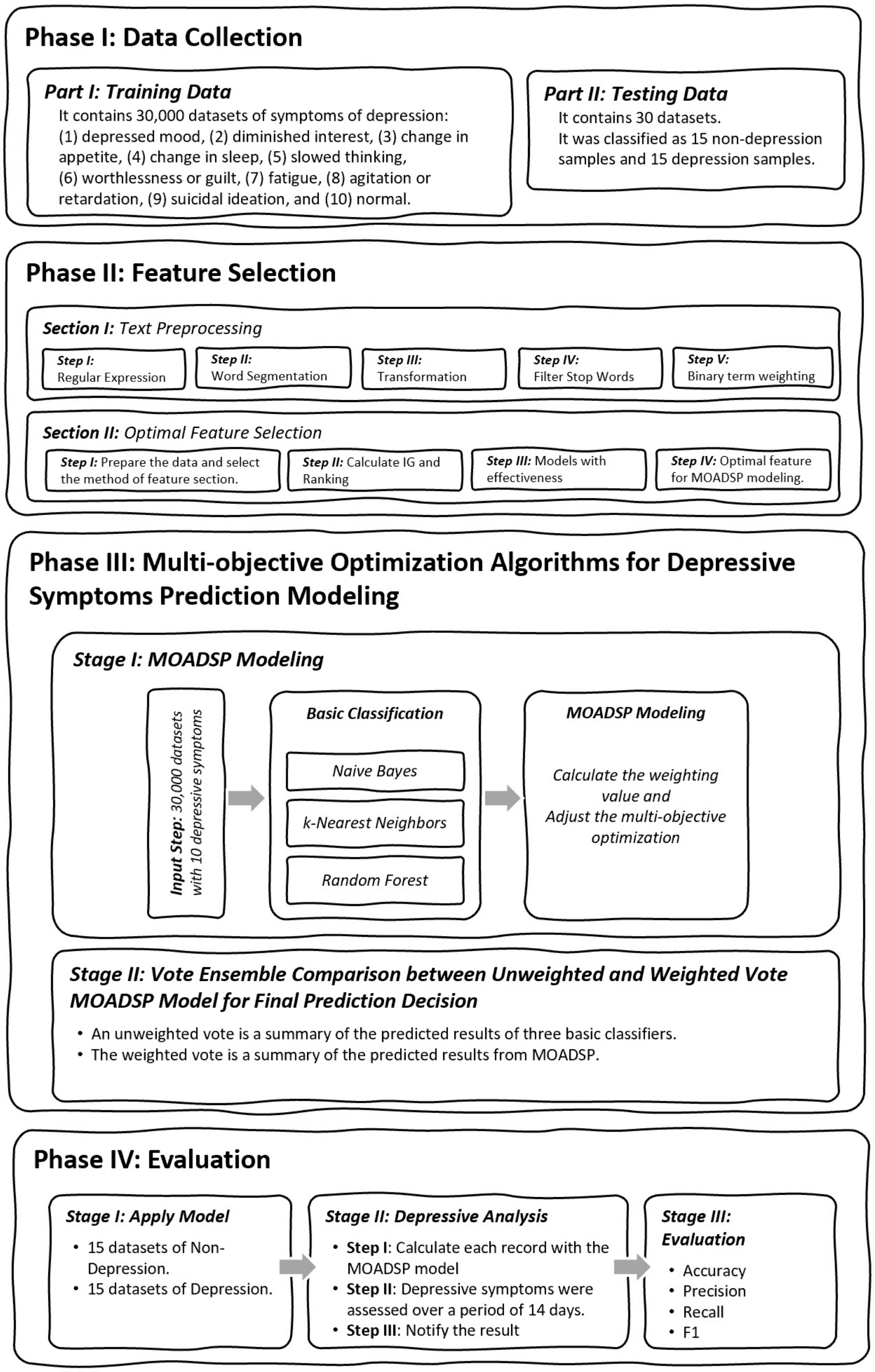

Phase I: Data Collection
Part I: Training Data
It contains 30,000 datasets of symptoms of depression: (1) depressed mood, (2) diminished interest, (3) change in appetite, (4) change in sleep, (5) slowed thinking, (6) worthlessness or guilt, (7) fatigue, (8) agitation or retardation, (9) suicidal ideation, and (10) normal.
Part II: Testing Data
It contains 30 datasets.
It was classified as 15 non-depression samples and 15 depression samples.
Phase II: Feature Selection
Section I: Text Preprocessing
Step I: Regular Expression
Step II: Word Segmentation
Step III: Transformation
Step IV: Filter Stop Words
Step V: Binary term weighting
Section II: Optimal Feature Selection
Step I: Prepare the data and select the method of feature section.
Step II: Calculate IG and Ranking
Step III: Models with effectiveness
Step IV: Optimal feature for MOADSP modeling.
Phase III: Multi-objective Optimization Algorithms for Depressive Symptoms Prediction Modeling
Stage I: MOADSP Modeling
Input Step: 30,000 datasets with 10 depressive symptoms
Basic Classification
Naive Bayes
k-Nearest Neighbors
Random Forest
MOADSP Modeling
Calculate the weighting value and Adjust the multi-objective optimization
Stage II: Vote Ensemble Comparison between Unweighted and Weighted Vote MOADSP Model for Final Prediction Decision
• An unweighted vote is a summary of the predicted results of three basic classifiers.
• The weighted vote is a summary of the predicted results from MOADSP.
Phase IV: Evaluation
Stage I: Apply Model
• 15 datasets of Non-Depression.
• 15 datasets of Depression.
Stage II: Depressive Analysis
• Step I: Calculate each record with the MOADSP model
• Step II: Depressive symptoms were assessed over a period of 14 days.
• Step III: Notify the result
Stage III: Evaluation
• Accuracy
• Precision
• Recall
• F1

Figure 1. The research framework

Optimal feature selection performs an efficient model, which can be divided into four steps, i.e. Step I: Preparing the data and selecting the method of feature section, Step II: Calculating IG and ranking, Step III: Finding the model effectiveness, and Step IV: Applying the optimal feature for MOADSP modeling as follows.

Step I: Preparing and performing the feature selection using IG by finding the appropriate k constant, the resultant number of features enabling the most efficient model and fast processing time.

Step II: Calculating and ranking the features using top k from the follow values: 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 4000, and 6000 to find the number of features increasing the accuracy in balance with the processing time.

Step III: Finding the model with the effectiveness created from the NB, RF, and KNN classification techniques. The construction of the model utilizes the 10-fold Cross-validation method, which constitutes the training data of 90% and the testing data of 10%. The results have provided the accuracy and time effectiveness, as shown in Figure 2 and Table 3.

Step IV: The best way to balance models from step III is applied to MOADSP modeling. Figure 2 shows the 1,300 features used to model the accuracy and the processing time. The recall of the 1,300 features modeling (shown in Table 6) was applied in phase III.

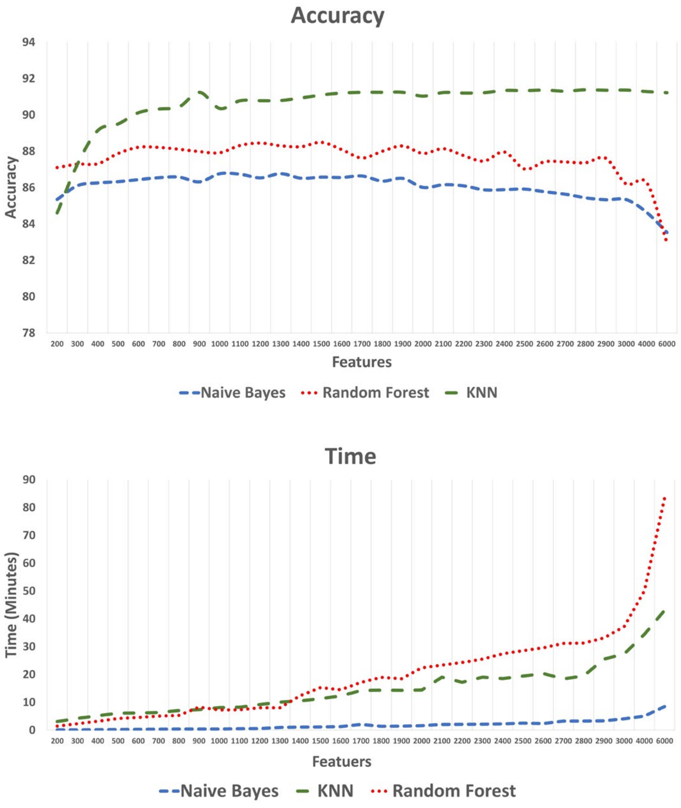


Figure 2. The accuracy and time consuming represented from the three primary classifications with various numbers of features

Table 3. The results of the accuracy and time

Features	Accuracy			Time (minute)		
	NB	RF	KNN	NB	RF	KNN
200	85.35	87.10	84.61	0.10	1.44	3.12
300	86.11	87.29	87.29	0.13	2.34	4.26
400	86.26	87.31	89.14	0.18	3.18	5.19
500	86.33	87.88	89.50	0.24	4.20	6.09
600	86.44	88.22	90.12	0.29	4.58	6.15
700	86.55	88.21	90.32	0.34	5.07	6.32
800	86.58	88.11	90.44	0.37	5.28	7.06
900	86.32	87.98	91.24	0.38	8.27	7.38
1000	86.77	87.92	90.35	0.40	7.27	8.10
1100	86.74	88.32	90.78	0.51	7.35	8.28
1200	86.54	88.45	90.78	0.55	8.02	9.18
1300	86.77	88.30	90.79	1.01	8.03	10.05
1400	86.52	88.25	90.93	1.09	12.38	10.52
1500	86.58	88.49	91.09	1.18	15.32	11.32
1600	86.56	88.09	91.20	1.25	14.51	12.25
1700	86.64	87.63	91.24	2.03	17.19	14.28
1800	86.36	88.00	91.24	1.40	19.00	14.38
1900	86.51	88.29	91.24	1.45	18.51	14.33
2000	86.01	87.88	91.03	1.58	22.37	14.42
2100	86.16	88.14	91.23	2.06	23.37	19.02
2200	86.10	87.76	91.20	2.11	24.34	17.22
2300	85.88	87.46	91.22	2.17	25.54	19.01
2400	85.89	87.96	91.34	2.24	27.44	18.52
2500	85.92	87.02	91.33	2.52	28.53	19.42
2600	85.77	87.43	91.36	2.38	29.60	20.36
2700	85.64	87.41	91.30	3.25	31.19	18.41
2800	85.44	87.37	91.38	3.24	31.30	19.45
2900	85.33	87.61	91.35	3.34	33.12	25.51
3000	85.34	86.19	91.36	4.07	37.22	27.42
4000	84.64	86.28	91.28	5.09	50.06	34.43
6000	83.53	82.97	91.22	8.48	83.40	43.08

2.5.3 Multi-objective optimization algorithms for depressive symptoms prediction (MOADSP)

The MOADSP modeling stage was divided into two steps: 1) Basic classification uses of the outputs of the model from phase II of the training data, and 2) Calculation of the weighting values and adjustment of the multi-objective optimization (MOO); the details are as follows:

Calculating the weighting values that were calculated from the number true positives (TP) dividing the number of members in each class (6) called TPrate. The probability multiplied by TPrate of the class n was called the TP weight (7), and the average maximum of TP weight of class n was called the AVG TP weight (8).

Adjusting the multi-objective optimization that the MOADSP chose for the AVG TP weight of each class n for final prediction.

Calculating the weighting values algorithm and adjusting the MOO steps to compute as shown in Table 4.

The vote ensemble compares unweighted and weighted votes of the MOADSP model for the prediction decision stage are unweighted, TP weight, and AVG TP weight by the testing data. An example of the principle of calculating the results of each model is in Figures 3-5.

Table 4. The calculation of the TP weight and AVG TP weight value algorithm

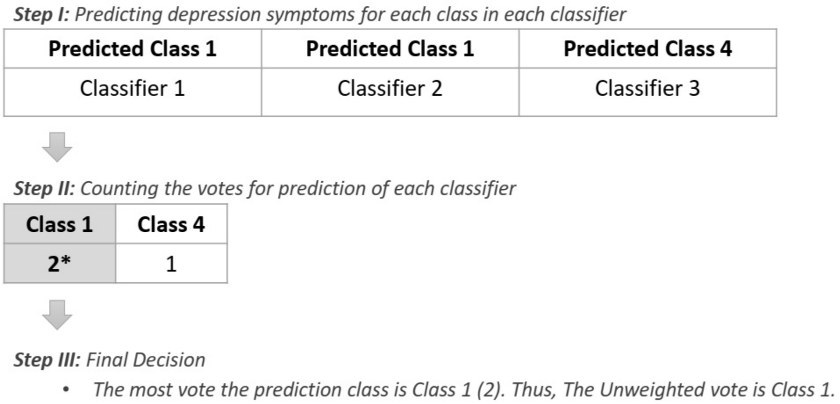
Algorithm calculation the weighted vote ensemble

```
1.  Calculating $\text{TPrate}_n$ of each classifier $c$
        TPrate = $(\textit{true positive class n}) / (\textit{total member of class n})$ where $n$ is class

2.  Calculating $\text{TPweight}_{(m,n)} = (\textit{Probability x TPrate})_n$ where $m$ is member and $n$ is class
3.  Assigning the weight of each classifier c the $\text{NewTPweight}_{(m,n)}$ = TPweight
4.  Calculating AVGTPweight of each classifier $c$
        $\text{AVGTPweight}_{(m,n)}$= sum $(\textit{TPweight}_{(m,n)})$ / $(\textit{number of classifier c})$
        where $m$ is member, $n$ is class, and $c$ is the number of classifiers

5.  Assigning the weight of each classifier c the $\text{NewAVGTPweight}_{(m,n)}$ = AVGTPweight
6.  Adjusting the MOO for the final prediction of the TP weight voting ensemble
    6.1  $\text{PreTP}_n$= max($\text{NewTPweight}_{(m,n)}$)
    6.2  Assigning the final prediction class of $\text{PreTP}_n$ = class $n$

7.  Adjusting the MOO the final prediction for the AVGTP weight voting ensemble
     MOADSP
    7.1  $\text{PreAVG}_n$= max($\text{NewAVGTPweight}_{(m,n)}$)
    7.2  Assigning the final prediction class of $\text{PreAVG}_n$ = class $n$
```



Step I: *Predicting depression symptoms for each class in each classifier*

Predicted Class 1	Predicted Class 1	Predicted Class 4
Classifier 1	Classifier 2	Classifier 3

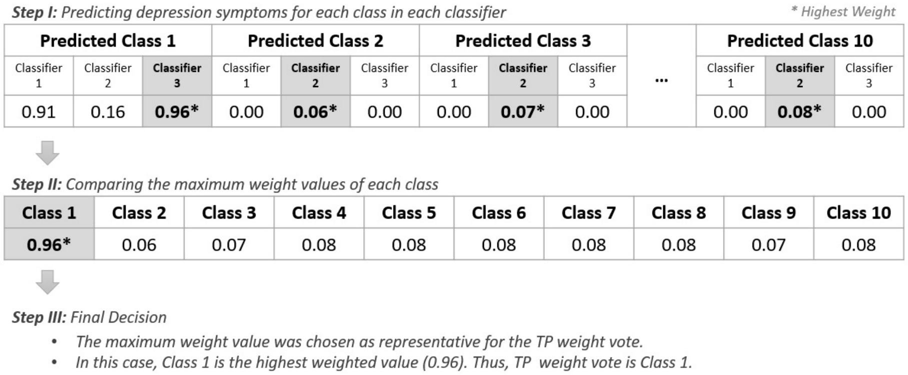
Step II: *Counting the votes for prediction of each classifier*

Class 1	Class 4
2*	1

Step III: *Final Decision*

- *The most vote the prediction class is Class 1 (2). Thus, The Unweighted vote is Class 1.*

Figure 3. Unweighted vote algorithm



Step I: *Predicting depression symptoms for each class in each classifier* ** Highest Weight*

Predicted Class 1			Predicted Class 2			Predicted Class 3			...	Predicted Class 10		
Classifier 1	Classifier 2	**Classifier 3**	Classifier 1	**Classifier 2**	Classifier 3	Classifier 1	**Classifier 2**	Classifier 3		Classifier 1	**Classifier 2**	Classifier 3
0.91	0.16	**0.96***	0.00	**0.06***	0.00	0.00	**0.07***	0.00		0.00	**0.08***	0.00

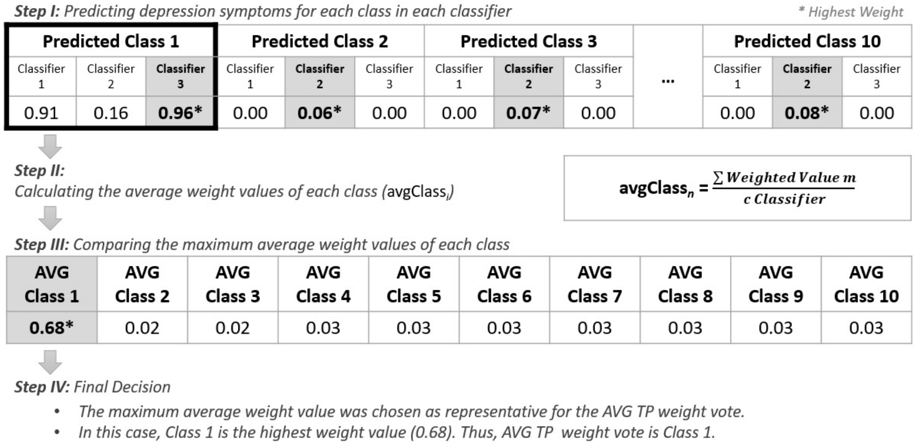
Step II: *Comparing the maximum weight values of each class*

Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9	Class 10
0.96*	0.06	0.07	0.08	0.08	0.08	0.08	0.08	0.07	0.08

Step III: *Final Decision*

- *The maximum weight value was chosen as representative for the TP weight vote.*
- *In this case, Class 1 is the highest weighted value (0.96). Thus, TP weight vote is Class 1.*

Figure 4. TP weighted vote algorithm



Step I: *Predicting depression symptoms for each class in each classifier* ** Highest Weight*

Predicted Class 1			Predicted Class 2			Predicted Class 3			...	Predicted Class 10		
Classifier 1	Classifier 2	**Classifier 3**	Classifier 1	**Classifier 2**	Classifier 3	Classifier 1	**Classifier 2**	Classifier 3		Classifier 1	**Classifier 2**	Classifier 3
0.91	0.16	**0.96***	0.00	**0.06***	0.00	0.00	**0.07***	0.00		0.00	**0.08***	0.00

Step II:
*Calculating the average weight values of each class (*avgClass$_i$*)*

$$\text{avgClass}_n = \frac{\sum Weighted\ Value\ m}{c\ Classifier}$$

Step III: *Comparing the maximum average weight values of each class*

AVG Class 1	AVG Class 2	AVG Class 3	AVG Class 4	AVG Class 5	AVG Class 6	AVG Class 7	AVG Class 8	AVG Class 9	AVG Class 10
0.68*	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03

Step IV: *Final Decision*

- *The maximum average weight value was chosen as representative for the AVG TP weight vote.*
- *In this case, Class 1 is the highest weight value (0.68). Thus, AVG TP weight vote is Class 1.*

Figure 5. AVG TP weighted vote algorithm

2.5.4 Evaluation

The evaluation phase was divided into three stages, and the details are as follows:

Applying the model: deployment to classify the depressive symptoms that were contained in 30 datasets as 15 non-depressive and 15 depressive samples for checking and demonstrating the classification model performance.

Depressive analysis: Using the results of a model predicting nine symptoms of depression. This involved counting the frequency and symptom duration in the system for 14 days, with moving up of the period with a one-day moving window, and calculating the new frequency until completed. Then the sum of the scores was calculated to measure the severity level of depression.

Evaluation: This research evaluates the efficiency of the models using Accuracy, Precision, Recall, and F1 as shown in equations (9) (10) (11) and (12), respectively.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$\text{Precision}=\frac{TP}{TP+FP} \tag{10}$$

$$\text{Recall}=\frac{TP}{TP+FN} \tag{11}$$

$$\text{F1}=\frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{12}$$

where *TP* is an outcome for the model correctly predicted event values, *FP* is an outcome for the model incorrectly predicted event values, *TN* is an outcome for the model correctly predicted no-event values, and *FN* is an outcome for the model incorrectly predicted no-event values.

3. Results and Discussion

This section is divided into three parts: 1) training data part to modeling for the feature for application of MOADSP, 2) testing data performance to evaluate the model, and 3) paired samples t-test [39] to compare the significance of the result between the unweighted and weighted voting ensemble method. The parameter settings and the results of the experiment are shown in Table 5.

Table 5. Parameter setting for each basic classification

Method	Parameter Setting
NB	estimation mode = greedy, minimum bandwidth = 0.1, number of kernels = 10, use application grid = true, application grid size =200
RF	number of trees = 100, criterion = gain ratio, maximal depth =10, guess subset ratio = true, voting strategy = confidence vote
KNN	k=3, measure type = NumericalMeasures, numerical measure = CosineSimilarity, weighted vote = true

3.1 Training data performance

The results of the text classification of the training data are shown in Table 6. The best accuracy and recall values came from KNN, and were 90.79% and 98.60%, respectively. It works and produces well when decision conditions are complex and it is robust for noisy data [40]. However, RF performed with the best precision for six classes, namely depressive, appetite, thinking, guilt, suicidal, and normal classes. KNN gave the best recall in the loss of interest class, equal to 98.60%. NB had the best precision in the loss of interest class, equal to 98.95%. The processing time directly varied with the number of features.

Table 7 shows that the unweighted vote method gave the best recall for thinking and tiredness classes; the TP weighted vote method offered the best recalls for loss of interest, appetite, sleep, and guilt. The AVG TP weighted vote method gave the highest recall equal to 98.43% in the loss of interest class, and the best recall for the depressive, movement, suicidal, and normal classes.

3.2 Testing data performance

This process used the testing data of 15 people with depression (Yes) and 15 non-depression (No) people to measure the efficiency of the MOADSP modeling.

Table 8 shows that RF had the highest recall, equal to 86.67% for the depression class in the single model case. The AVG TP weighted gave the best recall equal to 80.00% in the depression class, and accuracy equal to 66.67% in the ensemble model. The unweighted voting had the highest processing time. Table 8 also shows the results of the data testing, and it was found that the group of non-depression showed a high overfitting rate because the number of posts per day affected the predictions due to comments or condolences made in reference to the news, events, and important days, as can be seen in the training data in comments such as:

Table 6. Single modeling with 1,300 features performance

Class	NB		RF		KNN	
	Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Depressive	87.90	91.07	**96.56**	89.83	84.08	96.30
Loss of interest	**98.95**	87.70	96.33	97.17	97.17	**98.60**
Appetite	94.30	87.13	**97.09**	90.17	92.82	95.20
Sleep	82.02	78.30	69.30	83.07	73.34	**91.07**
Thinking	91.97	86.23	**93.46**	89.53	91.49	78.17
Guilt	87.28	89.40	**96.54**	88.40	95.44	93.53
Tired	66.21	91.70	75.47	88.60	**94.28**	85.77
Movement	87.58	83.00	78.43	86.43	**95.23**	88.43
Suicidal	91.21	83.67	**96.46**	83.67	95.94	89.77
Normal	88.79	87.63	**95.56**	87.60	95.02	91.63
Accuracy	86.77		88.30		**90.79**	

Table 7. The recall of the 1,300 features of the vote ensemble modeling

Class	Unweighted	TP Weighted	AVG TP Weighted
Depressive	96.30	96.33	**96.47**
Loss of Interest	97.97	98.40	**98.43**
Appetite	94.47	**95.17**	95.13
Sleep	87.07	**91.17**	89.50
Thinking	**91.27**	75.30	91.13
Guilt	92.93	**94.87**	94.63
Tired	**89.97**	87.33	87.63
Movement	86.90	89.30	**90.00**
Suicidal	87.73	90.83	**91.17**
Normal	90.40	92.13	**92.37**

Table 8. The testing data performance modeling

	Models	Precision		Recall		F1		Accuracy	Time
		Yes	**No**	**Yes**	**No**	**Yes**	**No**		**(Minute)**
Single	NB	56.25	57.14	60.00	53.33	58.06	55.17	56.67	10:02
	RF	60.00	52.00	20.00	**86.67**	30.00	65.00	53.33	15:58
	KNN	47.37	45.45	60.00	33.33	52.94	38.46	46.67	60:46
Ensemble	Unweighted	45.00	45.00	60.00	26.67	51.43	32.00	43.33	72:13
	TP weighted	61.11	66.67	73.33	53.33	66.67	59.26	63.33	-
	AVG TP weighted	63.16	72.73	**80.00**	53.33	70.59	61.54	**66.67**	-

"We are lost. Our children are seeing yet another school shooting",

" How to reduce mass shooting deaths? Experts rank gun laws",

"Another school shooting last Thursday w/3 dead and I'm only now hearing about it"

"Chants of no more guns break out as thousands attend candlelight vigil honoring Florida shooting victims and death ", and

" Watching yet another school shooting, big loss, condolences to their families".

These are not related to the symptoms but contains words that be within the scope of the depression symptoms, such as loss, shoot, dead, and gun. Moreover, it was found for groups with depression that their messages were routinely related to work. It had nothing to do with expressing feelings or symptoms of depression, and then it was an analysis based on DSM-5 that provided the results that were different from the training data. Example were "Throwback Thursday: this photo was taken yesterday?", "Fashion today show", and "Almost time for the next Flower Shop Mystery: Snipped in the Bud! Mark calendars Sunday". The same applied for differences in the training data and testing data of the features such as tiredness, sadness, lethargy, loafing, depressive, fatigued, and laziness.

3.3 Paired samples t-test

This research compared the effectiveness of the models for the unweighted and weighted voting methods by the paired samples t-test and the hypotheses that expressed the differences and the significance of the deployment performance as follows:

H_0= The unweighted has a better performance method than TP weighted voting method.

H_1= The TP weighted has a better performance method than the unweighted voting method.

And H_0= The unweighted has a better performance method than AVG TP weighted voting method.

H_1= The AVG TP weighted has a better performance method than the unweighted voting method.

The precision, recall, and F1 from the results of the testing data were calculated as shown in Table 9. From Table 9, the difference between the unweighted and TP (Yes and No) methods that was statistically very significant (p-value < 0.05), meaning rejection of the H_0 hypothesis and acceptance of the H_1 hypothesis. The difference between the unweighted and AVG TP (Yes and No) methods was statistically very significant (p-value < 0.05), indicating rejection of the H_0 hypothesis and acceptance of the H_1 hypothesis. Therefore, the TP weighted and AVG TP weighted methods were suitable for the major depressive disorder classification models.

Table 9. The paired samples t-test between the performance of the unweighted and weighted vote ensemble

	Methods	**Mean**	**S.D.**	**SEM**	**95% Confidence Interval of This Difference**		**t**	**Sig (2-tailed)**
					Lower	Upper		
Yes	Unweighted	52.143	7.525	4.344	-18.426	-11.360	18.140	0.0030
	TP weighted	67.037	6.118	3.532				
	Unweighted	52.143	7.525	4.344	-22.878	-13.336	35.926	0.0008
	AVG TP weighted	71.250	8.439	4.873				
No	Unweighted	34.557	9.429	5.444	-32.820	-17.573	14.221	0.0049
	TP weighted	59.756	6.684	3.859				
	Unweighted	34.557	9.429	5.444	-31.593	24.3604	33.286	0.0009
	AVG TP weighted	62.533	9.738	5.622				

4. Conclusions

In this research, we found the appropriate number of features, improved the weighted vote ensemble using multi-objective optimization, and compared the performance models of single and vote ensemble models for depressive disorder from Twitter. Creating the model for the number of features was done using the IG method, and a binary term occurrence weighting was appropriate for the short term. The single classification techniques used the NB, RF, and KNN techniques for creating the classification model. The vote ensemble models used the unweighted, TP weighted, and

AVG TP weighted methods. The weighting values were calculated from the True Positive from the single model applied to MOADSP. The feature selection process revealed the 1,300 features modeling to apply MOADSP to calculate each class's weighting. Then the deployment phase used the test set for comparing the performance model, paired-samples t-test of the results between the unweighted and TP weighted, the unweighted and AVG TP weighted voting methods.

The results of the 1,300 features modeling for the single modeling showed that the best precision of all classes was NB (98.95%) of the loss of interest class, but the best recall and accuracy of all classes was KNN with (98.60%) and (90.79%) for the loss of interest class, respectively. The resulting proposed model was better and used less features than described by Doenribram *et al*. [13]. After that, the application to the MOADSP phase for the vote ensemble modeling showed that the best recall of all classes was AVG TP weighted (98.43%) for the loss of interest class. The results for all classes except the sleep and tired classes were greater than 90% because the words were ambiguous and found in both classes. The example words are "tired" and "time". This model was found to be more effective and accurate when the text contained more than three terms. The result of the deployment phase using the test data of 15 people with depression and 15 non-depressive persons showed that the best accuracy was the AVG TP weighted (66.67%). The highest recall in the non-depressive class was RF (86.67%), and the highest recall in the depressive class was AVG TP weighted (80.00%). The result of the paired samples t-test process of the TP and AVG TP weighted in the depression of both classes was statistically significant, and the p-value was less than 0.05.

This research can solve the unstable problem of unweighted voting in case of a non-majority vote; the inefficiency results occur because of weak classifiers. The training data reached high performance, and conversely with the testing date because some people fed many Tweets and retweets of news that were social news such as "bored with the time #bored", which affected the model.

Finally, this research can resolve the problems of the single classification model and the unweighted voting ensemble and we chose the appropriate number of attributes for the dataset. This made the model effectively when the number of terms was more than three terms and took a short time to process. Therefore, in future work, we would like to analyze the keywords that are within the scope of levels 1-3 for depression and the sequence of symptoms leading to depression.

5. Acknowledgments

This research project was financially supported by Mahasarakham University. The authors would like to thank our advisors, lecturers, technicians, and respondents for their full support.

References

[1] Negrão, A.B. and Gold, P.W., 2007. Major depressive disorder. *Encyclopedia of Stress*, 28, 640-645.

[2] Mousavian, M., Chen, J. and Greening, S., 2018. Feature selection and imbalanced data handling for depression detection. *International Conference on Brain Informatics*, Arlington, USA., December 7-9, 2018, pp. 349-358.

[3] World Health Organization, 2021. *Depression*. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/depression.

[4] World Population Review, 2021. *Depression Rates by Country 2021*. [online] Available at: https://worldpopulationreview.com/country-rankings/depression-rates-by-country.

[5] UNICEF, 2021. *SOS Helplines for Parents and Children – Essential Services for Preventing Suicides*. [online] Available at: https://www.unicef.org/montenegro/en/stories/sos-helplines-parents-and-children-essential-services-preventing-suicides.

[6] Lin, L.Y., Sidani, J.E., Shensa, A., Radovic, A., Miller, E., Colditz, J.B., Hoffman, B.L., Giles, L.M. and Primck, B.A., 2016. Association between social media use and depression among U.S. young adults. *Depression and Anxiety*, 33(4), 323-331.

[7] Simon, K., 2020. *Digital 2020: July Global Statshot*. [online] Available at: https://datareportal.com/reports/digital-2020-july-global-statshot.

[8] Hand, D., Mannila, H. and Smyth, P., 2001. *Principles of Data Mining*. Cambridge: MIT Press.

[9] Jimenez-Marquez, J.L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J.L., and Ruiz-Mezcua, B., 2019. Towards a big data framework for analyzing social media content. *International Journal of Information Management*, 44(C), 1-12.

[10] Harrigian, K., Aguirre, C. and Dredze, M., 2020. On the state of social media data for mental health research. *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology*, Mexico City, Mexico, June 11, 2020, pp. 15-24.

[11] Coppersmith, G., Dredze, M., and Harman, C., 2014. Quantifying mental health signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Baltimore, USA, June 27, 2014, pp. 51-60.

[12] Mowery, D., Park, A., Conway, M. and Bryan, C., 2016. Towards automatically classifying depressive symptoms from Twitter data for population health. *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, Osaka, Japan, December 12, 2016, pp. 182-191.

[13] Doenribram, D., Jareanpon, C., Jiranukool M.D. and Jariya, S., 2019. Major depressive disorder classification from user behaviors from twitter. *The Twenty-Fourth International Symposium on Artificial Life and Robotics 2019 (AROB)*, Beppu, Japan, January 23-25, 2019, pp. 241-246.

[14] Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Chua, T.-S. and Zhu, W., 2017. Depression detection via harvesting social media: a multimodal dictionary learning solution. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, Melbourne, Australia, August 2017, pp. 3838-3844.

[15] Wolohan, J.T., Hiraga, M., Mukherjee, A. and Sayyed, Z.A., 2018. Detecting linguistic traces of depression in topic-restricted text: attending to self-stigmatized depression with NLP. *Proceedings of the First International Workshop on Language Cognition and Computational Models*, New Mexico, USA., August 20, 2018, pp. 11-21.

[16] Onan, A., Korukoğlu, S. and Bulut, H., 2017. A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing and Management*, 53(4), 814-833.

[17] Zul, M.I., Yulia, F. and Nurmalasari, D., 2018. Social media sentiment analysis using K-means and naïve Bayes algorithm. *Proceedings of 2nd International Conference on Electrical Engineering and Informatics: Toward the Most Efficient Way of Making and Dealing with Future Electrical Power System and Big Data Analysis (ICon EEI)*, Batam, Indonesia, October 16-17, 2018, pp. 24-29.

[18] Shatte, A.B.R., Hutchinson, D.M. and Teague, S.J., 2019. Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, 49(9), 1426-1448.

[19] Ahmed, Z., Mohamed, K., Zeeshan, S. and Dong, X. Q., 2020. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database : The Journal of Biological Databases and Curation*, 2020, DOI:

10.1093/database/baaa010.

[20] Bhakta, I. and Sau, A., 2016. Prediction of depression among senior citizens using machine learning classifiers. *International Journal of Computer Applications*, 144(7), 11-16.

[21] Zhang, W., Liu, H., Silenzio, V.M.B., Qiu, P. and Gong, W., 2020. Machine learning models for the prediction of postpartum depression: Application and comparison based on a cohort study. *JMIR Medical Informatics*, 8(4), DOI: 10.2196/15516.

[22] Onan, A., Korukoğlu, S. and Bulut, H., 2016. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications*, 62, 1-16.

[23] American Psychiatric Association, 2013. *Diagnostic and Statistical Manual Psychiatric of Mental Disorder DSM-5*. 5th ed. Arlington: American Psychiatric Association.

[24] Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y. and Ohsaki, H., 2015. Recognizing depression from twitter activity. *Conference on Human Factors in Computing Systems*, Seoul, Korea, April 18-23, 2015, pp. 3187-3196.

[25] Kang, K., Yoon, C. and Kim, E. Y., 2016. Identifying depressive users in Twitter using multimodal analysis. *2016 International Conference on Big Data and Smart Computing (BigComp)*, Hong Kong, China, January 18-20, 2016, pp. 231-238.

[26] Burdisso, S.G., Errecalde, M. and Montes-y-Gómez, M., 2019. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133, 182-197.

[27] Alabdulkreem, E., 2020. Prediction of depressed Arab women using their tweets, *Journal of Decision Systems*, 30(2-3), 102-117.

[28] Aldarwish, M.M., and Ahmad, H.F., 2017. Predicting depression levels using social media posts. *Proceedings of IEEE 13th International Symposium on Autonomous Decentralized Systems (ISADS)*, Bangkok, Thailand, March 22-24, 2017, pp. 277-280.

[29] Dash, M. and Liu, H., 1997. Feature selection for classification. *Intelligent Data Analysis*, 1(3), 131-156.

[30] Pintas, J.T., Fernandes, L.A.F., Cristina, A., and Garcia, B., 2021. Feature selection methods for text classification : a systematic literature review. *Artificial Intelligence Review*, 54, 6149-6200.

[31] Rastogi, S., 2018. Improving classification accuracy of automated text classifiers. *7th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions (ICRITO),* Noida, India, August 29-31, 2018, pp. 239-245.

[32] Zhu, L., Wang, G. and Zou, X., 2017. Improved information gain feature selection method for Chinese text classification based on word embedding. *Proceedings of the 6th International Conference on Software and Computer Applications (ICSCA)*, Bangkok, Thailand, February 26-28, 2017, pp. 72-76.

[33] Han, J., Kamber, M. and Pei, J., 2011. *Data Mining: Concepts and Techniques*. 3rd ed. Amsterdam: Elsevier.

[34] Moonpen, U., Mungsing, S. and Banditwattanawong, T., 2021. Classification model development based on cluster-to-class distance mapping for tourism form prediction of Inbound tourism market in Thailand. *Current Applied Science and Technology*, 21(2), 393-407.

[35] Nuankaew, P., Chaising, S. and Temdee, P., 2021. Average weighted objective distance-based method for type 2 diabetes prediction. *IEEE Access*, 9, 137015-137028.

[36] Nuankaew, W. and Thongkam, J., 2020. Improving student academic performance prediction models using feature selection. *17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Phuket, Thailand, June 24-27, 2020, pp. 392-395.

[37] Dehzangi, A. and Karamizadeh, S., 2011. Solving protein fold prediction problem using fusion of heterogeneous classifiers. *Information*, 14(11), 3611-3621.

[38] Rojarath, A. and Songpan, W., 2021. Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems. *Applied Intelligence*, 51(7), 4908-4932.

[39] Ndirangu, D., Mwangi, W. and Nderu, L., 2019. A hybrid ensemble method for multi-class classification and outlier detection. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, 45(1), 192-213.

[40] Hassan, S.U., Ahamed, J. and Ahmad, K., 2022. Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Computers*, 3, 238-248.