Big data analytics on social networks for real-time depression detection : A Review

Gurleen Kaur, Manik Nagpal

DAV INSTITUTE OF ENGENEERING AND TECHNOLOGY, JALANDHAR

Abstract— After the coronavirus epidemic, cases of depression increased significantly. Many depressed people publicly share their true feelings on social media. Therefore, big data analytics in social networks are recommended for real-time detection of depression. This research used the opinions and demographics of Twitter users. This was collected over a period of two months after completing the Patient Health Questionnaire-9, which served as the outcome measure of the study, and identified depressive symptoms. Machine learning techniques were used to build the recognition model. Support Vector Machines, Decision Trees, Naive Bayes, Random Forests, and Deep Learning are the five machine learning methods examined in this research. This study contributes to the body of knowledge by presenting an entirely new model based on surveys of Twitter user demographics and text sentiment. Thus, this work is a step towards reducing depression-induced suicide rates.

Keywords: Big data analytics, Depression detection, Social networks

I. Introduction

Depression is defined as "a mental state that reflects mood disorders such as depression, unhappiness, boredom, loss of appetite, and difficulty concentrating." Depression affects quality of life and can eventually lead to suicide. The World Health Organization predicts that in 2021, 280 million people worldwide will suffer from depression. In low- and middle-income countries, 80% of them received no treatment. The number of depression cases in the United States tripled during the coronavirus epidemic. Meanwhile, the prevalence of depression is rising in Thailand, with depression being responsible for 60% of suicides. Therefore, it is important to prevent and treat mental illness and promote mental health.

As established, all people are at risk of developing depression and most people are unaware of it and therefore do not seek treatment. It often expresses itself. Social networks allow people to express themselves through messages and comments on photos, rather than eye contact and facial expressions. Social media data can be used to identify people's grief based on the negative comments they have shared.

Machine learning techniques were applied as the detection model construction. There are five machine learning techniques explored in this research which are SuPport Vector Machine, Decision Tree, Random Forest, Naïve Bayes, and Deep Learning. Support Vector Machine tries to choose the optimal decision boundary (i.e., hyper- plane) by maximizing the margin distance between classes using Hinge Loss function as defined by the equation:

$$l(y) = max\left(0{,}1 + \max_{y \neq t} w_y x - w_t x\right) \quad \text{Eq. 1}$$

where t is the target label, w_t and w_y are the model parameters. Decision tree divides the dataset based on the attribute that divides the dataset most effectively. The attribute which provides maximum Information Gain is selected for splitting. Information Gain is the entropy of the parent node minus the sum of entropy of the child nodes, where entropy is defined as :

$$\text{Entropy} = \sum_{i=1}^{c} -P_i * \log_2 P_i \quad \text{Eq. 2}$$

Random forest is composed of several decision trees, which operate as an ensemble. Each decision tree predicts a class outcome. The class outcome with the most number of votes is the prediction of random forest. Thus, the decision trees should be least correlated with each other, i.e., each decision tree predicts using different features. Naïve Bayes classifies data based on Bayes's theorem.

If C_k rep- resents possible classes, and vector x represents features, the conditional probability can be defined as:

$$P(C_k \mid x) = \frac{P(x|C_k)P(C_k)}{P(x)} \quad \text{Eq.3}$$

where $P(C_k|x)$ is posterior, $P(x|C_k)$ is likelihood, $P(C_k)$ is prior, and $P(x)$ is evidence. Deep Learning is based on artificial neural networks, inspired by information processing and distributed communication nodes in biological brain. It uses multiple hidden layers to extract features from inputs. An output of aneural is the activation function f of a weighted sum of the neuron's input i.e., output is

$$f\left(b + \sum_{i=1}^{n} x_i w_i\right) \quad \text{Eq. 4}$$

where w_i is a corresponding weight of input x_i, and b is bias.

II. Literature review

There are several existing research papers focused on detecting depression using social networks. This work can be divided into three groups according to purpose. The first group (such as [10–14]) is aimed at analyzing sentiment from social network data. Barhan and Shakhomirov developed a model based on n-grams to classify emotions from Twitter data using emoticons and moods. The results showed that Support Vector Machines outperform Naive Bayes. It has 81% cure and 74% recall. This result was consistent with the work of Hutto and Gilbert, who presented a performance comparison of sentiment analysis algorithms for analyzing social network messages. We analyzed 4,000 Twitter messages using Support Vector Machines, Naive Bayes, and Maximum Entropy algorithms.

We used the behavioral and language data of 10,102 social network users to build classification and regression models. They used data mining techniques to predict mental health and depression levels from comment messages, with an accuracy value of 63.30%. Reese et al. used a Bayesian logistic regression algorithm to create a depression prediction model based on Instagram images. A sample of 166 participants and 43,950 images of her were collected for color analysis and facial recognition. Experimental results showed that the model achieved better efficiency than conventional diagnostics, with an overall efficiency of 61%. San et al. proposed a hybrid model that combines the long-term short-term memory of convolutional neural networks with the Markov chain Monte Carlo method. This model was used to identify user emotions, mood changes and mood disorders. Experimental results showed that the model could detect irregular transformations and emotional disturbances.

Table 1 A comparison of research related to depression detection using social network

Depression detection	Related work												This work
	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]	[21]	
Data source													
Facebook				✓						✓			
Instagram											✓		
Sina Weibo									✓				
Twitter	✓	✓	✓		✓	✓	✓	✓				✓	✓
Input feature													
Message	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓
Sentiment	✓	✓	✓	✓	✓	✓	✓			✓		✓	✓
Emoticon	✓					✓			✓				
Image											✓		
Behavior									✓				✓
Profile						✓			✓				✓
Methodology													
Decision Tree				✓	✓								✓
Deep Learning			✓					✓				✓	✓
K-Nearest Neighbors				✓									
Maximum Entropy		✓											
Naïve Bayes	✓	✓	✓							✓			✓
Random Forest											✓		✓
Regression						✓			✓				
Support Vector Machine	✓	✓	✓	✓						✓			✓
Result													
Positive/negative	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Depression level													✓

III. Real-time depression detection methodology

Real-time depression detection is the process of collecting streaming data from the Twitter application programming interface (API) by extracting, transforming, and loading the data into a data store on a Hadoop cluster. This data is used as an invisible input to the depression detection model. This model is used as a classifier to identify levels of depression in individuals. The system has four user modes: self, parents, counselor, and her employer. Self mode allows users to recognize their level of depression using their Twitter ID as input for the system. Parent mode allows parents to track their child's status by providing the child's Twitter ID to the system. In Employer Mode, an employer can also track an employee's status by providing the employee's Twitter ID to the system. Figure 1 shows the real-time depression detection process.

IV. Depression detection methodology

This research aims to develop a depression detection model using demographic characteristics and sentiment analysis from tweets. The research methodology consists of 5 processes as shown in Fig. 2 which are data acquisition, data transformation, data storage, model construction, and model performance evaluation. The details of each process are described as follows

- **Data acquisition**

Data collection for model building in this study comes from three sources: PHQ-9, personal information questionnaires, and the Twitter API. First, the data obtained from PHQ-9 are survey response dates and depression rating scores for each participant.

Second, the data collected from the personal information questionnaire includes Twitter ID and information such as gender, age, weight, education, congenital diseases, career, income, number of family members, cohabiting couple status, and parents' marital status. Includes demographics.

Finally, we use the Twitter API to collect her real-time Twitter user information from her Twitter page. Twitter IDs collected from the first source are used to access Twitter users' information. Information was collected two months prior to the survey response date. There were 222 tweets, 1522 retweets, and 16 hashtags collected from 192 Twitter users, of which 50 had moderate depression, 74 had mild depression, and 68 had evidence of depression.

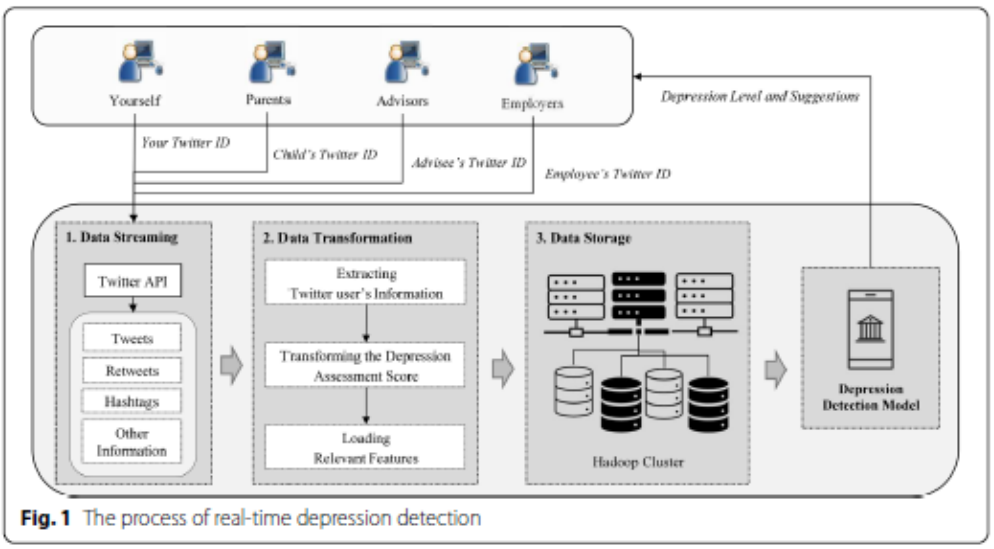


Fig. 1 The process of real-time depression detection

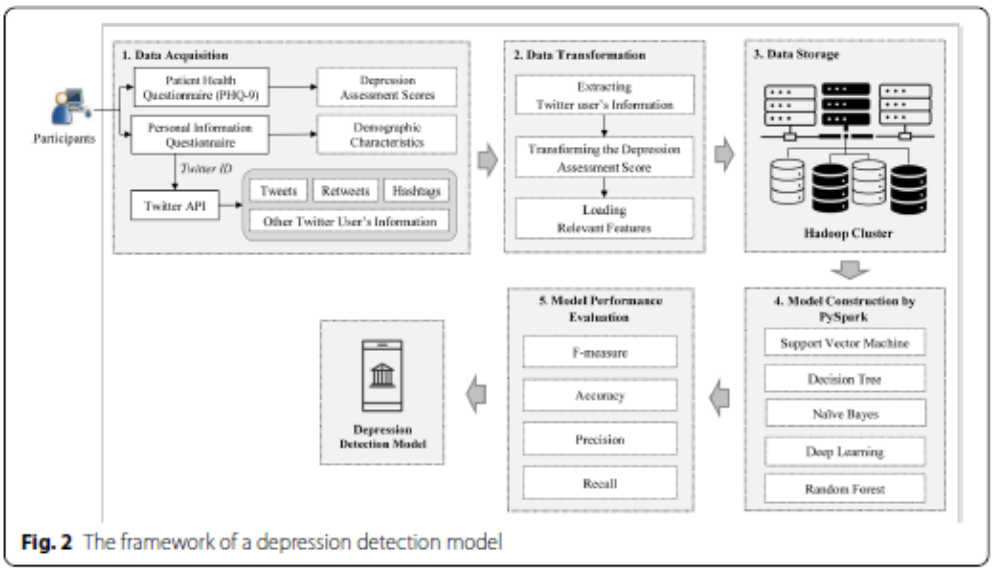


Fig. 2 The framework of a depression detection model

- **Data transformation**

The information obtained from Twitter will be processed to find sentiment attributes andtheir values. The process of extraction is illustrated in Fig. 3 and can be explained as follows. Firstly, the Twitter ID is used to access the Twitter user's information via Twitter API.The extracted information consists of tweets, retweets, hashtags, the number of friends, the number of followers, and periods of tweets.

Secondly, terms in the tweets, retweets, and hashtags are extracted using an NLTK library. Thirdly, the extracted terms in Thai are translated to English using Google translation API. Finally, the sentiment numerical scores of each term are derived from opinion lexicons in the WordNet database , where each term has three sentiment numerical scores: *Pos*(*s*), *Neg*(*s*) and *Obj*(*s*), according to Eq. (1).

$$Pos(s) + Neg(s) + Obj(s) = 1$$

The extracted sentiment attributes in this research comprise the number of positive and negative tweets, retweets and hashtags, the number of tweets, retweets and hashtags expressing depression, and sentiment score of all tweets.

The sentiment score of all tweets _ is derived from sentiment scores of each tweet *i* as shown in Eq. (2).

$$\Theta = \frac{\sum_{i=1}^{n} \text{Tweet}(s)_i}{n}$$

The sentiment score of each tweet $\text{Tweet}(s)_i$ is calculated from an average of positive and negative scores of all terms in the tweet as shown in Eq. (3).

$$\text{Tweet}\ (s) = \frac{\sum_{j=1}^{m} \text{Pos}(s)_j}{m} - \frac{\sum_{j=1}^{m} \text{Neg}(s)_j}{m}$$

The number of positive tweets Tweet_{pos} is the total number of tweets that have Tweet(s) > 0 . On the other hand, the number of negative tweets Tweet_{neg} is the total number of tweets that have Tweet(s) ≥ 0 .

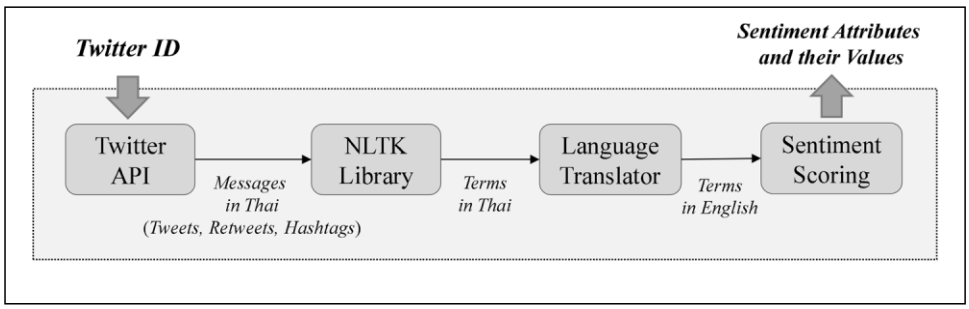


Fig. 3 The process of feature extraction

Table 2 Data characteristics of relevant features

Attribute ID	Attribute name	Data type	Description
x_1	Gender	Nominal	0: female, 1: male, 2: others
x_2	Age	Nominal	1: < 20, 2: 20–29, 3: 30–39, 4: 40–49, 5: 50–59, 6: ≥ 60
x_3	Weight	Nominal	1: < 40, 2: 40–59, 3: 60–79, 4: 80–99, 5: ≥ 100
x_4	Education	Nominal	1: high school diploma or less, 2: bachelor's degree, 3: graduate degree
x_5	CongenitalDisease	Nominal	1: no, 2: yes
x_6	Career	Nominal	1: government official, 2: state enterprise employee, 3: businessman, 4: student, 5: homemaker, 6: not currently employed
x_7	Income	Nominal	1: sufficient, 2: insufficient
x_8	FamilyMember	Numeric	Number of family members
x_9	Self-CoupleStatus	Nominal	1: no couple, 2: stay together with couple, 3: separated with couple
x_{10}	ParentMaritalStatus	Nominal	1: stay together with couple, 2: separated with couple, 3: couple died
x_{11}	PositiveTweets	Numeric	Number of positive tweets
x_{12}	NegativeTweets	Numeric	Number of negative tweets
x_{13}	DepressionTweets	Numeric	Number of tweets expressing depression
x_{14}	PositiveRetweets	Numeric	Number of positive retweets
x_{15}	NegativeRetweets	Numeric	Number of negative retweets
x_{16}	DepressionRetweets	Numeric	Number of retweets expressing depression
x_{17}	PositiveHashtags	Numeric	Number of positive hashtags
x_{18}	NegativeHashtags	Numeric	Number of negative hashtags
x_{19}	DepressionHashtags	Numeric	Number of hashtags expressing depression
x_{20}	SentimentScore	Numeric	Sentiment score of tweets
x_{21}	Friends	Numeric	Number of friends
x_{22}	Followers	Numeric	Number of followers
x_{23}	TweetPeriod1	Numeric	Number of tweets between 6 am and midnight
x_{24}	TweetPeriod2	Numeric	Number of tweets between midnight and 6 am
y	DepressionLevel	Nominal	Level 0: no depression, level 1: slight depression, level 2: depression

- **Data storage**

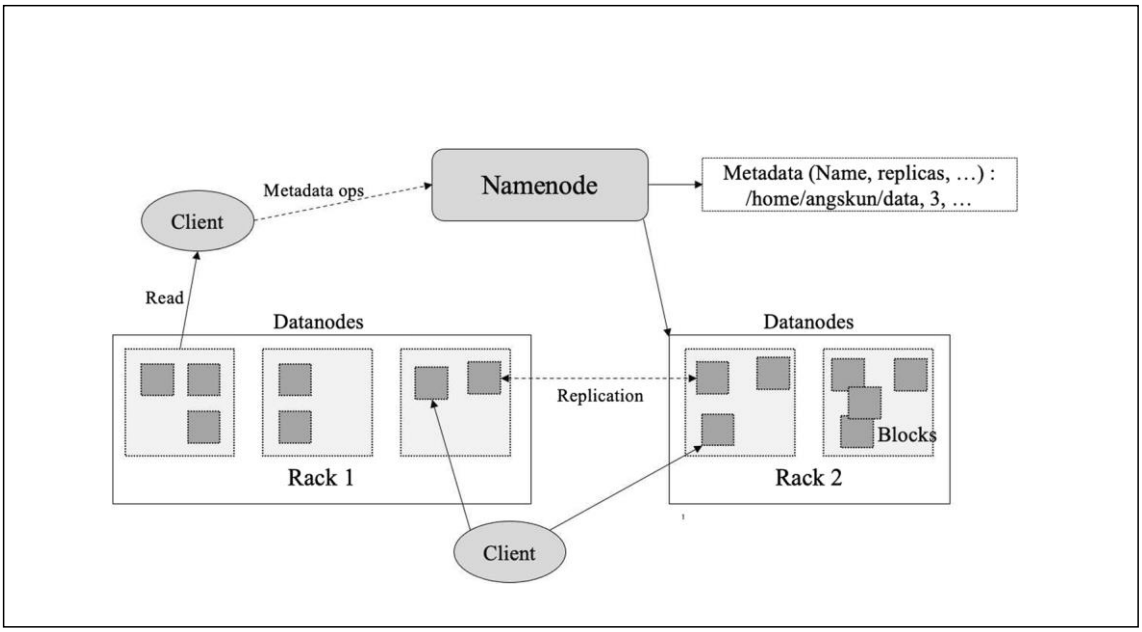
Data from the previous process is loaded into a Hadoop cluster, a specialized computing cluster designed to store and analyze large amounts of data. For this study, we used Hadoop's open-source software suite called Cloudera, which runs on commercial computers. Data collected by Twitter is stored in the Hadoop Distributed File System (HDFS). Saved. HDFS consists of two main components called name nodes and data nodes. Name nodes are provided as metadata, and data nodes contain the actual data blocks. As shown in Figure 4, files are stored in 128MB data chunks and each chunk is replicated to three different data nodes. A Hadoop cluster acts as a data lake for storing unstructured data from Twitter.

- **Model construction**

Create an automated depression detection model using machine learning. The type of machine learning used to build models is a classification technique. The data set, called the model's training set, consists of features or input variables (x1, . . . , x24) and the target or output variable (y), all listed in Table 2. More specifically, this work uses supervised machine learning techniques for model building, including support vector machines, naive Bayes, decision trees, random forests, and deep learning techniques. Model building is implemented using Spark ML lib, Spark's machine learning library with a Python interface. The advantage of Spark ML lib is that you can load huge datasets directly from HDFS.

This allows you to process large numbers of tweets using machine learning.

Fig. 4 HDFS architecture



V. Experimental evaluation

- **Testing environment**

The evaluation uses the same test set as the model-building training set and 192 Twitter users consisting of three classes. Twitter user information consists of 222 tweets, 1,522 retweets, and 16 hashtags.

Depression detection models were evaluated in three dimensions. The first focuses on the selection of attributes (called traits) that contribute to depression. This evaluation compares the use of demographic trait traits, traits extracted from Twitter user information, and mixed traits to build a depression detection model. The second aspect aims to reduce the number of features and computational complexity of modeling in order to improve model performance. The third aspect deals with performance comparisons of several machine learning techniques. Specifically, support vector machines with kernel radial basis functions (RBFs), decision trees with the C4.5 algorithm, naive Bayes, random forests with 100 trees in the forest, and hidden depths of 100 A feedforward network with layers and ReLU activations. These techniques are compared using scores that include F-score, accuracy, precision, and recall [26], as in Eq. (4)–(7):

$$\text{Accuracy} = \frac{TP+TN/}{(TP+TN+FP+FN)} \tag{4}$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \tag{5}$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \tag{6}$$

$$F - \text{measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

For depression level 0 (no depression), the *TP* (True Positive) means that persons with no sign of being depressed are correctly predicted as no depression. The *FP* (False Positive) means that persons who are slightly or moderately depressed are incorrectly predicted as no depression (actual depression level is 1 or 2). The *FN* (False Negative) means that persons with no sign of being depressed are incorrectly predicted as depression level 1 or 2 (actual depression level is 0). The *TN* (True Negative) means that persons who are slightly or moderately depressed are correctly predicted as depression level 1 or 2. The *TP*, *FP*, *FN*, *TN* of depression level 1 and 2 are derived in the same way as those of depression level 0.

- **Experimental results and discussion**

This section presents and discusses experimental results based on the test environment in the previous section. The first experiment aims to determine the impact of different properties on the performance of various machine learning techniques such as Support Vector Machines, Decision Trees, Naive Bayes, Random Forests and Deep Learning techniques. The results shown in Figure 5 show that the mixed trait, which is a combination of demographic and informational traits of Twitter users, achieved the highest average f-value among all machine learning methods of 0.728. Therefore, in this study, 24 traits from mixed traits were used to build a depression detection model.

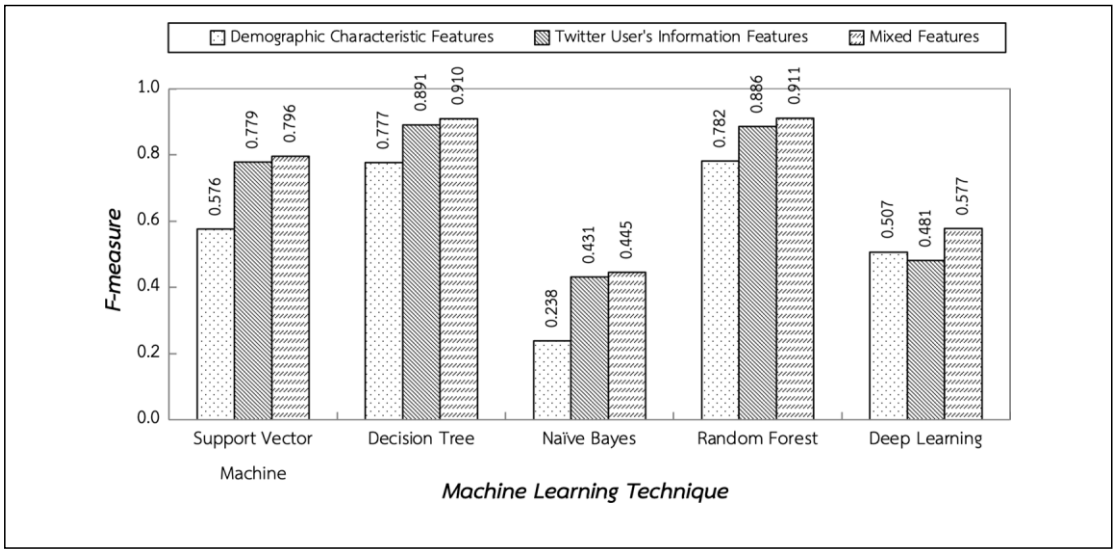


Fig. 5 The effect of different features on performance of a variety of machine learning techniques

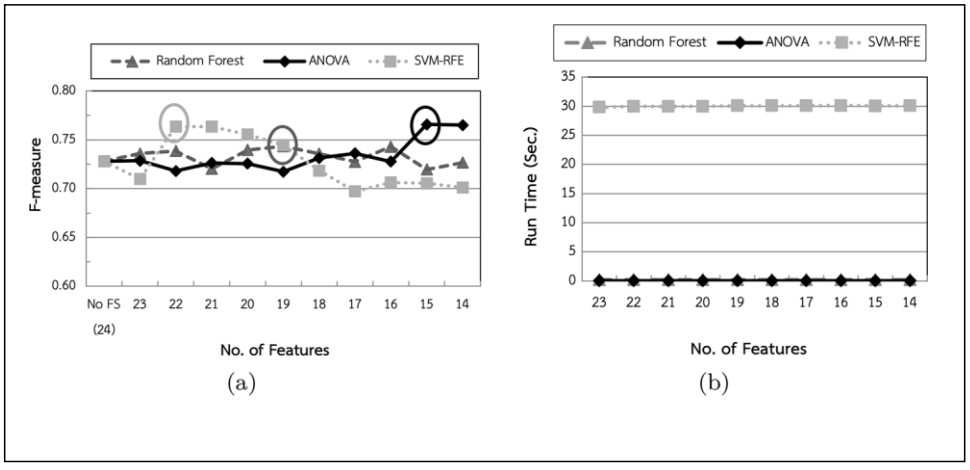


Fig. 6 Feature selection performance: **a** F-measure; **b** processing time

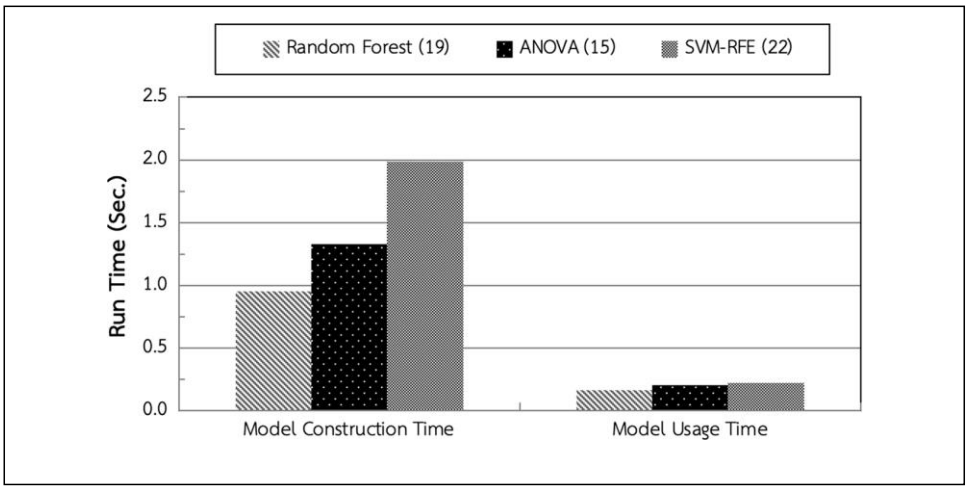


Fig. 7 Model construction time and model usage time of feature selection

Figure 6a shows that the optimal number of features for random forest, ANOVA, and SVM-RFE decreased from 24 features to 19, 15, and 22 features, respectively, while Figure 6b shows that feature selection by ANOVA decreased to 7 times and 917 times. It was faster than Random Forest and his SVM-RFE. Our results also showed that the number of features does not affect the processing time of feature selection. In this experiment, we concluded that 19, 15, and 22 features are adequate numbers for the Random Forest, ANOVA, and SVM-RFE methods, respectively. The results show that the proposed feature selection model using random forest (F measure = 0.7435), ANOVA (F measure = 0.7657), and SVM-RFE (F measure = 0.7646) outperforms existing models without feature selection. is also better (F measure = 0.728).

Experimental results showed that model building and using ANOVA is faster than SVM-RFE, but slightly slower than random forest, as shown in Figure 7. However, ANOVA achieved the best feature selection performance, as shown in Figure 6.

The final experiment aimed to find the most appropriate machine learning method for detecting depression using standard measures of F-value, accuracy, precision, and recall. timid. Figure 8 shows the performance of decision trees, naive Bayes, random forests, and deep learning techniques compared to the support vector machine chosen as the baseline. Results showed that the decision tree and random forest methods outperform baseline accuracy. Therefore, random forest was the most suitable machine learning technique for detecting depression.

VI. Conclusion and future work

This article presents a model for detecting depression in social networks using big data analytics. The main theoretical implication of this study is a new model based on analysis of Twitter user demographics and text sentiment. Two of his feature selection techniques, called ANOVA and SVM-RFE, are applied to improve the performance of the model. Experimental results showed that ANOVA achieved higher f-values with less processing time than his SVM-RFE. Machine learning techniques are then applied to build depression detection models. This research explores five machine learning techniques: Support Vector Machines, Decision Trees, Naive Bayes, Random Forests, and Deep Learning. Experimental

results showed that the random forest method can detect depression with higher accuracy than other machine learning methods.

There are some improvements that can be made in the near future. First, the number of participants can be increased to increase accuracy. Second, additional characteristics such as time spent on social networks and number of likes per post can be examined for modeling. Third, refinement of the model building process can be performed using other data such as emoticons and images. Finally, data can be collected from multiple social networks.

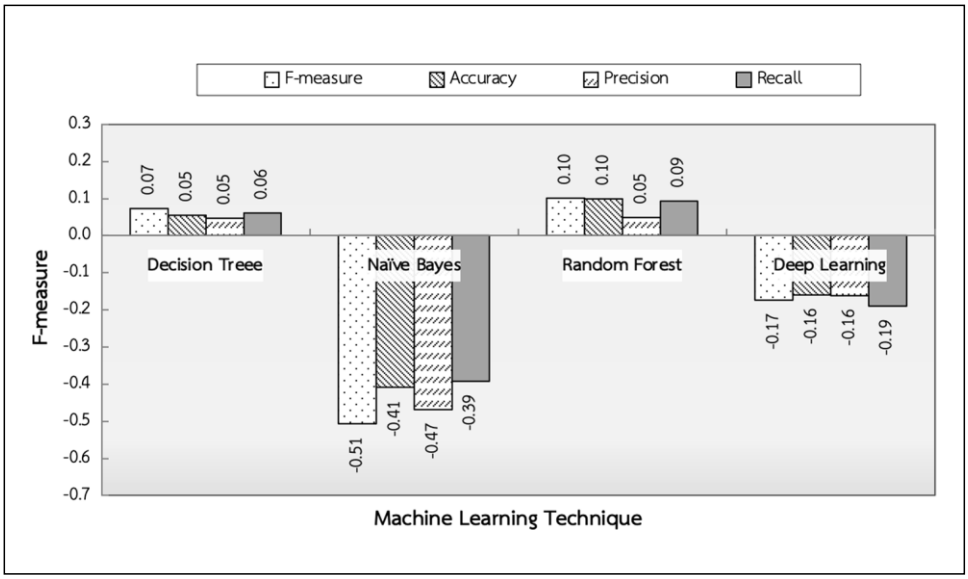


Fig. 8 Performance comparison of machine learning techniques compared with Support Vector Machine

VII. Abbreviations:

ANOVA: Analysis of Variance; API: Application Programming Interface; HDFS: Hadoop Distributed File System; PHQ-9: Patient Health Questionnaire-9; SVM-RFE: Support Vector Machine-Recursive Feature Elimination

VIII. References

1. Hongsrisuwan N. Depression. HCU. J Health Sci. 2016;19(38):105–18.
2. WHO. Depression fact sheets. (2021). http://www.who.int/news-room/fact-sheets/detail/depression.
3. Khubchandani J, Sharma S, Webb FJ, Wiblishauser MJ, Bowman SL. Post-lockdown depression and anxiety in the USA during the COVID-19 pandemic. J Public Health. 2021;43(2):246–53. https://doi.org/10.1093/pubmed/fdaa250.
4. Wongpiromsarn Y. Mental health and the COVID-19 crisis in Thailand. J Ment Health Thail. 2020;28(4):280–91.
5. Phanichsiri K, Tuntasood B. Social media addiction and attention deficit and hyperactivity symptoms in high school students in Bangkok. J Psychiatr Assoc Thail. 2016;6(13):191–204.

6. Choudhury MD, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: Proceed- ings of the 5th annual ACM web science conference. New York: ACM; 2013. p. 47–56. https://doi.org/10.1145/2464464.2464480.
7. Tsai C-W, Lai C-F, Chao H-C, Vasilakos AV. Big data analysis: a survey. J Big Data. 2015;2(21):1–32. https://doi.org/10.1186/s40537-015-0030-3.
8. Kolajo T, Daramola O, Adebiyi A. Big data stream analysis: a systematic literature review. J Big Data. 2019;6(47):1–30. https://doi.org/10.1186/s40537-019-0210-7.