JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A survey on Mental Health Prediction

1st Rajat Singh
Student, Department of Computer Science And Engineering
Buddha Institute of Technology,
GIDA, Gorakhpur

2nd Pranav Srivastava
Student, Department of Computer Science and Engineering
Buddha Institute of Technology,
GIDA, Gorakhpur

3rd Pragati Mani Tripathi
Student, Department of Computer Science and Engineering
Buddha Institute of Technology,
GIDA, Gorakhpur

4th Jaya Agrawal
Student, Department of Computer Science and Engineering
Buddha Institute of Technology, GIDA, Gorakhpur
5th Pallavi Dixit
Assistant Professor in
Department of Computer Science and Engineering
Buddha Institute of Technology,
GIDA, Gorakhpur

abstract - In today's fast-paced modern world, mental health issues such as anxiety, depression, and stress have become very common in the general population. In this article, predictions of anxiety, depression, and stress were made using machine learning algorithms. By applying these algorithms to, data was collected from employed and unemployed people across cultures and communities via the Depression, Anxiety and Stress Scale Questionnaire (DASS 21). Anxiety, depression, and stress were predicted by five different machine learning algorithms at five levels of severity - and because of their high accuracy they are particularly good at predicting psychological problems. After applying different methods, classes were found to be unbalanced in the confusion matrix. Therefore, the f1-score metric has been added, which helps to determine the best accuracy model among the five algorithms applied as a random forest classifier. Additionally, the specificity parameter indicates that the algorithm is also particularly sensitive to negative results.

Keywords: Decision Tree (DT); K neural network; Naive Bayesian (NB); Random Forest Tree (RFT); Support Vector Machine (SVM)

1. INTRODUCTION

Today, people become naturally ambitious and seek every possible opportunity to advance the professional level of Anxiety, depression, stress-related frustration and dissatisfaction have become so common that people now believe them to be part of professional life. The World Health Organization (WHO) observes that depression is the most common mental disorder, affecting more than 300 million people worldwide, and the scale of the problem has led many health researchers to focus their research in this area. Distinguishing between anxiety, depression, and stress is problematic for machines; therefore, appropriate learning algorithms are required for accurate diagnosis. According to the World Health Organization, a healthy person has a healthy brain and good physical health. [1] The standard diagnostic criteria for depression is the Patient Health Questionnaire (PHQ), while for depression the Anxiety and Stress Scale (DASS 21) has 21 questions to screen for symptoms associated with these psychiatric disorders. [5-6]. Clinically, the main symptoms of depression [3] are memory loss; inability to concentrate; inability to make decisions; loss of

interest in leisure activities and hobbies, including sex; depression and weight loss; feelings of guilt, worthlessness, helplessness, restlessness, and irritation; and suicidal thoughts. In case [2], these symptoms were found to have a significant impact on in important areas of an individual's life - such as education, employment and social activities, which provided clues important for to make a clinical diagnosis. GAD (Generalized Anxiety Disorder) [3] Symptoms are irritability, nervousness, fatigue, insomnia, gastrointestinal problems, panic and sense of impending danger, increased heart rate heart rate, sweating, shortness of breath and difficulty concentrating. Symptoms of stress [4] are feeling restless or irritable, inability to relax, lack of energy, chronic headaches, frequent overreactions, and persistent colds or infections. As a result, stress, anxiety, and depression have many common symptoms, including insomnia, chest pain, fatigue, racing heartbeat, and lack of concentration, all of which are difficult for people to categorize. machinery.

The article is structured as follows: Section 2 explores relevant research on anxiety, depression, and stress, and the methods and techniques employed by. Section 3 describes the materials and methods used in this study, while section 4 presents the results obtained after applying the classification algorithm. Finally, section 5 is the conclusion, which summarizes the whole study.

2. LITERATURE SURVEY IN THE FIELD OF RECENT TRENDS

Many researchers have used machine learning algorithms to predict anxiety and depression, such as Random Forest Trees (RFT), Support Vector Machines (SVM), and Convolutional Neural Networks (CNN) Collecting data from blog posts and subsequent ranking. For text encoding, several techniques are used, namely Topic Modeling, Bag of Words (BOW), and Term Frequency-Inverse Document Frequency (TF-IDF). Additionally, modeling experiments using Python programming performed best among the classifiers [2] generated by CNN, with precision and recall of 78% and 0.72, respectively. Different machine learning algorithms such as logistic regression, Catboost, Naive Bayes, RFT and SVM have been applied in [7] for classification. In this study, 470 seafarers were interviewed and participants' occupational, socio-demographic and health information was collected across 16 characteristics, including age, education qualifications, monthly income, employment status, BMI, years of service, family type, marital status, presence of hypertension, diabetes or ischemic heart disease (if applicable), job profile, rank within the organization, job type and dummy variables for education and marital status. As a result, researchers found that Catboost provided the highest level of accuracy and precision of all classifiers - 82.6% and 84.1%, respectively, Xiu et al. (2017) manually collected data on 630, elderly people from medical colleges and hospitals in Kolkata, West Bengal, of whom 520 received special care. After applying different classification methods Bayesian Network, Logistic, Multilayer Perceptron, Naive Bayes, Random Forest, Random Tree, J48, Random Sequential Optimization, Random Subspace and K-Star, they observed that Random Forest is the best accuracy. of 91 gives % and 89% in the two data sets of 110 and 520 people, respectively. For feature selection and classification, the WEKA tool was used in [1]. Today, social media is rapidly evolving into a medical assessment tool for predicting various types of diseases. Saha et al. [8] Selected themes and psycholinguistic features appearing in publications on the LiveJournal website. This was then fed into a joint modeling framework to classify psychological issues found in

online communities interested in depression. The proposed conjoint modeling framework outperforms existing single-task learning (STL) and multi-task learning (MLT) baselines, and studies show that discussions in online communities go beyond emotions depressed. Rees et al. [9] Predictors of depression and post-traumatic stress disorder (PTSD) following Twitter users. Hidden Markov Models (HMM) were used to identify an increased likelihood of PTSD. Of the in the dataset, 31.4% and 24% were observed to be affected by depression and PTSD.

Braithwaite et al. [10] tweets collected from 135 participants recruited from Amazon Mechanical Turk (MTurk) and application of the decision tree classification to measure suicide risk. It has been observed to predict suicide rate with 92% accuracy. Du et al. [11] extract streaming data from Twitter and annotate tweets considered suicidal using psychostressors. Convolutional Neural Networks (CNNs) outperform Support Vector Machines (SVMs) and Supplementary Trees (ETs), among others. It was 78% accurate in identifying suicidal tweets. The audio-text approach was also used to model depression, and the researchers collected data from depressed patients. The LSTM neural network model was used in [12] for depression detection, and context-free models were observed to produce the best results for audio (weighted, sequential, and multi-model). Social media content also predicted early stage depression [13].

Data collection was performed using CLEF eRisk. After evaluating five systems, it was found that a combination of machine learning and information retrieval gave the best results. Hou et al. used a big data approach to predict 4,444 depressions based on a person's reading habits. Chinese text feature extraction developed book classifiers, and after applying five classifications, Naive Bayes was found to be the most suitable [14]. [15] Post-traumatic stress disorder detected using a supervised machine learning classifier. They studied 4,444 former conscripts among the British militants, and the parameters used in their study were alcohol abuse, gender and deployment status. Like results, several supervised machine learning classifiers achieved satisfactory sensitivity, but results were not very sensitive to false negative diagnoses. mood and anxiety disorders were detected in [16] by scanning patients' facial emotions and applying cross-validation, and found more accurate results, which were validated by different statistical measures. Unbalanced classification is applied in [20] and an ensemble machine learning approach is discussed in [21]. Different researchers have applied different machine learning algorithms to predict psychological disorders and found that the performance of different algorithms varies across scenarios; no fixed algorithm has been identified as the most appropriate in all situations. Therefore, in this study, all machine learning

algorithms were applied to identify symptoms of anxiety, depression, and stress.

3. Materials and Methodology

This study focused on the detection of anxiety, depression and stress using the Depression, Anxiety and Stress Scale (DASS 21) questionnaire. Data was collected from a total of 348 participants via Google Forms and then
participants were ranked using five machine learning algorithms, namely Decision Trees, Random Forest Trees, Naive Bayes, support vector machines and KNN.

3.1 Participants

A total of 348 participants, men and women, aged 20 to 60, employed and unemployed, with a wide range of responsibilities ranging from domestic to professional tasks participated in this study.

3.2 Questionnaire

Data for this study were collected from the DASS-21 Depression Scale Questionnaire, of anxiety and stress. The DASS 21 contains 21 questions, and the stress, anxiety, and depression scales have 7 questions each.
The possible answers to each question - which can be given in text or numerical form - are as follows:

0 does not apply to me
1 applies to me somewhat or occasionally.
2 works for me to some extent or most of the time.
3 Applies to me often or most of the time.

After data collection, participants' responses were coded using a numerical value from 0 to 3, then the values associated with each question were summed and a score was calculated as follows:

score Sum of each category's scoring points = *2 (1)

Once the final scores are calculated, they are labeled by level of severity - i.e. Normal, Mild, Moderate, Severe and Extremely Severe.

3.4. Classification

Application of machine learning algorithms to the R programming language using Rstudio version 3.5. This predicts the percentage of people with symptoms of stress, anxiety, and depression at the severity level. The data set is split 70:30, representing the training and test sets respectively. The operating principles of the machine learning algorithms are described in the following subsections.

3.4.1. Decision Tree

The decision tree method of machine learning makes decisions at different levels using tree data structure – this is suitable for predictive problems because they are easy to interpret and the structure is stable. It covers both classification (tree models with a volatile target for a distinct set of values) as well as regression (a volatile target for an endless set of values) [17]. In figure 1 decision tree example is shown. In this a question is divided into yes or no (binary; 2 alternatives) into two branches (yes and no) driving out of the tree. One can get a greater number of choices than 2.

3.4.1. Random Forest

The Random forest classifier creates multiple decision trees from randomly selected subset of training dataset as shown in figure 2. Then it aggregates the votes from different decision trees to decide the final class of test objects [19]. A random forest classification was proposed in [18] with reduced number of trees.

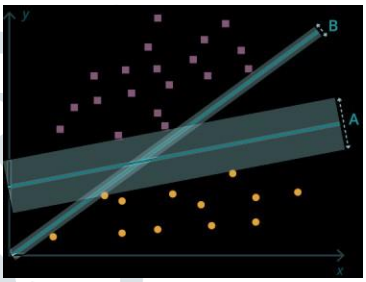
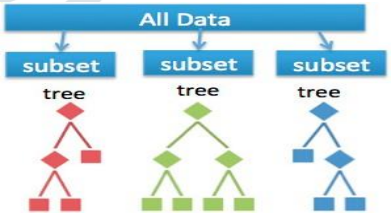


Fig. 2. Random forest

3.4.2. Support Vector Machine (SVM)

A support vector machine [22] is a machine learning algorithm that is suitable for both regression and classification tasks, but is mainly used for classification. This classifier [23] has been used in many applications recently due to the excellent classification ability and presentation



quality of, which linearly divides data into two distinctclasses (also called superclasses), between which the maximum distance is as suggested. In Figure 3.

Fig. 3. Support vector machine representation

3.4.1. Naïve Bayes

This classifier uses Bayes' Theorem supervised machine learning algorithm and assumes that the
features are analytically independent. This claim is based on the naive assumption that the
input factors are independent of each other [24-25]. The formula for naïve berries is:

$$p(H|D)= p(H)p(D|H) / p(D) \quad (2)$$

Where,
p(H | D) = it is later than
p(H) = this is the previous i.e. What do you believe before seeing the evidence
p(D|H) = This is the probability of seeing the evidence if your hypothesis is correct
p(D) = This is the normalization of the evidence in all cases

1.4.1. K- Nearest Neighbour (K-NN)

K-NN is one of the simplest algorithms for classification and regression
problems in machine learning. KNN obtains information based on the nearest metric and ranks the nearest information points. The information is then assigned to the class with nearest neighbors. The diagram is shown in Figure 4.
KNN [26] is often used to classify future information due to its ease of implementation and suitability.As the picture shows.

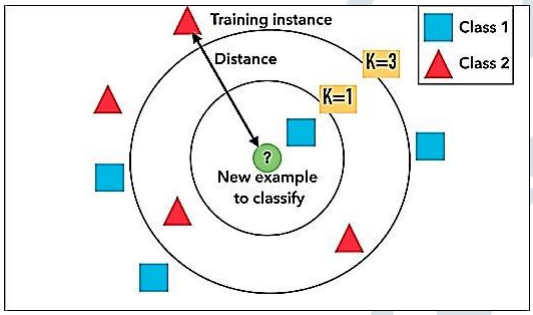


Fig. 4. K- nearest neighbour representation

4. Results and discussions

Application of the five methods, viz.
Decision Trees (DT), Random Forest Trees (RFT), Naive Bayesian (NB), Support Vector Machines (SVM) and K Nearest Neighbors (KNN) - the three categories of stress, anxiety and depression, resulting in Table 3 confusion matrix described in. The rows of the matrix display the actual class, while the columns display the predicted class. Numbers 1, 2, 3, 4 and 5 in rows and columns represent normal, mild, moderate, severe and very severe cases respectively. Equations 3 through 8 below are used to calculate precision and error rates, precision, recall, and specificity in each confusion matrix.

Accuracy Rate = Sum of diagonals (TP) / Total number of instances

Error Rate = 1 – Accuracy Rate

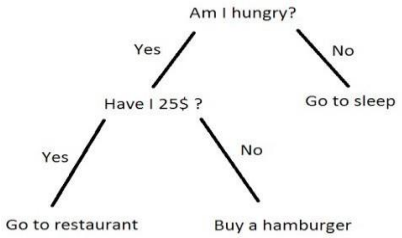
Precision = TP / (TP + FP)

Recall = Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

F1 Score = 2 *(Precision * Recall) / (Precision + Recall)

Whereas,
TP (True positive) = diagonal of the matrix
FN (False negative) = sum of the rows consistent with the class (excluding TP for this class)
FP (False positive) = sum of the columns corresponding to the class (class TP excluding this class)



TN (True negative) = sum of all rows and columns (rows and columns excluding this class)

According to the observations, Naive Bayes achieved the highest accuracy for all three scales of anxiety, depression, and stress. Nevertheless, the results in Table 3 show that the classes are unbalanced, since the confusion matrices for anxiety, depression and stress yielded 25, 37 and 43 cases respectively, of which were normal but 7, 12 and 19 a minor case. Again, 35, 25, and 23 moderate; 11, 11 and 16 severe; and 27, 19, and 4 extremely severe for the anxiety, for depression and stress scales; the classes here are therefore unbalanced. In this case, accuracy alone is not enough of a metric, and the f1 score becomes an important metric for determining the best model. The f1 score is the harmonic mean of precision and recall, with higher precision and recall having higher values.
Therefore, in case of class imbalance, the best model is the one with the highest f1 score even though it has a lower precision of. Random forests for stress and Naive Bayes had the highest f1 scores for depression and the lowest for all algorithms for anxiety. In all three cases, the specificity of all algorithms was found to be around 90% or greater. Moreover, it is an important parameter in healthcare because it shows that negative cases (patients without disease) were also correctly classified. All the algorithms applied in this study also produced very accurate results for negative cases.

(a)

Variable importance	
Scared_without_any_good_reason	26
Breathing_difficulty	20
Close_to_panic	19
Aware_of_the_action_of_heart_in_the_absence_of_physical_exertion	15
Dryness_of_mouth	15
Experienced_trembling	5

(b)

Variable importance	
Life_was_meaningless	19
Nothing_to_look_forward_to	18
Down_hearted_and_blue	18
Wasnt_worth_much_as_a_person	16
Unable_to_become_enthusiastic_about_anything	14
could_nt_experience_any_positive._feeling	11
Difficult_to_work_up_the_initiative_to_do_things	5

(c)

Variable importance			
Difficult_to_relax	23	Getting_agitated	20
I_felt_that_I_was_rather_touchy	20	Hard_to_wind_down	13
nervous_energy	12	Intolerant_to_getting_what_I_was_doing	12

Fig. 5. Variable importance (a) Anxiety (b) Depression (c) Stress.

Figure 5 shows the varying importance ratings for anxiety, depression, and stress, respectively; the higher the number, the more important the variable. That is, the "Fear_without_any_good_reason" variable turned out to be and was the most important on the anxiety scale; the variable "Life_was_useless" was found to be significant on the depression scale at , while "Difficult_to_relax" was found to be significant on the stress scale.

4. Conclusion

In this article, a machine learning algorithm was applied to identify five different levels of severity for anxiety, depression and stress. Data were collected using a standard questionnaire measuring general symptoms of anxiety, depression and stress (DASS-21) in subjects. Five different classification techniques were then applied - Decision Trees (DT), Random Forest Trees (RFT), Naive Bayes, Support Vector Machines (SVM) and K-Nearest Neighbors (KNN). Naive Bayes accuracy was found to be the highest, although Random Forest was identified as the best model. Since this problem produces unbalanced classes, the best model selection is made based on the f1 score, which is used in the case of unbalanced partitions. For the anxiety, depression and stress scales, the significant variables were found to be 'scared_for_no_good_reason', 'Life_was_pointless' and 'Difficult_to_relax' respectively. Therefore, these variables are considered the most important for detecting psychological disorders.

References

[1] Sau, A., Bhakta, I. (2017)"Predicting anxiety and depression in elderly patients using machine learning technology."*Healthcare Technology Letters* **4 (6)**: 238-43.

[2] Tyshchenko, Y. (2018)"Depression and anxiety detection from blog posts data."*Nature Precis. Sci., Inst. Comput. Sci., Univ. Tartu, Tartu, Estonia.*

[3] https://adaa.org/understanding-anxiety/depression/symptoms

[4] https://www.webmd.com/balance/stress-management/stress-symptoms-effects_of-stress-on-the-body#1

[5] Oei, T. P., Sawang, S., Goh, Y. W., Mukhtar, F. (2013) "Using the depression anxiety stress scale 21 (DASS-21) across cultures." *International Journal of Psychology* **48 (6):** 1018-1029.

[6] Kroenke, K., Spitzer, R. L.,Williams, J. B. (2001) "The PHQ-9: validity of a brief depression severity measure."*Journal of generalinternalmedicine* **16 (9)**: 606-613.

[7] Sau, A., Bhakta, I. (2018) "Screening of anxiety and depression among the seafarers using machine learning technology."*Informatics in Medicine Unlocked* :100149.

[8] Saha, B., Nguyen, T., Phung, D., Venkatesh, S. (2016) "A framework for classifying online mental health-related communities with an interest in depression."*IEEE journal of biomedical and health informatics* **20 (4)**: 1008-1015.

[9] Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., Langer, E. J.(2016) "Forecasting the Onset and Course of Mental Illness with Twitter Data."*Scientific reports* **7 (1)**: 13006.

[10] Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., Hanson, C.L. (2016) "Validating machine learning algorithms for Twitter data against established measures of suicidality."*JMIR mental health* **3 (2)**: e21.

[11] Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., Xu, H. (2018) "Extracting psychiatric stressors for suicide from social media using deep learning."*BMC medical informatics and decision making* **18 (2)**: 43.

[12] Al Hanai, T., Ghassemi, M. M., Glass, J.R. (2018) "Detecting Depression with Audio/Text Sequence Modeling of Interviews." *InInterspeech* : 1716-1720.

[13] Ramiandrisoa, F., Mothe, J., Benamara, F., Moriceau, V. (2018) "IRIT at e-Risk 2018."*E-Risk workshop*: 367-377.

[14] Hou, Y., Xu, J., Huang, Y., Ma, X. (2016) "A big data application to predict depression in the university based on the reading habits."*3rd IEEE International Conference on Systems and Informatics (ICSAI)*: 1085-1089.

[15] Leightley, D., Williamson, V., Darby, J., Fear, N.T. (2019) " Identifying probable post-traumatic stress disorder: applying supervised machine learning to data from a UK military cohort." *Journal of Mental Health*. **28 (1)**: 34-41.

[16] Young, C., Harati, S., Ball, T., Williams, L. (2019) "Using Machine Learning to Characterize Circuit-Based Subtypes in Mood and Anxiety Disorders."*Biological Psychiatry* **85 (10)**: S310.

[17] Li, L., Zhang, X. (2010) "Study of data mining algorithm based on decision tree." *In2010 International Conference On Computer Design and Applications IEEE* **1**: V1-155.

[18] Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., Kundu, S. (2018)"Improved random forest for classification."*IEEE Transactions on Image Processing* **27 (8)**: 4012-4024.

[19] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J. P. (2012) "An assessment of the effectiveness of a random forest classifier for land-cover classification."*ISPRS Journal of Photogrammetry and Remote Sensing* **67**: 93-104.

[20] Liu, X. Y., Wu, J., Zhou, Z. H. (2009)"Exploratory undersampling for class-imbalance learning."*IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **39 (2)**: 539-50.

[21] Dietterich, T. G. (2000) "Ensemble methods in machine learning."*InInternational workshop on multiple classifier systems*Springer, Berlin, Heidelberg: 1-15.

[22] Saitta, L., (2000) "Support-Vector Networks." *Kluwer Acad. Publ. Bost.* : 273–297.

[23] Hamed, T., Dara, R., Kremer, S. C. (2014) "An accurate, fast embedded feature selection for SVMs." *In2014 13th International Conference on Machine Learning and Applications*IEEE :135-140.

[24] Martinez-Arroyo, M., Sucar, L. E. (2006) "Learning an optimal naive bayes classifier."*In18th International Conference on Pattern Recognition (ICPR'06)* IEEE **3**: 1236-1239.

[25] Cheng, J., Greiner, R. (1999) "Comparing Bayesian network classifiers."*InProceedings of the Fifteenth conference on Uncertainty in artificial intelligence*Morgan Kaufmann Publishers Inc : 101-108.

[26] Taneja, S., Gupta, C., Goyal, K., Gureja, D. (2014) "An enhanced k-nearest neighbor algorithm using information gain and clustering."

Fourth International Conference on Advanced Computing & Communication Technologies IEEE : 325-329.