ANALYTICS OF BEHAVIOURAL HEALTH CONDITIONS USING MACHINE LEARNING AND DEEP LEARNING

[1]Wilona Maria Furtado, [2]Shantala Devi Patil

School of Computer Science and Engineering, *REVA University*
[1]wilonafurtado2@gmail.com, [2]shantaladevi.patil@reva.edu.in

Abstract.

In this paper, we have collected thebehavioural health conditions dataset of normal and depressive comments on which NLP techniques like tokenization, lemmatization are applied on text to get clean text using the NLTK toolkit. The words with most frequency are visualized through word cloud matplotlib package. The dataset is balanced using oversampling technique on which TF-IDF is applied. Machine learning techniques like Random Forest classifier, XG Boosting and deep learning techniques like ANN are implemented where Confusion matrix, Accuracy, F1-Score and ROC are used to view the performance of the model. The model will be deployed using FLASK python where a user will enter their condition in the form of text where the model will detect and suggest if the person is normal or suffering from depression and needs to visit a counsellor. Theperformance of the modelswas analysed using accuracy and error.Random Forest Classifier model has accuracy of 99.83% which is better compared to XG Boosting and ANN.

Keywords. Tokenization, Lemmatization, Oversampling, TF-IDF, Random Forest Classifier, XG Boosting, ANN, Confusion Matrix, Accuracy, F1 score, precision, recall, classification report, FLASK python

1. INTRODUCTION

Mental illness, usually referred to as mental health diseases, is a condition that affects one's mood, though process, and the way they behave. Mental illnesses include depression, anxiety and eating disorders, addiction and schizophrenia.A mental health condition becomes a mental illness when unattended symptoms create stress and hamper one's ability to function.Mental illness can make you feel awful, causing problems at home, at work, and in your relationships.Medication and counselling might be used to treat symptoms.

Feeling depressed, having trouble concentrating, having intense emotions of guilt, having dramatic mood fluctuations, withdrawing from friends and hobbies, and having suicidal thoughts are all signs and symptoms of mental illness. Inherited features, prenatal exposure to the environment, and brain chemistry are all potential causes of mental illness.The risk factors of mental illness are a history of mental illness in a blood relative, tough situations in life, chronic medical condition, childhood history of abuse, use of alcohol or recreational drugs and previous mental illness.Mental illness can be prevented by paying attention to warning signs, go for regular checkups for medical care and top it all by taking good care of oneself.

It is quite a debatable topic to come to a conclusion if a person is suffering from depression or not. Hence, we train the model with the behavioural health conditions dataset by using ML and DL techniques. The model will help in predicting if the person is normal or suffering from depression. This will help the person to go and get some help at the earliest leading to reduction of suicidal cases.

The following is the layout of the paper. The literature survey is discussed in Section II. NLP, ML, and DL approaches employed in the paper are discussed in Section III. Section IV presents the proposed work. In Section V, you'll see the results. Section VI concludes with a discussion of the future scope.

2. LITERATURE REVIEW

Anamika Ahmed et al. [1] developed a model that integrated traditional psychological tests with machine learning algorithms to detect the severity of mental illnesses at various levels.Two datasets, anxiety and depression, were used to test CNN, SVM, LDA, K-NN Classifier and Linear Regression. The CNN algorithm was used to achieve accuracy of 96% for anxiety and 96.8% for depression.Data is processed by Rahul Katarya et al., [2] to uncover the features that influence employee mental health which could be personal or professional. To discover the model with the highest accuracy, ML techniques such as SVM, KNN, Regression, DecisionTree and Random Forest are used. Decision tree classifier had the best performance with 84% and 83 precision whereas KNN had the worst.MS. Purude Vaishali Narayanrao et al., [3] used questionnaires, social media posts, verbal communication text, and facial expressions to collect data. Students in high school, college and working professionals were the target categories for identification with the findings indicating whether or not the person need assistance. Machine learning algorithms and classifiers such as Decision Tree, SVM, Naïve Bayes Classifier, Logistic Regression, and KNN Classifier are used to detect the mental state in a targeted group. Using the Twitter scraping tool Twint, the tweet is classified as depressive or not.Yara E. Alharahsheh et al. [4] employed a strong, reliable supervised ML classifier with the best performance to predict whether or not a person in Kenya is likely to suffer from depression. SVM, Random Forest, Ada Boosting, and voting ensemble methods had the greatest f1-score 0.78 and 85% accuracy while Naïve Bayes, Logistic Regression, Decision Tree, Gradient Boosting, Bagging, XGBoost, and Stack methods were also applied.Nur E Jannat Asha et al., [5] have developed a low-cost heart rate monitoring system based on sensors and IoT devices. The sensor will be attached to the finger and the colour variation will be chosen when the interval is monitored. The signal is processed using an Arduino microcontroller and devices are used to track blood. CSV files are used to store the Arduino-measured heart rate data. Heart rates are recorded and emotions are classified as positive, negative, or neutral using the Geneva affective picture database. SVM with polynomial kernel, a machine learning method, is used to predict mental stress from heart rate data exhibiting the best accuracy.

Using machine learning, Carlos Alfonso V. Palattao et al. [6] discovered factors leading to stress, depression, and anxiety in the Philippines population. To measure mental health, the data from 2119 participants who responded to an online survey was analysed using feature selection methods and ML classifiers such as Random Forest, Naïve Bayes, SVM and Logistic Regression.Data from electronic health records was used to create a tagged list of words relevant to the disease matched against symptoms of psychological disorders for prediction by YamuAryal et al., [7]. The output of machine learning models is compared to the prediction of psychological disorder based on fMRI and PET images collected from the patient's EHR. In order to process complicated mental health data, artificial neural networks and machine learning are used.V. Uday Kumar et al., [8] collected data from working people and asked them a variety of questions in order to determine despondency. In comparison to SVM and Decision Tree, the dataset is passed through ML algorithms, with Random Forest providing the highest accuracy of 87.02%.ML and DL techniques were used by Pramod Bobade et al., [9], to detect stress in individuals using a multimodal dataset collected from wearable psychological and motion sensors in order to prevent stress-related health problems. Sensor data such as ACC, BVP, ECG, TEMP, RESP, EDA and EMG are utilised to determine three psychological states: amusement, neutral and stress. Machine learning algorithms such as K-NN, LDA, Random Forest, Decision Tree, AdaBoost, and Kernel Support Vector Machine were used to evaluate and compare the accuracies of all three classes and binary classification. The accuracies for three class and binary classification were 81.65% and 93.20%, respectively. The results for the simple feed forward deep learning artificial neural network were 84.32% and 95.21%, respectively.Sangeeta R.Kamite et al., [10] where a system capable of analysing syntactical markers related to onset and perpetual symptoms of depression was developed. Algorithm was developed to help in prediction of depression in an effective manner with an approach that syntactical markers used in tweets were used to frame statistical model in depression prediction. Random forest achieved 99.89% accuracy.

Payel Bhattacharjee et al., [11] where sedentary, sleep or rest behaviour of healthy adult with aid of physical activity was analysed. A relation is obtained between parameters affecting sleep and sedentary behaviour with the physiological signals obtained from commercial wearable devices. The techniques employed for analysis were random forest, XGBoost, SVM, and K-NN.Apple watch and Fitbit are the commercial wearables used for data analysis. XGBoost provides best accuracy.Kuhaneswaran A/L Govindasamy et al.,

[12] where the users depression was detected using their social media data. The sentiment score was determined using sentiment analysis, which classified it as positive, negative, or neutral. Labeled tweets were fed into ML algorithms, with Naïve Bayes and a hybrid mode, NBTree, both providing 97.31% accuracy.Tahmid Hasan Sakib et al., [13] where a person's tweet is analysed whether it has suicidal intention or not using machine learning. Various sets of word embedding and tweet features were used along with comparison with models like Voting Classifier, CatBoost Classifier, XGBoost Classifier, Gradient Boosting Classifier, Logistic Regression, Bagging Classifier, Multi-layer Perceptron, Decision Tree Classifier, SVM, AdaBoost Classifier, K-Nearest Neighbor and Naïve Bayes Classifier.Shivangi Yadav et al., [14], used a routine survey in which people were asked about their home and work situations, as well as their family history of mental illness.K-NN, Decision Tree, Multinomial Logistic Regression, Random Forest Classifier, Bagging, Boosting, and Stacking were among the algorithms employed to predict depression in humans. Best performance was by Boosting with 81.75% accuracy.Faisal Muhammad Shah et al., [15], suggested a hybrid algorithm to detect depression using textual posts from users. The reddit dataset was used to train and test DL models. Bidirectional Long Short Term Memory (BiLSTM) was proposed, along with a variety of word embedding algorithms and metadata elements.Word2VecEmbed+Meta features performed well.

In all the above research papers, they have not deployedall these models for any user use in a simple way by creating a user interface to capture a comment from the user about their state of mind.The comparison between the base paper [1] and this paper is that the base paper made use ofalgorithms like CNN, SVM, LDA, K-NN Classifier and Linear Regressionand achieved highest accuracy of 96% and 96.8% on dataset of anxiety and depression respectively using CNN. This paper made use of techniques like tokenization, lemmatization, oversampling, TF-IDF, Random Forest, XG Boosting, ANN and finally deployed using FLASK python to provide a GUI for the end users and the highest accuracy achieved is 99.83% using Random Forest.

3. NATURAL LANGUAGE PROCESSING, MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

NLP techniques help machines to understand the language humans communicate in by breaking it down. Machine learning methods are used to train the model and perform prediction on unseen data. Ensemble methods aid in learning from multiple weak models and take a decision on the output. Bagging makes use of parallel learning while boosting uses sequential learning. Deep learning algorithms imitate the structure of the human brain and functionalities in the aim to make machines intelligent enough to do complicated tasks.

A. Tokenization

Tokenization is a technique for breaking down large amounts of text into small chunks such as words or sentences called as tokens.Tokens aid in the comprehension of context or the development of an NLP model.By analysing the word sequence, tokenization aids in determining the meaning of the text. The commonalgorithm used for tokenization is Word Tokenization where a piece of text is separated into individual words.

B. Lemmatization
Lemmatization is a stemming-like technique that gives context to words by associating words that have similar meanings to one another.Lemmatization is the process of getting rid of inflectional endings from a word and returning it to its base form known as lemma by making use of vocabulary and morphological analysis.

C. Tf-idf
Term frequency-inverse document frequency, a text vectorizer which convertstext into a vector that may be used. The frequency of each word present in the document is indicated by Term Frequency (TF) while the relevance of the word in the document is indicated by Inverse Document Frequency (IDF).

D. Oversampling
Random oversampling is the most basic form of oversampling method used to balance an imbalanced dataset. The minority class samples are duplicated in order to balance the data. Although this method does not result in loss of data, the dataset is prone to overfitting due to the repetition of the same information.

E. Random Forest
Random forest takes a combination of multiple decision tree models and does the classification based on majority votes of prediction and predicts the final output. This method is also known as bootstrapping where the number of rows and columns from the dataset can be chosen.

F. XG Boosting (Extreme Gradient Boosting)
Gradient Boosted Decision Trees are implemented in XGBoost. In XGBoost, decision trees are built sequentially with weights playing a significant role. All independent variables are given weights which are then fed into decision trees to predict results. The weight of variables that the tree predicts incorrectly is increased and fed to the following decision tree. Individual classifiers are then combined to produce a more accurate model.

G. ANN (ArtificialNeutral Network)
ANN borrows the concept from a biological neural network. To facilitate communication between units, there is a large collection of units connected to the pattern. Units are called nodes or neurons which act as simple processors that operate in parallel. Each neuron is associated with a weight and is connected to another neuron by a link that conveys the input signal information. An activation signal is an internal state that each neuron has.The output signal is formed by combining the input signal supplied to other units with the activation rule.

4. PROPOSED WORK

As shown in Fig.6, we take the behavioural health conditions dataset having a message and label on which we apply tokenization, lemmatization to get clean text. This imbalanced dataset is balanced using oversampling technique followed by TF-IDF. 75% data is used for training while 25% data is used for testing. Model is trained for machine learning algorithms like Random Forest Classifier, XG Boosting and deep learning methods like ANN. Using confusion matrix, accuracy, f1-score, ROC curve we view the performance of the model. A web application is developed using FLASK python to provide a graphical user interface to the user to enter their comment and the system will state if it's a normal or depressive comment by giving the input to the model on server side. Hence stating if they are required to pay a visit to the counsellor / psychiatrist or not.

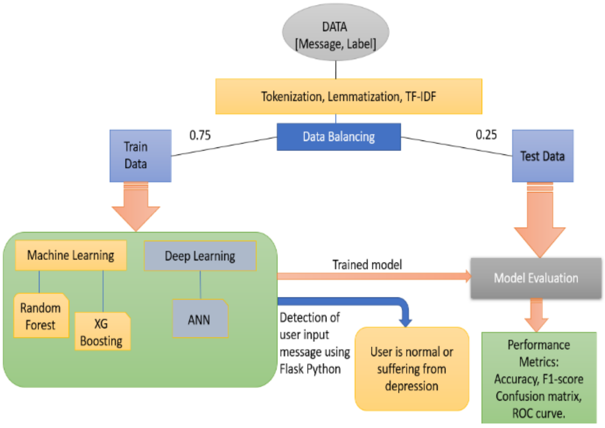

DATA
[Message, Label]
Tokenization, Lemmatization, TF-IDF
0.75
Data Balancing
0.25
Train Data
Test Data
Machine Learning
Deep Learning
Random Forest
XG Boosting
ANN
Trained model
Model Evaluation
Detection of user input message using Flask Python
User is normal or suffering from depression
Performance Metrics: Accuracy, F1-score Confusion matrix, ROC curve.

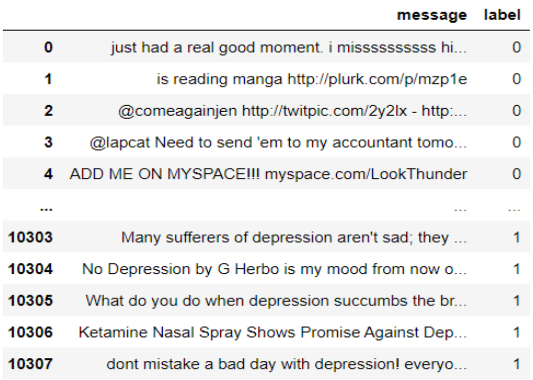
Fig.1. Flow Chart of Proposed Model

5. IMPLEMENTATION AND RESULTS

IMPLEMENTATION:

The behavioural health conditions dataset was used. Using Jupyter Notebook we imported all the necessary packages like pandas, numpy, matplotlib and seaborn. We read the csv file and performed necessary operations. Data is explored using seaborn where we plot the count of target variable using countplot(). If label==0 then it's a normal comment else if label==1 it's a depressive comment. Plot WordCloud after removing stopwords for each normal and depressive comments. Import re, nltk, stopwords, string and WordNetLemmatizer packages to perform NLP techniques. A function process_text() is defined where urls, mentions and punctuation marks are removed and finally we apply lemmatization on the clean words. RandomOverSampler, Counter and TfidfVectorizer packagesare used to balance data using Over Sampling technique on which we apply Tf-idf to clean text. The train_test_split package divides data into train and test segments, with 75% of the data used to train the model and 25% used to test it. Machine learning techniques like Random Forest and XG Boosting and deep learning techniques like ANN are used to train the model and predict test data by importing RandomForestClassifier, XGBClassifier, Sequential and layers packages.Performance is evaluated using confusion_matrix, ConfusionMatrixDisplay, classification_report, accuracy_score and roc_curve packages. Best model is saved by importing pickle package. The model is deployed using FLASK python by importing Flask, render_template and request packages to create a graphical user interface for the user to enter their comment to test if the comment is normal or depressive.

RESULTS:



	message	label
0	just had a real good moment. i missssssssss hi...	0
1	is reading manga http://plurk.com/p/mzp1e	0
2	@comeagainjen http://twitpic.com/2y2lx - http:...	0
3	@lapcat Need to send 'em to my accountant tomo...	0
4	ADD ME ON MYSPACE!!! myspace.com/LookThunder	0
...	...	...
10303	Many sufferers of depression aren't sad; they ...	1
10304	No Depression by G Herbo is my mood from now o...	1
10305	What do you do when depression succumbs the br...	1
10306	Ketamine Nasal Spray Shows Promise Against Dep...	1
10307	dont mistake a bad day with depression! everyo...	1

Fig.2. Data in the Dataset

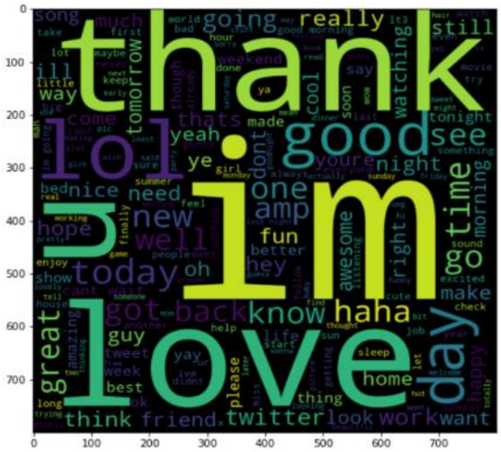


Fig.3. Words frequency in Normal Comments

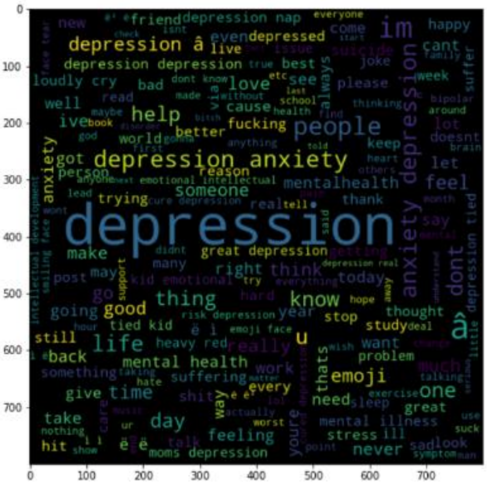


Fig.4. Words frequency in Depressive Comments

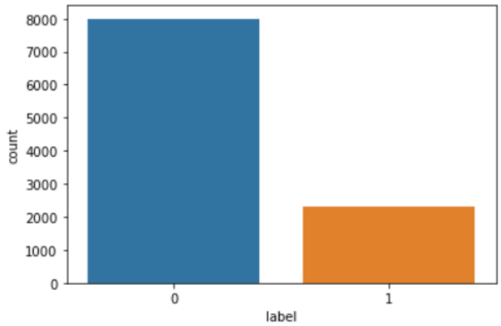


Fig.5. Illustration of a count plot of label column

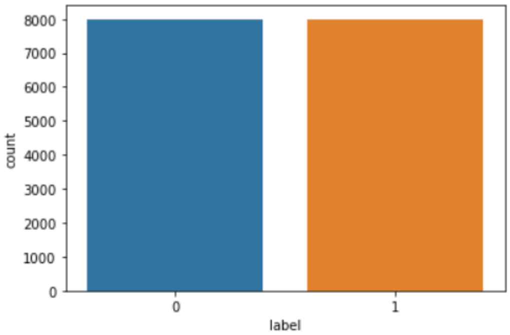
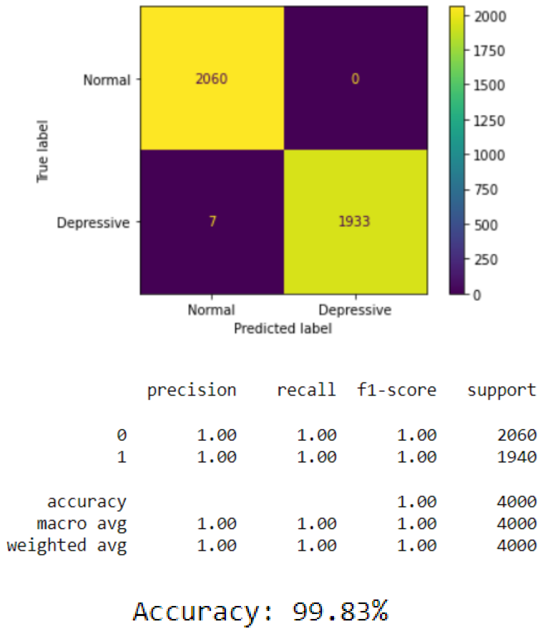


Fig.6. Balanced Data



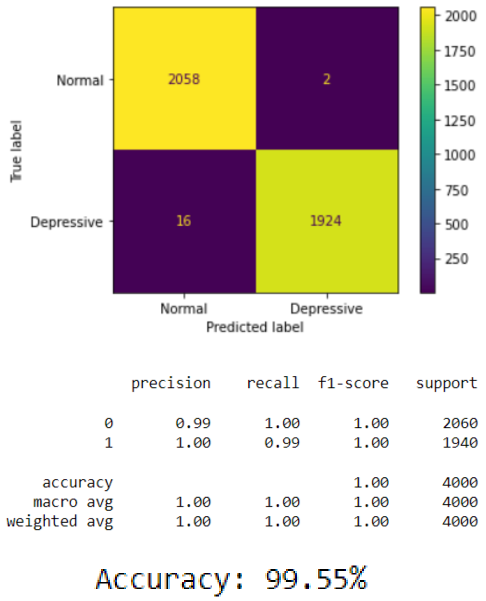
```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2060
           1       1.00      1.00      1.00      1940

    accuracy                           1.00      4000
   macro avg       1.00      1.00      1.00      4000
weighted avg       1.00      1.00      1.00      4000

Accuracy: 99.83%
```

Fig.7. Random Forest Classifier



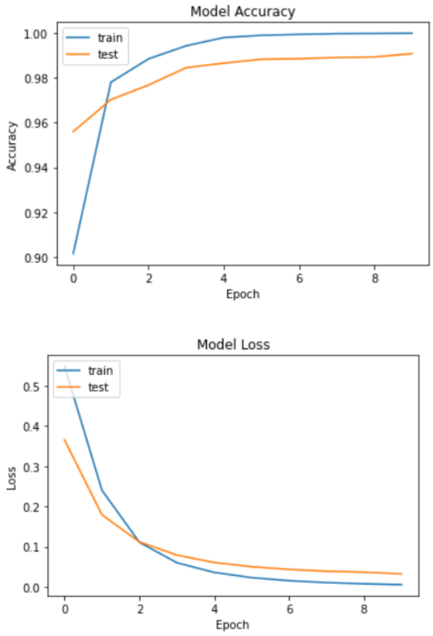
```
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      2060
           1       1.00      0.99      1.00      1940

    accuracy                           1.00      4000
   macro avg       1.00      1.00      1.00      4000
weighted avg       1.00      1.00      1.00      4000

Accuracy: 99.55%
```

Fig.8. XG Boosting



```
Testing Accuracy: 99.0750
```

Fig.9. ANN

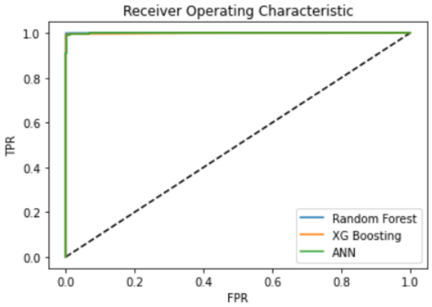
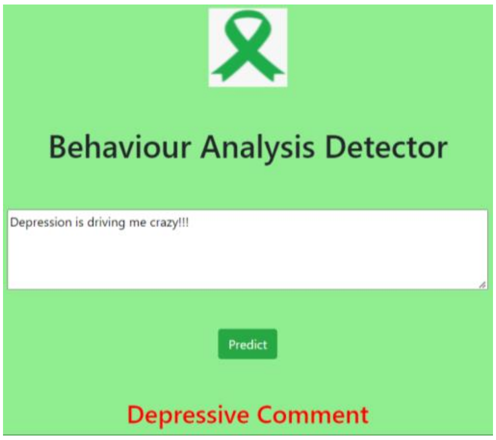


Fig.10. ROC Curve

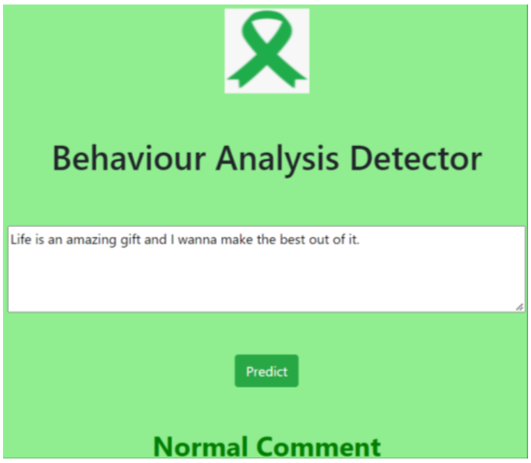


Behaviour Analysis Detector

Depression is driving me crazy!!!

Depressive Comment

Fig.11. User enters a Depressive comment



Behaviour Analysis Detector

Life is an amazing gift and I wanna make the best out of it.

Normal Comment

Fig.12. User enters a Normal comment

6. CONCLUSION& FUTURE SCOPE

Behavioural health conditions affect a person's mood, thinking which causes Depression, Anxiety, etc. By using some machine learning and deep learning approaches aids in improved model training. The dataset considered has normal and depressive comments. NLP techniques like tokenization and lemmatization are applied on text to get clean text. The dataset is balanced using oversampling technique on which we apply TF-IDF. Random Forest Classifier gave the best accuracy. Finally, the model is deployed using FLASK python for the user to easily interact with this system. This recommendation system will suggest to the person if they are normal or suffering from depression and need to visit a counsellor / psychiatrist which helps in early detection of behavioural health conditions reducing drastic steps like suicide.

In future work,the dataset could be created through surveys taken from people of diverse backgrounds. This system could be linked with counsellors such that if the person shows signs of depression, then the counsellor will get alerted and can contact the patient.

7. REFERENCES

[1] A. Ahmed, R. Sultana, M. T. R. Ullas, M. Begom, M. M. I. Rahi and M. A. Alam, "A Machine Learning Approach to detect Depression and Anxiety using Supervised Learning," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411642.

[2] R. Katarya and S. Maan, "Predicting Mental health disorders using Machine Learning for employees in technical and non-technical companies," 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering(ICADEE), 2020, pp. 1-5, doi: 10.1109/ICADEE51157.2020.9368923.

[3] P. V. Narayanrao and P. Lalitha Surya Kumari, "Analysis of Machine Learning Algorithms for Predicting Depression," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1-4, doi: 10.1109/ICCSEA49143.2020.9132963.

[4] Y. E. Alharahsheh and M. A. Abdullah, "Predicting Individuals Mental Health Status in Kenya using Machine Learning Methods," 2021 12th International Conference on Information and Communication Systems (ICICS), 2021, pp. 94-98, doi: 10.1109/ICICS52457.2021.9464608.

[5] N. E. J. Asha, Ehtesum-Ul-Islam and R. Khan, "Low-Cost Heart Rate Sensor and Mental Stress Detection Using Machine Learning," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), 2021, pp. 1369-1374, doi: 10.1109/ICOEI51242.2021.9452873.

[6] C. A. V. Palattao, G. A. Solano, C. A. Tee and M. L. Tee, "Determining factors contributing to the psychological impact of the COVID-19 Pandemic using machine learning," 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021, pp. 219-224, doi: 10.1109/ICAIIC51459.2021.9415276.

[7] Y. Aryal, A. Maag and N. Gunasekera, "Application of Machine learning algorithms in diagnosis and detection of psychological disorders," 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), 2020, pp. 1-10, doi: 10.1109/CITISIA50690.2020.9371801.

[8] V. U. Kumar, A. Savithri, M. J. Bhavani, A. M. Priya, K. V. S. B. Jahnavi and N. D. N. Lakshmi, "Finding Psychological Instability Using Machine Learning," 2020 7th International Conference on Smart Structures and Systems (ICSSS), 2020, pp. 1-4, doi: 10.1109/ICSSS49621.2020.9202009.

[9] P. Bobade and M. Vani, "Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data," 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 51-57, doi: 10.1109/ICIRCA48905.2020.9183244.

[10] S. R. Kamite and V. B. Kamble, "Detection of Depression in Social Media via Twitter Using Machine learning Approach," 2020 International Conference on Smart

Innovations in Design, Environment, Management, Planning and Computing (ICSIDEMPC), 2020, pp. 122-125, doi: 10.1109/ICSIDEMPC49020.2020.9299641.

[11] P. Bhattacharjee, S. P. Kar and N. K. Rout, "Sleep and Sedentary Behavior Analysis from Physiological Signals using Machine Learning," 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2020, pp. 240-244, doi: 10.1109/ICIMIA48430.2020.9074883.

[12] K. A. Govindasamy and N. Palanichamy, "Depression Detection Using Machine Learning Techniques on Twitter Data," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 960-966, doi: 10.1109/ICICCS51141.2021.9432203.

[13] T. H. Sakib, M. Ishak, F. F. Jhumu and M. A. Ali, "Analysis of Suicidal Tweets from Twitter Using Ensemble Machine Learning Methods," 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021, pp. 1-7, doi: 10.1109/ACMI53878.2021.9528252.

[14] S. Yadav, T. Kaim, S. Gupta, U. Bharti and P. Priyadarshi, "Predicting Depression From Routine Survey Data using Machine Learning," 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, pp. 163-168, doi: 10.1109/ICACCCN51052.2020.9362738.

[15] Ahmed, S. T. (2017, June). A study on multi objective optimal clustering techniques for medical datasets. In 2017 international conference on intelligent computing and control systems (ICICCS) (pp. 174-177). IEEE.

[16] Nagashree N, Premjyoti Patil, Shantakumar Patil, Mallikarjun Kokatanur, “Performance Metrics for Segmentation Algorithms in Brain MRI for Early Detection of Autism”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2S, December 2019.

[17] N. Nagashree, Premjyoti Patil, Shantakumar Patil, and Mallikarjun Kokatanur, “InvCos Curvature Patch Image Registration Technique for Accurate Segmentation of Autistic Brain Images”, Springer Nature Singapore Pte Ltd. 2022 V. S. Reddy et al. (eds.), Soft Computing and Signal Processing, Advances in Intelligent Systems and Computing 1340, https://doi.org/10.1007/978-981-16-1249-7_62, pp 659-666.