IJPRSE
Progressive Research

International Journal of Progressive Research in Science and Engineering
Volume-1, Issue-5, August-2020
www.ijprse.com

Screening Depression in IT Industry Using Machine Learning

Sushma Koushik N.[1], Pareekshith US Katti[2], Ganesh Manu Mahesh Kashyap[2], Sanjay R Rao[2], Jitendra Kumar Mahto[2]

[1]*Associate Professor, Department of Computer Engineering, Maharaja Institute of Technology Mysore, Mandya, India.*

[2]*Student, Department of Computer Engineering, Maharaja Institute of Technology Mysore, Mandya, India.*

Corresponding Author: pareekshithk@gmail.com

Abstract: - Depression is at an all-time high. In some countries like the US, depression hasn't been this high since the Second World War. This is a call for urgent action – improvements in diagnosis and treatment of depression need to be brought about. Diagnosis is the first step in addressing depression. Our research revealed that current methods are insufficient in this regard, both traditional and electronic approaches. We wanted to introduce a considerably better depression detection system, for the IT industry in particular to start off with. The best electronic method at present is the questionnaire approach, which has achieved an accuracy of around 81%. Other e-methods like face recognition and sentiment analysis are infeasible in terms of accuracy and/or ease of implementation. We used the OSMI 2018 dataset to train and compare 9 machine learning models while using the same questionnaire approach. Random Forest turned out to be the best, with an accuracy of around 96%. We have thus arrived at a highly accurate 15-question quiz that helps determine if a tech employee has depression, based on his condition in the workplace and family. We hope to test the approach on site and expand it from the IT industry to the general public, ideally causing an improvement in mental health in the population and drive humanity towards better mental healthcare.

Key Words — ***Machine Learning, Random Forest, Target Encoding, Random Search, Mental Health.***

I. INTRODUCTION

Depression is a mental illness primarily characterized by sadness and loss of interest in hobbies and daily activities which persists for a period for two weeks or higher. Additionally, depression may cause people to have a loss of appetite, disruptions to sleep and energy levels, difficulty maintaining focus and concentration, have feelings of hopelessness, and self-harm or suicidal thoughts and tendencies. Depression is treatable with talking therapy. Talking therapy or Psychotherapy is a psychology-based depression treatment approach that involves a licensed professional counsellor talking face to face with the patient and bringing the patient to a state of balance, clarity, productivity and hope. Based on the latest data we have; global suicide rates have been high the past few years. In some countries like the US, suicide rates haven't been this high since the Second World War. The WHO estimates that each year, approximately one million people die from suicide, representing a global mortality rate of one death every 40 seconds. Just like physical illnesses, mental illnesses like depression don't discriminate based on factors like fame, wealth, or status, and nobody can be ruled out as being immune to them. But despite this, arguably, mental health isn't taken as seriously as physical health. The WHO says that the availability and quality of data on suicide and suicide attempts is poor, which is a testament to this. Our project is meant to be a part of the progress of humanity towards better mental healthcare. We use a machine learning program which poses questions, and determines the presence and level of depression based on the answers they give. The program also serves as a data collection platform to support the drive for more data on mental health. Since there's no human being paid for their time and presence, it's free. Also, it's a cold, non-judgmental software interface that the user interacts with – an anonymous Q&A platform. For these reasons, in theory, our solution is more accessible than therapy sessions or other conventional methods for depression diagnosis. Machine learning provides a better ability to upscale, upgrade and obtains results than hard coded algorithms. A machine learning model is an entity that understands the problem – this is obviously better for non-deterministic, real world problems like depression, compared to a pre-programmed system that can do nothing but go by the book. Intuitively, ML is the right approach for this problem, and we have made use of the same.

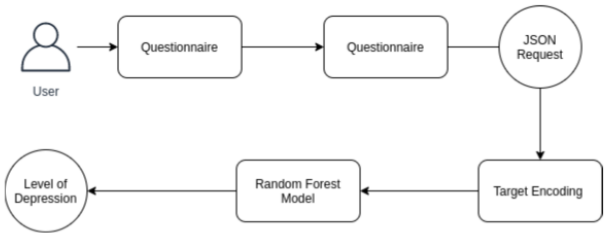
II. RELATED WORK

Among the different computer based approaches to detect depression, the prominent approaches we found were questionnaire, sentiment analysis and facial analysis. Questionnaire based approaches seen in papers "Machine

Learning Techniques for Stress Prediction in Working Employees" and "Prediction of Mental Disorder for employees in IT Industry" used boosting algorithms to achieve 75% and 81.7% accuracy respectively. Sentiment analysis using neural network seen in "Detecting Depression in Social Media Posts Using Machine Learning" achieved 81% accuracy while Facial analysis seen in "Designing a Framework for Assisting Depression Severity Assessment from Facial Image Analysis" achieved just 56% accuracy. Furthermore, Sentiment analysis was done using tweets from twitter making the sample size very low for a real world use case. A paper titled "Machine Learning Techniques for Stress Prediction in Working Employees" by U Srinivasulu Reddy and Aditya Vivek Thota claimed that there is an increasing amount of stress disorders among employees. We found that the questionnaire approach was the best way to solve the problem of depression. Other approaches such as sentiment analysis, facial expression analysis and using emotion signals to detect depression were considered. Most of the sentiment analysis approaches use social media sites such as twitter as data sources. This limits our sample set to the people who are using twitter. Further, the person should be posting tweets regularly and should show depressive symptoms in their tweets. This limits our sample size even more. This approach also suffers from issues such as from population bias, limited sample size and user privacy. Facial expression analysis is another approach which was considered. The major drawback of facial expression analysis is that it fails to differentiate sadness from depression. It is also hard to detect depression using facial expressions since people can hide depression. Detecting emotional signals make use of makes use of expensive hardware and doesn't do a good job in detecting grief which is the primary emotion of a depressed person, making this infeasible to implement. Majority of the papers used machine learning algorithms such as Gradient Boosting, Random Forests, Support Vector Machines (SVM), K Nearest Neighbors (KNN) and Decision Trees to detect depression levels. Majority of the papers do not use deep learning techniques. Recent advancements in deep learning mean that there is a possibility to improve the results which can be experimented. The questionnaire approach provides a way to get direct inputs from the person under test. Since the test can be taken anonymously, it removes the problems surrounding therapy sessions. Since we use questions, we can analyze and interpret the output of the model which is not possible by using a Facial expression approach and is barely possible if we use a sentiment analysis approach.

III. System Architecture

The passive components of the system include data preprocessing, feature engineering, model selection scripts which were used to train and develop machine learning model. The user answers a questionnaire. The questionnaire data is sent to the target encoding model via JSON request. The features are then encoded using the target encoding model and the encoded data is passed to the random forest model which predicts the level of depression.



IV. Methodology

We found and used a tech industry-based mental health data set, and considering the depression subset of it. The dataset contains the details of the annual survey conducted by an American organization called Open Sourcing Mental Illness, Illinois (OSMI). The survey contains company and mental health related questions asked to employees of various IT companies throughout the world. We chose the OSMI 2018 dataset as it was the latest dataset available when we started the project. The dataset we got had 51708 values in the shape of 417x124. Out of the 124 features, 71 features had more than 50% missing values. These columns were dropped. Then, we dropped all the rows which were empty. We also dropped 3 columns containing ID and dates. For the rest of the columns with missing values, we filled the missing values using the most frequent values in that column. The dataset was clean after filling all the missing values and we were left with 53 features to work with. We ran feature selection and feature importance and utilized random forest to select the top 15 most important features. We wrote a function to manually encode the target variable "Depression". Then, we used Target Encoding algorithm to encode rest of the categorical variables. After selecting top correlated variables, we ran the data through a random forest model to calculate top fifteen most important features. There is another common feature selection method which is to quite simply measure the impact of each feature on the accuracy of the model directly. This is done by permuting the values of each feature and observing the resulting effect on the accuracy of the model, wherein, obviously, unimportant variables have little to no effect on the accuracy while a significant decrease is seen with the important variables. We used the encoded data with selected features and ran 9 different models and recorded their precision, recall and accuracy metrics and chose the best

among them based on that. All the models were tuned using Random Search Hyperparameter Tuning Algorithm. To implement the screening test, we chose Python as the main language of the backend to remove the coding overhead since implementing machine learning-based backend code is challenging in itself. We used MongoDB for storing the data. We used HTML5, CSS3, JavaScript languages and Bootstrap framework to implement the frontend. Bootstrap was used for setting up the layout and not for other stylistic elements.

V. Result Analysis

We were able to achieve 96.19% Accuracy using a Target Encoding/Random Forest model which was tuned with Random Search algorithm to select hyper parameters. This is significantly higher than paper [1] which had 75% Accuracy with Boosting. This is also higher than paper [2] which achieved 81.7% accuracy with Boosting. Our results are significantly higher than sentiment analysis approach seen in paper [3] which had 81% accuracy through Artificial Neural Networks. Our model completely outclasses image related models seen in paper [4] which had a 55% accuracy. The results are compared below.

Approach	Algorithm	Accurac
Questionnaire	Boosting	75%
Questionnaire	Boosting	81.70%
Sentiment Analysis	Artificial Neural Network	81%
Facial Analysis	Nearest Neighbours	55%
Questionnaire (this project)	**Target Encoding + Random Forest tuned using Random Search**	**96.20%**

All the algorithms in the following table used Target Encoded data & random Search for tuning.

Algorithm	Accurac
Random Forest	96.19%
Gradient Boost	89.52%
SVM	88.57%
KNN	87.61%
Decision Tree	86.66%
Bagging	85.71%
XGBoost	84.76%
AdaBoost	82.85%
Naïve Bayes	79.04%

The below table shows Precision, Recall, F1-score and Support for the Random Forest model.

Leve	Precisio	Recall	F1-scor	Suppor
0	0.90	0.90	0.90	21
1	0.86	1.00	0.92	6
2	0.99	0.97	0.98	78

From the table above, we can see that our model is really good in predicting level 2 depression and does a pretty good job in predicting level 0 depression. The model struggles when trying to classify level 1 depression. This can be attributed to the following reasons:

- Level 1 lies between Level 0 - No Depression and Level 2 - Highly Likely Depression. So, it is difficult for the model to predict.
- This level of depression is also very rare. It was seen just 8 times in the training set and 6 times in the test set.

Having said that, 86% Precision is still really good but it doesn't compare well with the precision of Level 0 and Level 2 depression.

VI. Conclusion

From the results, we can conclude that Questionnaire and Machine Learning combination is the right approach when we are trying to build a depression screening system. We achieved 21% increase in accuracy compared to our base paper. This can be because of the different approach we took while solving this problem. Target encoding proved to be a real big factor as it preserves the correlation after encoding. Hyperparameter tuning using Random Search also helped to achieve a higher accuracy than the rest of the approaches. As more data is collected on depression, a Machine Learning based approach becomes more viable for detecting depression. One of the drawbacks in our model is lower precision in detected level 1 depression as it is a Grey area. This problem will be solved if more data is available on level 1 depression. We also did not explore deep learning based solutions mainly because we were satisfied with how our random forest model performed. With the advancements of deep learning, it can become a viable option to predict depression.

References

[1]. U Srinivasulu Reddy, Aditya Vivek Thota, A Dharun (2018), 'Machine Learning Techniques for Stress Prediction in Working Employees', IEEE.

[2]. Sandhya P, Mahek (2019), 'Prediction of Mental Disorder for employees in IT Industry', IJITEE.

[3]. Abhilash Biradar (B) and S. G. Totad (2019), 'Detecting Depression in Social Media Posts Using Machine Learning', Springer Nature Singapore Pvt. Ltd.

[4]. A. Pampouchidou, K. Marias, M. Tsiknakis, P. Simos and F. Yang, F. Meriaudeau (2015), 'Designing a Framework for Assisting Depression Severity Assessment from Facial Image Analysis', IEEE.

[5]. Gyeongcheol–Hwan Lee (2019), 'Review of Machine Learning Algorithms for Diagnosing Mental Illness', Korean Neuropsychiatric Association.

[6]. Munmun De Choudhury, Scott Counts, Eric Horvitz (2013), 'Social Media as a Measurement Tool of Depression in Populations', ACM.

[7]. Imen Tayari Meftah and Nhan Le Thanh (2012), 'Detecting Depression Using Multimodal Approach of Emotion Recognition', IEEE.

[8]. Shatte ABR, Hutchinson DM, Teague SJ (2019), 'Machine learning in mental health: a scoping review of methods and applications', Psychological Medicine.

[9]. S'adan MAHM, Pampouchidou A, Meriaudeau F (2018), 'Deep Learning Techniques for Depression Assessment', ICIAS.