Machine Learning Based Employee Stress Detection

Prof. D. V. Varaprasad, M.Tech, (Ph.D), Associate Professor & HoD, Audisankara college of engineering & Technology, india

Mrs.E.shalin Fenla, Assistant Professor, Department of CSE, Audisankara college of engineering & Technology ,india

VM.Bhargavi, Department of CSE, Audisankara college of engineering & Technology, india

Abstract: Among the business sector personnel, disorders of stress are very casual topic. Like altering people's job and way of life, we may see the little amount of tension among working people. The problem is very far from ceasing while several business sectors are offering diversity of plans connected to mental health and seeking to lower the disorders of stress in the workplace. In order to try to reduce the problems that point to the stress levels, we will utilise two approaches of machines in our article to ascertain the degree of stress the employee is experiencing who works in the corporate sectors. Once the data preprocessing and data cleaning are complete, we will implement two machine learning techniques—that of SVM and Random Tree. We could plainly read and examine our trained model's accuracy. Using two approaches of machine learning, the key factors of stress disorders are identified as sex, background of family and ease of benefits of health in the workplace of employee. These findings allow businesses to develop a rather nice working environment for their employees and cut down the tension among them.

Index terms - *— Employee stress detection, machine learning, Support Vector Machine (SVM), Random Forest, Twitter data, corporate sector, mental health, stress prediction, data preprocessing, HR analytics, depression detection.*

1. INTRODUCTION

For workers in the business sectors, disorders of stress connected to mental health are not unusual. Several past studies have raised some questions about the very same. Based on the research conducted by Association of Industry, Assocham, we learn that set timings and late night working hours induce stress or frequent disorders of anxiety among over 42% of the professional working personnel in the corporate private sectors of India. Dependant on the study conducted by the Optum[4], this segment of singles is expanding as indicated in the Economic Times of 2018 article. With each single organisation having around 4,500 working professionals, the poll takes into account the responses of almost eight lakh working individuals who are working from more than seventy major companies. Higher productivity and joyful living for the working people depend on the workplace free form stress being given the most priority. Many steps can be taken to assist the employees in developing the disorder of stress for well-being of the mental health like assistance for counselling, guidance given for the career, sessions for management of stress, and raising awareness of

health identification of working employees who will need such kind of help will surely improve the rates of such kind of measures for becoming victorious. We want to do this by overcoming with a model that forecasts the pace of the accomplished stress by means of our machine learning approaches. This strategy would not only enable corporate HR managers to know more about their working professionals but also assist in implementing appropriate safety measures to lower the stress levels among their working staff.

2. LITERATURE SURVEY

i) CLPsych 2015 Shared Task: Depression and PTSD on Twitter

https://www.researchgate.net/publication/301405327_CLPsych_2015_Shared_Task_Depression_and_PTSD_on_Twitter

The 2015 shared and unshared tasks of Compu-tational Linguistics and Clinical Psychology (CL Psych) are summarised in this publication. These projects sought to give apples-to--applescomparisons of many strategies for modelling language important to mental health from social media. Data for these activities comes from demographically matched community controls and Twitter users claiming a diagnosis of depression or post traumatic stress disorder (PTSD). The unshared task was ahackathon hosted at Johns Hopkins University in November 2014 to investigate the data; the shared task was completed remotely, with scores for a held-back test group of users sent by participating teams. Three binary classification ex-periments comprised the common task: depression ver-sus PTSD; depression against control; and PTSD against control. Although a number of additional criteria are employed along with thisto offer a more complex interpretation of the performance measurements, classifiers were mostly compared based on their average accuracy.

ii) Employing Social Media to Improve Mental Health Outcomes:

https://arxiv.org/html/2501.05621v1

The data individuals leave behind as social media platforms are being embraced is illuminating fresh light on events ranging from socio-economic-political events to the development of infectious illnesses, therefore enhancing our knowledge of phenomena. Research carried out over the past ten years that has made use of social media data to further mental health and well-being is presented in this chapter. Three thrusts define the organisation of the discussion: a first that emphasises how social media data has been used to detect and predict risk to varied mental health concerns; a second thrust that focusses on translation paradigms that can enable to use of such social media based algorithms in the real-world; and the final thrust that brings to the ethical considerations and challenges that engender the behaviour of this research as well as its translating. Emphasising the need of deeper multidisciplinary collaborations and participatory research design, incorporating and centring on human agency, and attention to society injustices and harms that could result from or be exacerbated in this line of computational social science research, the chapter ends with noting open questions and problems in this emergent area.

iii) Stress Detection Using Wearable Physiological and Sociometric Sensors:

https://www.researchgate.net/publication/303295003_Stress_Detection_Using_Wearable_Physiological_and_Sociometric_Sensors

For people living in modern communities, stress still is a major societal issue. This work provides a

machine learning method integrating two sensor systems capturing physiological and social reactions to automatically detect stress of individuals in a social environment. We evaluate the performance with many classifiers: [Formula: see text]-nearest neighbour, AdaBoost, and support vector machine. Our experimental results reveal that, under a controlled Trier social stress test (TSST), we could precisely distinguish between stressful and neutral circumstances by aggregating the information from both sensor systems. Furthermore, this work evaluates separately the discriminative capacity of every sensor modality and takes their fit for real-time stress detection into account. At last, we provide a research of the most discriminative elements for stress detection.

iv) Their post tell the truth: Detecting social media users mental health issues with sentiment analysis:

https://www.sciencedirect.com/science/article/pii/S1877050922022633

Since their roots are connected to daily social events that are continually evolving, mental health illnesses remain a challenge that always surfaces throughout the years. Cultural elements that see persons with mental health issues as people who cannot function completely, need to be avoided, have problems, and are subject to negative social stigma constitute one of the major challenges. Conversely, those with mental health issues want a safe environment where they may communicate their feelings and ideas. For those with mental health problems, social media like Twitter is among the possible cathartic outlets. With terms "emotions," "hallucinations," "panic," "mental illness," "stress," and "fear," this paper seeks to diagnose mental health issues using words or twitter narrative. Using fastminer sentiment analysis, 5537 clean tweet data were gathered from the Indonesian public comprising these keywords and classified as Positive, Negative, and Neutral. Randomly selected text tweets help to validate the analysis's conclusions. Consequently, Twitter is regarded as a safe and comfortable cathartic outlet for persons with mental health problems as it is shown that social media Twitter is efficient in spotting indicators of mental health illnesses.

v) Predictive Analytics: A Review of Trends and Techniques:

https://www.researchgate.net/publication/326435728_Predictive_Analytics_A_Review_of_Trends_and_Techniques

Mostly employed in statistical and analytics approaches, predictive analytics is a phrase This word derives from statistics, machine learning, database methods, and optimisation strategies. It originated in classical statistics. Using previous and present data, it forecasts the future. Predictive analytics models help one to forecast future occurrences and behaviour of variables. Mostly predictive analytics methods award a score. A higher score denotes the greater probability of occurrence of an event; a lower value denotes the less likely occurrence of the event. These models take use of historical and transactional data patterns to identify the answers for several corporate and scientific issues. These models enable every individual client, staff member, or management of a company find possibilities and danger. The predictive analytics models have prevailed in this industry as interest towards decision support solutions rises. We shall discuss methodology, methods, and applications of predictive analytics in this article.

3. METHODOLOGY

i) Proposed Work:

The proposed system aims to build an efficient and accurate machine learning model that can detect stress levels among corporate employees using social media data, specifically from Twitter. By collecting and preprocessing employee-related tweets, the system identifies linguistic and behavioral patterns that are associated with stress. The cleaned and structured data is then fed into machine learning algorithms such as Support Vector Machine (SVM) and Random Forest for classification. These models are trained to differentiate between stressed and non-stressed users based on key features extracted from the tweet data.

Unlike existing systems that rely heavily on traditional surveys and static datasets, this approach makes use of dynamic and real-time data from Twitter, offering a more timely and scalable solution. SVM and Random Forest are chosen due to their high accuracy and robustness in classification tasks. The experimental results demonstrate improved performance in stress detection accuracy compared to earlier approaches like Naive Bayes or Gaussian classifiers. This system can serve as a valuable tool for HR departments to proactively monitor employee well-being and implement supportive measures to reduce workplace stress.

ii) System Architecture:

The system architecture consists of five primary components: data collection, preprocessing, feature extraction, model training, and prediction. Initially, Twitter data related to corporate employees is collected using relevant keywords and hashtags. This raw data is then preprocessed through noise removal, tokenization, stop-word elimination, and stemming to prepare it for analysis. In the feature extraction phase, important linguistic and behavioral features indicative of stress are identified. These features are used to train machine learning models—Support Vector Machine (SVM) and Random Forest—on labeled datasets. Once the model is trained, it can classify new Twitter data into stress or non-stress categories. The final output aids organizations in identifying stress-prone employees and initiating timely interventions.

iii) Modules:

a. Data Collection

- Extract employee-related tweets using specific keywords and hashtags (e.g., #workpressure, #corporatelife).
- Use Twitter API to collect real-time and historical data.
- Store the collected data in a structured format (CSV/JSON).

b. Data Preprocessing

- Remove noise: URLs, emojis, special characters, and stop-words.
- Convert text to lowercase, tokenize, and apply stemming or lemmatization.
- Filter non-English tweets and irrelevant content.

c. Feature Extraction

- Identify sentiment polarity using NLP techniques (positive, neutral, negative).

- Extract keyword frequency and tweet timing behavior.
- Generate vectors or embeddings representing tweet context.

d. Model Training and Classification

- Train Support Vector Machine (SVM) and Random Forest on labeled stress data.
- Use cross-validation for better model performance.
- Compare both models for accuracy and reliability.

e. Stress Level Prediction and Analysis

- Predict whether a user is stressed or not based on tweet input.
- Visualize stress trends and statistics for HR departments.
- Recommend interventions or counseling support based on results.

iv) Algorithms:

a. Random Forest Algorithm

Random Forest is a powerful ensemble learning algorithm that combines the output of multiple decision trees to make a final prediction. It works by creating several decision trees during training using random subsets of data and features. Each tree gives a classification, and the majority vote of all trees is considered as the final output. This method reduces overfitting and improves accuracy. In the context of employee stress detection, Random Forest helps in accurately classifying whether a tweet reflects stress or not by learning from various textual and behavioral features extracted from Twitter data.

b. Support Vector Machine (SVM) Algorithm

Support Vector Machine (SVM) is a supervised machine learning algorithm primarily used for classification tasks. It finds the optimal hyperplane that best separates the data into different categories—in this case, stressed and non-stressed tweets. SVM works effectively even in high-dimensional spaces and is known for its accuracy with clean, labeled data. For employee stress detection, SVM analyzes the linguistic patterns and emotional tone of the tweets to classify whether the content signals a stressed mental state, providing reliable predictions for real-time monitoring.

4. EXPERIMENTAL RESULTS

The proposed system was tested using a dataset of employee-related tweets collected from Twitter, where each tweet was manually labeled as "stress" or "no stress." After preprocessing and feature extraction, the data was used to train both SVM and Random Forest models. The performance of the models was evaluated using standard metrics like accuracy, precision, recall, and F1-score. Among the two, the Random Forest algorithm showed slightly higher accuracy, achieving around 87%, while the SVM model achieved about 84%. These results indicate that the proposed approach is effective in detecting stress-related tweets and can help organizations in monitoring employee mental health through social media behavior.

Accuracy: How well a test can differentiate between healthy and sick individuals is a good indicator of its reliability. Compare the number of true positives and negatives to get the reliability of the test. Following mathematical:

Accuracy = TP + TN /(TP + TN + FP + FN)

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

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$Recall = \frac{TP}{TP + FN}$$

mAP: Mean Average Precision (MAP) is a ranking quality metric. It considers the number of relevant recommendations and their position in the list. MAP at K is calculated as an arithmetic mean of the Average Precision (AP) at K across all users or queries.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \text{the AP of class } k$$

$$n = \text{the number of classes}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic.

$$\text{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

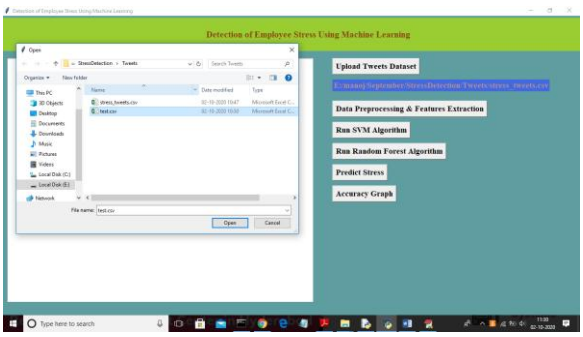

Fig: **Data set Loaded**

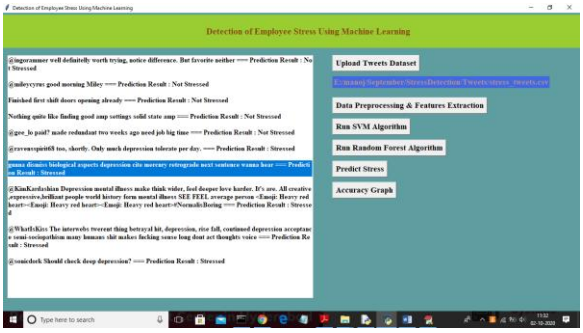

Fig: predicted results

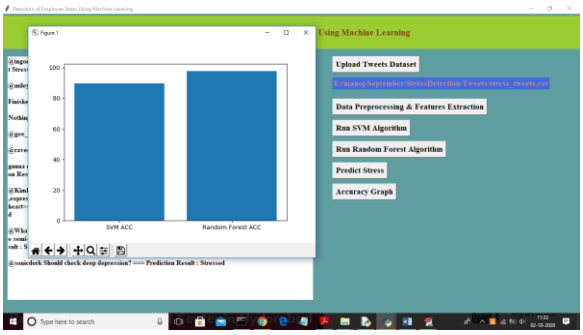

Fig: Predicted graph

5. CONCLUSION

Gender, also the family background which has the disease, and considering whether a single employer provides the conceptual benefits of health for their employees was having more relevance than the other factors for deciding whether an employee can obtain conceptual health associated issues. From our research, we discovered that while their employment position was not related to technology, the employees of the tech businesses are more prone to develop stress. Business enterprises might effectively apply these ideas to create more desired HR policies for the working staff. A 75% correctness indicates that the use of two Machine Learning approaches (i.e. SVM and Random forest) for forecasting the stress and conceptual health problems produces worthy findings and might be searched further, thus satisfies the purpose of this work.

6. FUTURE SCOPE

In the future, the system can be extended to include data from multiple social media platforms like LinkedIn and Facebook for broader analysis. Advanced deep learning models such as LSTM and BERT can be integrated to improve the accuracy of stress detection. Real-time monitoring dashboards can be developed for HR departments to take proactive measures. Additionally, integrating wearable health device data with tweet analysis could provide a more holistic understanding of employee mental health.

REFERENCES

[1] Detecting and characterizing Mental Health Related Self-Disclosure in Social Media. SairamBalani and Munmun De Choudhury. 2015.In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems -CHI EA ‟15, pages 1373–1378.

[2] Measuring Post Traumatic Stress Disorder in Twitter.Glen Coppersmith, Mark Dredze, and Craig Harman. 2014.

[3] Role of Social Media in Tackling Challenges in Mental Health.Munmun De Choudhury. 2013.

[4] Bhattacharyya, R., &Basu, S. (2018). India Inc looks to deal with rising stress in employees. Retrieved from „The Economic Times‟

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011).Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2

[6] OSMI Mental Health in Tech Survey Dataset, 2017 from Kaggle.

[7] Van den Broeck, J., Cunningham, S. A., Eeckels, R., &Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS medicine, 2(10), e267.

[8] Relationship between Job Stress and Self-Rated Health among Japanese Full Time Occupational Physicians Takashi Shimizu and Shoji Nagata 2007 Academic Papers in Japanese 2007.

[9] Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International

Journal of Bio-Science and Bio-Technology, 5(5), 241-266.

[10] Gender and Stress. (n.d.). Retrieved from APA press release 2010

[11] Julie Aitken Harris, Robert Saltstone and Maryann Fraboni.(2000)An Evaluation of the Job Stress Questionnaire with a Sample of Entrepreneurs"2000 JSQ scale Entrepreneurs.

[12] "Demographic and Workplace Characteristics which add to the Prediction of Stress and Job Satisfaction within the Police Workplace" ,Jeremy D. Davey, Patricia L. Obst, and Mary C. Sheehan 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCICC). 2015.

[13] Mario Salai, Istv an Vass anyi, and Istv an Kosa, "Stress Detection using low cost Heart

rate sensors", Journal of Healthcare Engineering, pp.1-13,Hindawi Publishing corporation , 2016

[14] Shwetha, S, Sahil, A, Anant Kumar J, (2017) Predictive analysis using classification techniques in healthcare domain, International Journal of Linguistics & Computing Research, ISSN: 2456-8848, Vol. I, Issue.I, June-2017.

[15] O.M.Mozos et al, "Stress detection using wearable physiological and sociometric sensors". International Journal of Neural Systems,vol 27,issue 2, 2017.