Stress Detection Using NLP & DL Models

Dr. Anil Vithalrao Turukmane
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh
anil.turukmane@vitap.ac.in

A Lakshmi Narayana
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh
lakshminarayana.21bce9147@vitapstudent.ac.in

G Chaitanya
School of Computer Science and Engineering
VIT-AP University
Amaravati, Andhra Pradesh
chaitanya.21bce9572@vitapstudent.ac.in

D Reddy Nithish Kumar
School of Computer Science and Engineering
VIT-AP University Amaravati,
Andhra Pradesh
nithish.21bce9054@vitapstudent.ac.in

A Somnath
School of Computer Science and Engineering
VIT-AP University Amaravati,
Andhra Pradesh
somnath.21bce9501@vitapstudent.ac.in

Abstract - Many people use social media platforms these days to post tweets about their everyday lives that reveal their mental health. Stress must be identified and dealt with before it becomes a serious issue. a considerable of casual communications shared every day on blogs, chat rooms, and social networking sites. This study suggests a method for calculating stress levels through data which were from the social media like Twitter. This project handled data collection, cleaning, system training, and determining people's stress levels. Accomplished by utilizing the algorithms or approaches and several other models like natural language processing (NLP) and ML models, XG-Boost, Random Forest, Decision Tree, SVM, Bernoulli-NB. People's health is at risk due to psychological stress. Proactive care necessitates quick stress assessment. Online social network data can be utilized to detect stress because users are accustomed to interacting with friends and sharing details of their everyday lives on these sites. heavy, we find a considerable relationship between the stress level of the person and their friends on social media. Our systematic methodology makes use of a sizable dataset from actual social networks. In order to improve outcomes, we first develop a series of stress-related tests and then use machine learning techniques to train the system. After that, the suggested system determines whether or not to highlight a tweet depending on the input tweets.

Keywords: BERT Model, SVM, XG-Boost, deep learning, machine learning, and natural language processing.

I. INTRODUCTION

Social networking platforms include well-known websites like Facebook, Twitter, WhatsApp and Instagram, where users create and utilize programs to express their thoughts and sentiments as well as to connect with people on a range of subjects. These days, social media platforms have a significant impact on people of all ages and are transforming people's lives. Because consumers utilize social media constantly, it is much simpler to ascertain their psychological health by promptly gathering and examining communication logs and messages from social media users.

The physical and biological conditions associated with psychological stress can be harmful to an individual's health. Every year, people's stress levels rise, and too much stress can sometimes lead to suicide thoughts. In India, there were 15.7 suicides for per 100,000 individuals in 2018. Nevertheless, stress is becoming increasingly common in our daily lives and is bad for people's health. As a result, it's crucial to determine people's stress levels before they worsen. With the growing usage of social media across all age groups, machine learning techniques—which are far superior to old methods—make it far more feasible to identify a user's emotional or stress condition early on.

A. Motivation –

Social Media Influence: Because they provide an abundance of user-generated content, social networks such as Facebook, Twitter and WhatsApp have become essential communication tools. The analysis of these data can reveal information about the psychological state of people.

Rising Stress Levels: Psychological stress is a growingpublic health concern, with increasing rates of stress- related disorders and suicides. Traditional detection methods are limited, prompting the need for more proactive and scalable approaches.

B. Contribution

- Created four excellent datasets based on Reddit and Twitter, which were made available to the scientific community to improve the understanding of the prevalence of stress and mental health problems in the general public.

- Sought to address issues including the lack of large-scale datasets for model development and time-consuming data annotation, which are essential for creating a trustworthy stress detection system.

- The accuracy and resilience of the stress forecast are increased by combining SVM, XG-Boost, and further learning approaches.

- Stress detection accuracy can be improved by fusing literary analysis with characteristics learned from social interactions.

II. Literature Review

Sentiment analysis has witnessed significant advancements in recent years, evolving from simple binary classification to more nuanced approaches. Early models, such as the one proposed by Alexander Pak and Patrick Paroubek [5], categorized tweets into three classes: objective, positive, and negative. They utilized emoticons as indicators of sentiment and constructed a sentiment classifier using Naive Bayes with N-gram and POS-tag features. However, their approach was limited by the reliance on tweets containing emoticons.

Subsequent research [6-10] focused on effective data preprocessing techniques for social media content, especially tweets. Strategies include eliminating stop words, symbols and punctuation, and standardizing word forms.

Categories of sentiments positive, negative, and neutral, were introduced by Apoorv Agarwal and associates [11] in their sentiment analysis approach. During the research, a variety of models were examined, including unigram models, feature-based models and tree kernel-based models. The results acknowledge the tree kernel approach while reinforcing the significance of traits like word polarity and position-based tagging.

Recent developments in deep learning have improved stress analysis and detection procedures. Contextual deep word embeddings were created by Peters and Neumann for stress comprehension-oriented language learning [17]. Radford and Narasimhan improved language comprehension and stress detection with their generative pre-training approach [18]. Stress analysis benefited greatly from the BERT architecture developed by Devlin, Chang, Lee, and Toutanova, which pretrained deep bidirectional transformers [19]. By using BERT-based multi-labelled sentiment analysis and improved TF-IDF, Jin, Lai, and Cao's study made it easier to comprehend complicated stress [20].

It is now simpler to define and locate articles because to the widespread usage of tagging in web content management. Using the tag-LDA paradigm, Xiance et al. [15] demonstrated versatile tag identification technique. Krestel et al. [16] combined LDA trials with probabilistic techniques to recommend labels in a personalized way.

Stress analysis and detection have benefited greatly from recent developments in deep learning. Contextualized deep word representations were introduced by Peters and Neumann, which enhanced language comprehension connected to stress [17]. The generative pre-training developed by Radford and Narasimhan enhanced stress detection and language understanding [18]. The development of stress analysis pretraining deep bidirectional transformers was greatly aided by the BERT design of Devlin, Chang, Lee, and Toutanova [19]. In order to facilitate nuanced stress assessment, Jin, Lai, and Cao's work utilized modified TF-IDF and BERT to multi-label sentiment analysis [20].

Summary, by utilizing methods like deep learning architectures and sophisticated feature engineering, researchers have made tremendous progress in the detection and analysis of stress through advancements in sentiment analysis and natural language processing.

III. Stress Detection using deep learning & Natural Language Processing[Methodolgy]

- Preprocessing

To guarantee data integrity and purity, preprocessing textual data for stress detection requires a number of crucial processes. First, every text is transformed to lowercase in order to prevent case variances from causing conflicts and standardize the dataset. Links and URLs are eliminated from the text because they usually add noise and don't help with stress analysis. Furthermore, any HTML tags are removed so that the text is the only thing on display.

The text is then made simpler and less dimensional by eliminating punctuation, such as commas, periods, and exclamation points. Additionally, numeric characters are not included because they could not be important for identifying stress and might be regarded as noise in the dataset. Eliminating common stop words like "the," "is," and "and" highlights more significant words that are probably suggestive of material connected to stress while reducing noise.

In order to standardize and enhance uniformity in word representation, stemming methods also reduce words to their root forms. This method aids in vocabulary reduction while maintaining word meanings even when they change over time or in form. The procedure culminates with word tokenization, which facilitates feature extraction and further analysis. This separates the text into discrete tokens or words. Together, these preliminary steps guarantee that the text data is thoroughly cleaned and standardized to produce an efficient stress analysis.

- **Proposed methods architecture**.

Support Vector Machine (SVM):

The objects in the region are placed so that there is a significant amount of space between the groupings, and an SVM approach illustrates how it appears between them. Estimating the hyperplane of the maximum margin that results in the optimal class partition is the goal. Auxiliary vectors are cases that are often closer to a hyperplane with maximum limitations. Variable selection focuses on the part of the data set that matches the training set. Two hyperplanes can arise at the same time because to dual class support variables. Additionally, the classifier's generalizability decreases with the amount of space between the other two hyperplanes. SVMs were developed differently than several other machine learning techniques.

XG-Boost:

Extreme Gradient Boosting, also known as XG-Boost, is a versatile and excellent ML algorithm that has gained interest because of its reactivity and performance in a variety of predictive modelling paradigms, including classification, regression, and ranking. In the field of ensemble learning, XG-Boost is a member of the boosting technique class. By training many weak learners (usually decision trees) and then merging them, boosting creates a strong learner. XG boosting outperforms conventional gradient boosting techniques thanks to the improved model structure, which attempts to decrease overfitting and increase prediction accuracy.

Logistic Regression:

The fundamental classification method that is frequently employed in machine learning is called logistic regression. Logistic regression evaluates the given data and provides probability for potential outcomes using a logistic function for binary outcomes only, in contrast to random forest, which averages the predictions of several trees.

Logistic Regression as a classifier to analyze stress levels in textual data. Unlike Random Forest, which utilizes multiple decision trees, Logistic Regression focuses on estimating the probability of a particular outcome. In our implementation, we utilized logistic regression to model the relationship between the features extracted from the text data and the binary stress level classification.

Random Forest:

Using the average prediction from several decision trees constructed from various dataset subsets, a Random Forest classifier enhances the data's predictive power. To put it another way, it votes on both decision trees predictions rather than depending just on one. We used the random forest approach in our work and constructed a model with 21 estimators, sometimes known as decision trees or trunk trees.

Naive bayes:

For data classification, one supervised learning method that is accessible is the Naive Bayes. Its guiding concepts came from the Bayes Theorem. A large training sample is used for the primary application, which is text categorization. One of the most popular classification algorithms that makes it easier to quickly develop machine learning models with quick prediction services is the Naive Bayes classifier, which is simple and effective. It resembled a probabilistic fighter in that it depends on the likelihood of certain future events depending on the existence or non-existence of specific items. Sentiment analysis, email spam detection, and news classification are just a few of the applications for the Naive Bayes technique.

The two principles that make up the Bayes approach are naïve and uninformed, which go hand in hand with naïve. The theory's core tenet—that the existence of one quality has no bearing on the existence of another—is shown by the term naive. For example, take into consideration the colour, shape, and flavour of a certain fruit—an apple—that we all know by instinct to be red, spherical, and delicious. In this instance, the other qualities are not necessary to confirm that this is an apple because each attribute works independently. The Bayes technique is so named because it is based on the theoretical foundations of the Bayes theorem.

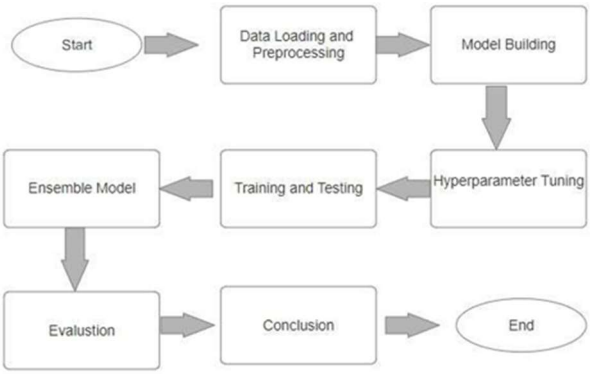

Fig- Workflow structure of the model

BERT (Bi-directional Encoder Representations from Transformers**):**

Architecture:

Based on a transformer model with multiple encoder layers that use self-attention mechanisms (12 layers in BERT-base and 24 in BERT-large). Token, positional, and segment embeddings are combined in the input.

Training Tasks:

Language Model with Masks (MLM)**:** predicts words at random while taking into account both left and right context. In order to comprehend sentence relationships, Next Sentence Prediction (NSP) makes predictions about whether one sentence logically follows another.

Pretraining: Completed on sizable datasets (such as Wikipedia), followed by optimization for particular NLP tasks (e.g., question answering, classification).

Hybrid Model:

Combining predictions from SVM and XG-Boost using weighted averaging.

Hyperparameter Tuning:

Grid Search or Random Search:

Searching for the best combination of hyperparameters for each model to optimize performance. Analyzing which features contribute the most to model predictions, especially in ensemble models like Random Forest and XG-Boost.

K-fold Cross-Validation:

Ensuring robustness of model performance by splitting data into multiple folds for training and testing.

Prediction:

Generating predictions for new/unseen data using the trained models. Combining predictions from multiple models for improved accuracy (if applicable).

IV. DATASET DESCRIPTION

It appears that you have provided a tabular dataset with 2838 rows and 112 columns. Below is the description of the columns:

id: An identifier for each entry.

label: A binary label indicating a classification, with 0 and 1 values.

confidence: Confidence level associated with the classification.

social_timestamp: Timestamp related to social activity.

social_karma: Karma or reputation score associated with social activity.

syntax_ari: Automated Readability Index related to syntax.

lex_liwc_WC: Word count through LIWC.

lex_liwc_Analytic: Analytical score based on LIWC.

lex_liwc_Clout: Clout score based on LIWC.

lex_liwc_Authentic: Authenticity score based on LIWC.
... (continues for 102 more columns.

The dataset seems to include features related to social activity, linguistic analysis (LIWC), readability, sentiment, and potentially other aspects. Each row likely represents a data point or observation, while each column represents a different feature or attribute.

For further analysis or interpretation, you may want to examine the specific meanings and interpretations of each column, especially those related to social metrics, linguistic analysis, and sentiment.

V. RESULTS AND DISCUSSION

Using the Dreaddit dataset, we assessed five distinct machine learning models for text classification. Among the models are Support Vector Classifier, Bernoulli Naive Bayes, XG-Boost, Random Forest, Logistic Regression and hybrid models. SMOTE was used to address the imbalanced nature of the data after TF-IDF vectorization was used as a preprocessing step on the dataset.

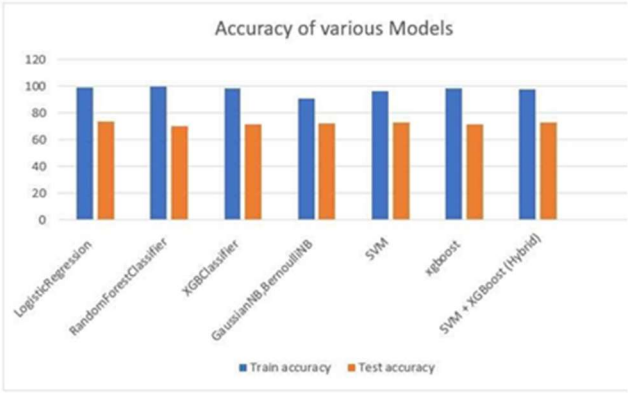

Fig1: Performance Comparison of Accuracy, testing and training ratio.

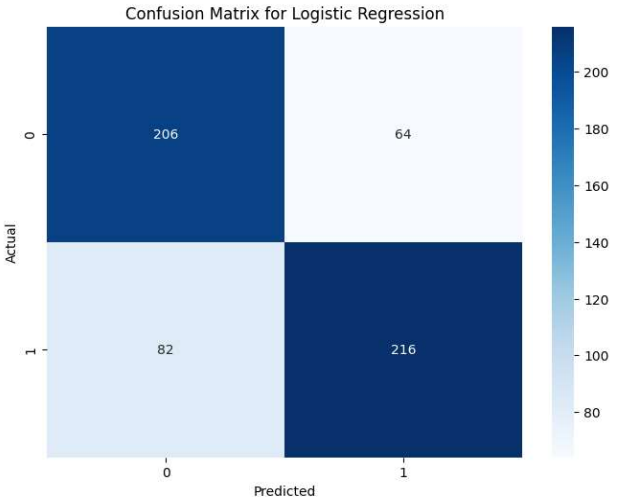

Fig2: Logistic Regression confusion matrix

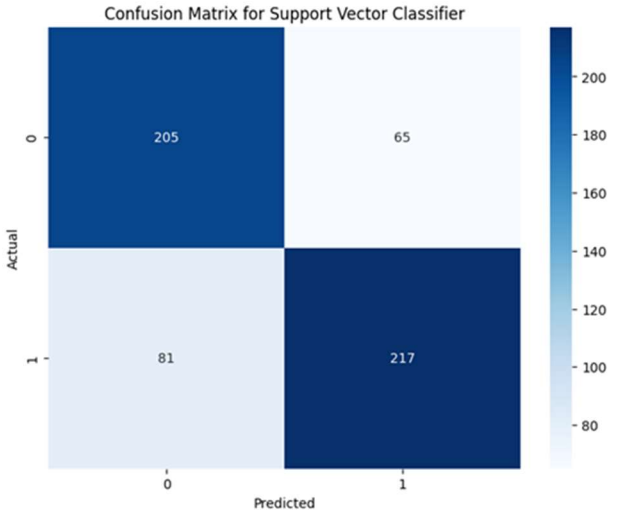

Fig3: Support Vector Machine confusion matrix

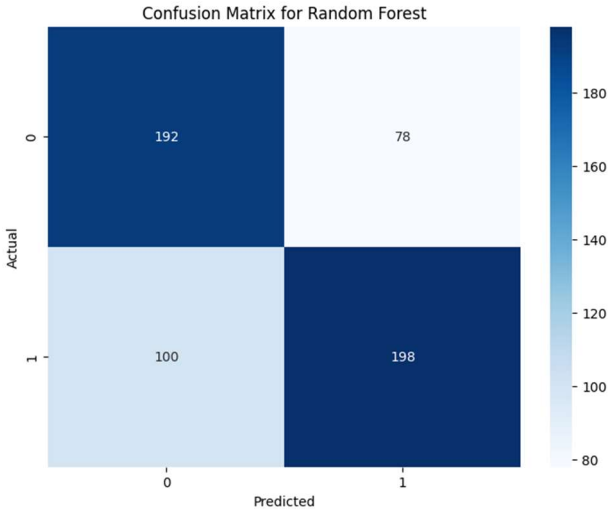

Fig4: Random forest confusion matrix

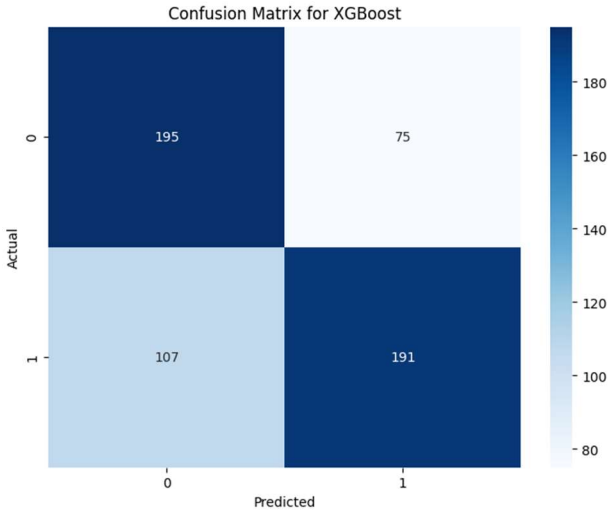

Fig5: XG-Boost confusion matrix

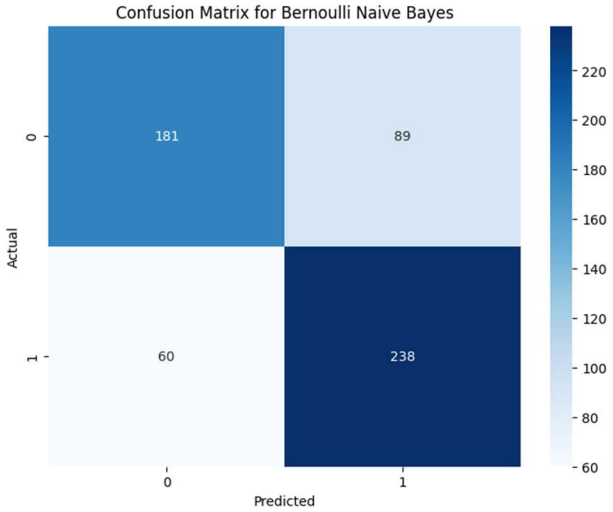

Fig6: Confusion Matrix of Bernoulli Naïve Bayes.

The experiment demonstrates that, for stress identification from online posts, both Bernoulli Naive Bayes and Logistic Regression provided balanced performance, with Logistic Regression turning out to be the most reliable method overall. Overfitting was evident in both Random Forest and XG-Boost, with XG-Boost remaining a viable choice because of its adaptability in spite of its lower test accuracy.

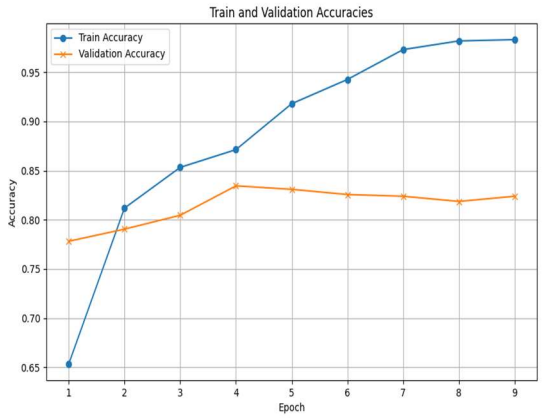

Fig7. Epoch vs Accuracy graph of BERT Model

Throughout ten epochs of training, BERT-based models for stress classification demonstrated a consistent increase in both training and validation accuracies. The ensemble was created by weighted voting, which allocated weights according to each model's performance. Weighted random sampling was used to address class imbalance, and early stopping was used to prevent overfitting. Accuracy plots were used to visualize performance trends, and the ensemble successfully classified stress with competitive accuracy, saving the best-performing model for final predictions.

Fig7. Dataset most repeated words

For the stress detection model, the discrete values 0 and 1 make up the label column of this data set. Stress is represented by a number one, and its absence by a number zero. In place of the numbers one and zero, respectively, the terms Stress and No Stress have been used. In this regard, the columns, text, and label that are crucial to the training process of the machine learning model must be properly chosen and positioned at this point:

	text	label
0	said felt way sugget go rest trigger ahead you...	Stress
1	hey rassist sure right place post goe im curr...	No Stress
2	mom hit newspap shock would know dont like pla...	Stress
3	met new boyfriend amaz kind sweet good student...	Stress
4	octob domest violenc awar month domest violenc...	Stress

VI. CONCLUSION AND FUTURE WORK

The majority of young people today—and everyone else, for that matter—struggle with tremendous stress as a result of heavy workloads, peer pressure, and other family-related problems. For this reason, assessing an individual's stress level is crucial. Thus, it's critical to recognize stress and eliminate it as soon as it manifests. This project was designed to increase their awareness of the stress issue.

Those who find it difficult to talk to others about their concerns can greatly benefit from our effort. This will assist these individuals in readjusting to the situation based only on their social connections and might also motivate them to seek medical attention. We used machine and human learning, as well as concepts from sentiment analysis. This system's primary advantage over earlier techniques is its quick deployment and non-invasiveness for stress sensing.

VII. REFERENCES

[1] Dataset - https://huggingface.co/datasets/asmaab/dreaddit

[2] McKeown, K., and Turkan, E. (2019, October 31). Dreaddit: A Reddit dataset for social media stress study. arXiv.org. retrieved from https://arxiv.org/abs/1911.00133 on November 7, 2021.

[3] Winata, G. I., Kampman, O. P., & Fung, P. Attention-Based LSTM for Distant Supervision-Based Psychological Stress Identification from Spoken Language. IEEE International Conference on Speech, Signal Processing, and Acoustics (ICASSP) in 2018. 10.1109/icassp.2018.8461990

[4] A Inf Process Manage, 2017;53(1):106–21; Thelwall M. TensiStrength: Stress and relaxation magnitude identification for social media texts.

[5] S. Rachele, R. Sahu, and V. Bhattacharjee's study "Mental Stress Prediction graph convolutional network-based framework“ was released in 2022. The article https://doi.org/10.1016/b978-0-323-919996-200007-7

[6] "XG-Boost Model for Chronic Kidney Disease Diagnosis," by Ogunleye and Q.-G. Wang, in IEEE/ACM Transactions on Computational Biology and Bioinformatics (2020). https://doi.org/10.1109/TCBB.2019.2911071

[7] Glass, James, Ghassemi, Mohammad, and Hanai, Tuka (2018). Using text Sequence modelling of Interviews to Identify Depression. 10.21437/Interspeech.2018-2522. 1716-1720

[8] Khoshgoftaar TM, Pomeranets A, Wang D, and Sohangir S. Big Data: Financial Sentiment Analysis using Deep Learning. J Big Data, 2018;5(1):1–25.

[9] "Physiological sensing-based stress analysis during assessment" by Aniruddha Sinha, Pratyusha Das, Rahul Gavas, Debatri Chatterjee, and Sanjoy Kumar Saha was presented at the Frontiers in Education Conference (FIE) in 2016, pp. 1–8.

[10] Nine Ways Stress is More Dangerous Than You Think", Healthline., Feb 2018, [online] Available: http://www.healthline.comlhealth- news/mental-eight-ways-stress-harms-your-health-082713.

[11] S. Jadhav, A. Machale, P. Mharnur, P. Munot, S. Math, Text based stress detection techniques analysis using social media, in: 2019 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA), IEEE, 2019, pp. 1–5.

[12] Elzeiny, M. Qaraqe, Blueprint to workplace stress detection approaches, in: 2018 International Conference on Computer and Applications (ICCA), IEEE, 2018, pp. 407–412

[13] Krestel R, Fankhauser P. Personalized topic-based tag recommendation. Neurocomputing. 2012;76:61–70. https://doi.org/10.1016/j.neucom.2011.04.034.

[14] Zucco C, Calabrese B, Cannataro M. Sentiment Analysis and Affective Computing for Depression Monitoring. In 2017 IEEE international conference on bioinformatics and biomedicine (BIBM). New York: IEEE. 2017. p. 1988–1995.

[15] Pradha S, Halgamuge MN, Vinh NQT. Effective text data preprocessing technique for sentiment analysis in social media data, 2019 11th International Conference on Knowledge and Systems Engineering (KSE). 2019. p.

[16] https://doi.org/10.1109/KSE.2019.8919368. C Chaturvedi S, Mishra V, Mishra N. Sentiment analysis using machine learning for business intelligence, 2017 IEEE International Conference on power, control, signals, and instrumentation engineering (ICPCSI). 2017. p. 2162–2166. https://doi.org/10.1109/ICPCSI.2017.8392100.

[17] Boyi X, Rebecca P, Owen R, Ilia V, and Apoorv A. sentiment analysis using data from Twitter. The Workshop on Languages in Social Media Proceedings. 2011.

[18] Jenya B, Matina H, Bruno P, van der Erik G, Mijail K, Vanni Z, Ralf S, and Alexandra B. news articles that analyze sentiment. LREC proceedings. (n-1). 2013

[19] Ahmad HF, Aldarwish MM. The IEEE 13th International Symposium on Autonomous Decentralized System (ISADS), Bangkok, 2017, pp. 277–280, Predicting Depression Levels Using Social Media Posts. 10.1109/ISADS.2017.41 (http://doi.org/10.1109).

[20] Zucco C, Calabrese B, Cannataro M. Using Affective Computing and Sentiment Analysis to Track Depression. In 2017, the IEEE held an international conference on biomedicine and bioinformatics (BIBM). IEEE, New York, 2017, pp. 1988–1995.