STM JOURNALS

Journal of
Computer Technology & Applications

ISSN: 2229-6964 (Online)
ISSN: 2347-7229 (Print)
Volume 16, Issue 1, 2025
January–April
DOI (Journal): 10.37591/JoCTA

https://journals.stmjournals.com/jocta

Review JoCTA

Textual Clues to Stress: A Machine Learning Approach

K. Purushotam Naidu[1,*], M. Prasanthi[2], Y.S.P. Kousalya[2], Trisha Jenna[2]

Abstract

Nowadays, numerous individuals utilize social media platforms to share tweets about their daily lives, which often reflect their mental well-being. Recognizing and managing stress is essential before it becomes a serious issue. Each day, a significant volume of informal messages is posted on discussion forums, blogs, and social networking sites. This study introduces a method for detecting stress using information gathered from social media, with a focus on Twitter. The project encompasses various tasks, including data collection, data cleaning, system training, and stress identification for users. Natural Language Processing (NLP) and Machine learning (ML) methods like SVM, Random Forest (RF), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Decision Tree (DT) will be used to do this. Detecting stress in a timely manner for preventive care is challenging. The proposed study consists of two main components: stress detection through machine learning methods and information extraction through natural language processing. The four primary stages of this study involve text mining, auto summarization, stress detection, and collection of social media data. The suggested model can predict an internet user's stress level or cognitive load. It incorporates several machine learning strategies, with Support Vector Machine demonstrating superior performance concerning F1 score, accuracy, recall, and precision compared to other methods. The early detection of stress offered by the current methodology will bring significant benefits to society. Therefore, the proposed system utilizes tweets as input to make informed decisions.

Keywords: SVM, random forest, k-nearest neighbor, naïve bayes, decision tree, text mining, auto summarization

INTRODUCTION

Over the last few years, the prevalence of stress-related issues has prompted observers to explore innovative methods for early detection and intervention. Leveraging the wealth of real-time data available on social media channels like Twitter, the present research work investigates the feasibility of utilizing social tweets to detect and understand stress levels among users. By analyzing linguistic patterns, sentiment and content; researchers aim to develop computational models and algorithms capable of accurately identifying indicators of stress. This study explores the methodological approaches, potential applications, and ethical considerations associated with stress detection using social tweets, contributing to the advancement of digital mental health research.

***Author for Correspondence**
K. Purushotam Naidu
E-mail: purushotam.k30@gmail.com

[1]Assistant Professor, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Gayatri Vidya Parishad College of Engineering for Women, Visakhapatnam, Andhra Pradesh, India
[2]Student, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Gayatri Vidya Parishad College of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

Received Date: October 05, 2024
Accepted Date: December 12, 2024
Published Date: December 31, 2024

Citation: K. Purushotam Naidu, M. Prasanthi, Y.S.P. Kousalya, Trisha Jenna. Textual Clues to Stress: A Machine Learning Approach. Journal of Computer Technology & Applications. 2025; 16(1): 72–76p.

Mental stress is a medical or physical condition that affects a person and their life. Each year, an increasing number of individuals are experiencing higher levels of stress, which can sometimes result in thoughts of suicide. There will be approximately 16.7 suicides per 100,000 people in India. Stress, which was considered harmful to human health, has become part of everyday life. Some of the current

stress sensing methods are slow in nature and have many limitations such as hysteresis, time and labor, and high cost. With the increasing use of social networking applications across all ages, it is increasingly possible to detect user emotions or stressful situations at an early stage using machine learning techniques way better than traditional methods.

LITERATURE SURVEY

NLP stands at the forefront of technology for identifying mental health concerns like anxiety and melancholy within social media content. Many studies focus on text prediction using NLP and machine learning [1]. Observers used content generated by social media users to predict stress using SVM, Logistic Regression (LR), RF, and NB, but SVM was used with RF to achieve a perfect accuracy of 91%.

Illahi *et al.* stacked up data from Reddit. Classical and ensemble ML approaches were used [2]. The traditional methods are DT, LR, SVM and NB. Entry methods include jump, climb and vote. The best model was logistic regression with an F1 score of 76.6%.

Anakha *et al.* detected stress by identifying emotions on human faces through facial expressions, video streaming, and audio recordings [3]. The dataset includes emotions such as happiness, anger, sadness, disgust, surprise, fear and prejudice. The dataset is processed through Convolutional Neural Networks (CNN) for alignment. This layer was implemented using the Progressive Model.

Another approach proposed by Zubler and Yoon is used to detect stress in plants with the help of imaging and fluorescence [4]. Advances in ML algorithms such as ANNs and DNNs have enabled image detection techniques for biotic and abiotic stressors.

Acharyulu *et al.* used ML techniques to assess users and their stressors [5]. The main techniques used in this project and development are eyebrow detection, blink detection and real-time facial expression recognition. They used ML techniques to implement this project on three platforms: Open CV, NumPy and Tenser flow.

METHODOLOGY

Support Vector Machine

SVM can handle nonlinear relationships between data and high-level features. In addition, SVM exhibits resilience against outliers and noise, rendering it appropriate for real-world datasets [6]. It identifies two categories by pinpointing the most effective dataset that enlarges the gap between neighboring data points belonging to distinct classes. As numerous planes are discovered to differentiate between categories, expanding the gap between points enables the algorithm to ascertain the optimal boundary separating the classes. The optimal line is called a support vector. This occurs because these vectors intersect the data points that establish the maximum margin.

Random Forest

Random Forest is an advanced and robust machine learning method employed for classification tasks and eliminating noise from data [7]. It functions by constructing numerous decision trees in the training phase. Each decision tree is trained using a randomly selected portion of the training dataset along with a random selection of features, resulting in a varied group of trees. When making predictions, each tree in the forest generates its own class predictions for classification tasks or continuous value predictions for regression tasks. The overall prediction is then made by taking a majority vote for classification and calculating the average for regression across all the trees.

Naïve Bayes

Naive Bayes is a Bayes based machine learning algorithm [8]. It is commonly employed for classification tasks in natural language processing (NLP) applications, including spam detection, sentiment analysis, and categorizing text. The algorithm computes the likelihood that a sample pertains to a category by

considering the existence or nonexistence of particular attributes or traits. A key benefit of this model is its capacity to manage both extensive and limited training datasets. However, performance may be compromised if independence is violated or there is a strong relationship between roles. Overall, NB is a powerful algorithm with many features that should be a key tool in your machine learning kit, especially in the areas of text classification and NLP.

K-Nearest Neighbor

The K-Nearest Neighbors (KNN) algorithm, a non-parametric learning technique, has a rich history of application in machine learning for both classification and regression assignments [9]. Here, new data points are clustered or predicted based on the average or majority votes of their neighbors in the category space. For classification purposes, the class label of the nearest neighbor is assigned to a new data point. Smaller k values provide smaller resolution limits but are more prone to noise, larger k values provide better resolution but smoother limits. Overall, KNN is a versatile algorithm suitable for many classification and regression tasks where simplicity and flexibility are important, especially for efficiency.

Decision Trees

Decision trees are widely favored in machine learning for their application in classification and inference tasks [10]. The algorithm subdivides the dataset into groupings guided by the characteristics that most effectively segregate the data according to specific standards, such as Gini impurity and information gain. DTs are transparent and capable of processing both numerical and categorical data. Methods like pruning, limiting the maximum depth, or employing clustering techniques such as random forests can help minimize redundancy and enhance generalization performance. In general, decision trees provide a simple and easy-to-understand model for multi-level decision making (Figure 1).

RESULTS AND DISCUSSION

Eqs. (1)–(3) reflect the metrics: Precision, Recall, F1-scores and Support, that are used to assess the effectiveness of our models. Table 1 displays the outcomes of these indicators.

$$\text{Precision} = TP/(TP+FP) \quad (1)$$

$$\text{Recall} = TP/(TP+FN) \quad (2)$$

$$\text{F1-score} = TP/(TP+(FP+FN)/2)) \quad (3)$$

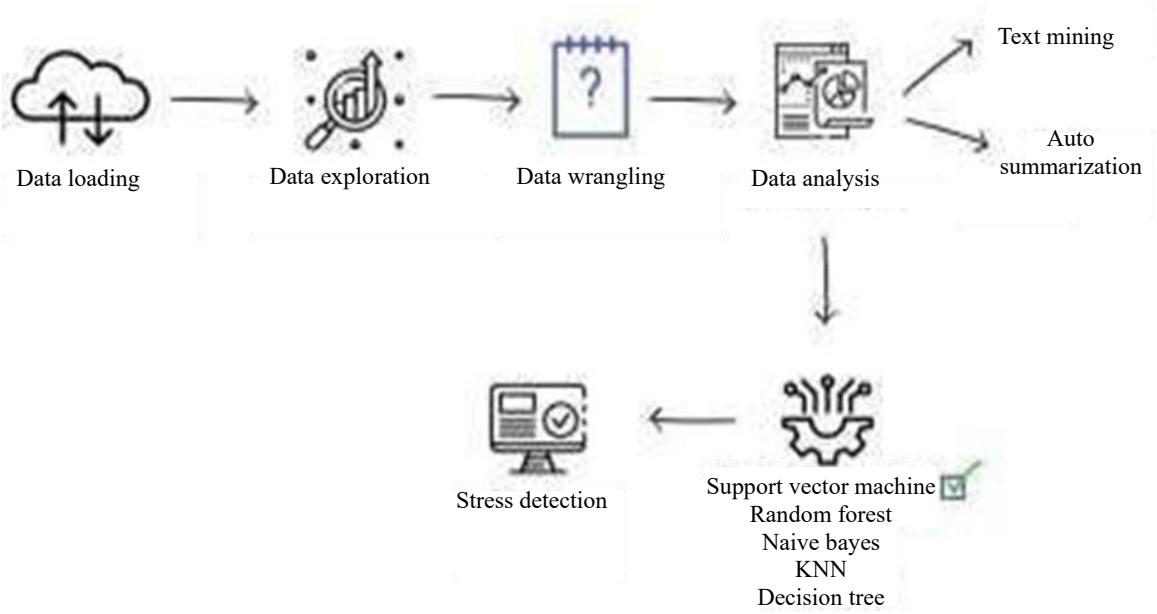

Figure 1. This architecture diagram shows the work flow.

Table 1. Shows the accuracy, macro average and weighted averages of varied machine learning algorithm.

Classifier	Accuracy	Macro avg.	Wt. avg.
SVM	0.742621	0.74	0.74
Random Forest	0.726092	0.72	0.73
Naïve Bayes	0.697757	0.69	0.70
KNN	0.644628	0.64	0.65
Decision Trees	0.590319	0.62	0.62

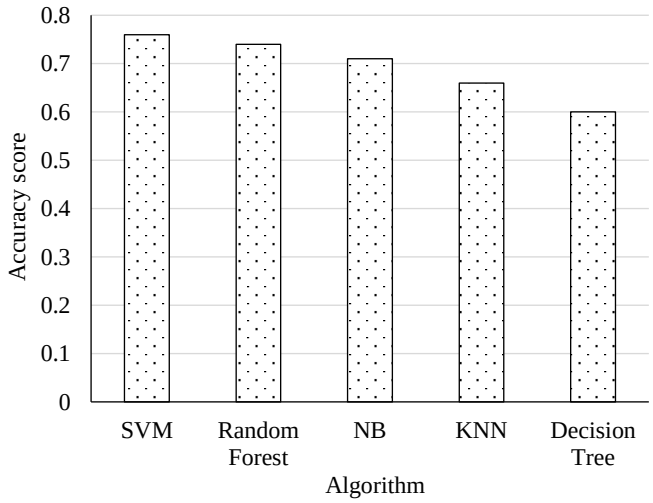

Figure 2. The histogram shows the accuracy of different algorithms, with SVM performing the best, followed by Random Forest and Naive Bayes.

A study on stress detection using NLP and ML in social media found that of the five algorithms evaluated, SVM was the most accurate at identifying stress levels in user-generated content (Figure 2). Because of SVM's ability to handle high-dimensional data and identify nonlinear relationships among features found in social media content, we evaluate the effectiveness of SVM classification using a carefully selected dataset of social media posts labeled with stress indicators. An SVM classifier was developed using this labeled data and subsequently applied to categorize new, unlabeled posts. Compared to SVM, RF, NB, KNN and DT algorithms, it is less accurate in the task of stress detection in social media content. Random Forest uses ensemble learning and Naive Bayes uses probabilistic principles, but both tend to capture the complex, non-linear relationships underlying the data and can lose accuracy. Furthermore, KNN's reliance on local similarity and the vulnerability of decision trees to overfitting may limit its ability to identify stress levels across multiple social media posts. The robustness of Support Vector Machines (SVM) stems from its adeptness in managing high dimensional datasets and capturing intricate non-linear associations, rendering it particularly well suited for this endeavor, as evidenced by its outstanding classification accuracy of 74.2%.

CONCLUSION

This research demonstrates the feasibility of leveraging social media data, particularly Twitter, to accurately detect stress levels using machine learning techniques. Among the algorithms assessed, Support Vector Machines (SVM) demonstrated the best performance in categorizing stress-related content. This study underscores the potential of early stress detection for preventive interventions. Future studies could explore ways to improve the feature engineering process, aiming to boost the accuracy of the model. Incorporating multimodal data, such as user profiles and behavioral patterns, could provide additional insights into stress indicators. Broadening the dataset to encompass a variety

of demographics and cultural backgrounds would enhance the model's applicability across different populations. Furthermore, exploring real-time stress detection and developing personalized intervention strategies based on stress levels are promising avenues for future investigation. By consistently enhancing stress detection models, we can aid in creating effective systems for mental health support.

REFERENCES

1. Kumari K, Das S. Stress Detection System using Natural Language Processing and Machine Learning Techniques. In WNLPe-Health@ ICON. 2022; 45–55.
2. Illahi M, Siddiqui IF, Ali Q, Alvi FA. Ensemble machine learning approach for stress detection in social media texts. Quaid-E-Awam University Research Journal of Engineering, Science & Technology, Nawabshah. 2022; 20(02): 123–8.
3. Anakha PS, Devi A, Nair AS, Suresh A, George N. Automated Stress Detection using Machine Learning. Int J Eng Res Technol (IJERT). 2022; 10(04): 163–174.
4. Zubler AV, Yoon JY. Proximal methods for plant stress detection using optical sensors and machine learning. Biosensors. 2020 Nov 29; 10(12): 193.
5. Acharyulu KV, Sampath Kumar N, Paavan Sampath K, Yaswanth Reddy B, Guna Sekhar G. Stress Detection using Machine Learning Technique. J Emerg Technol Innov Res (JETIR). 2023; 10(2): e195–e200.
6. Weerasinghe S, Erfani SM, Alpcan T, Leckie C. Support vector machines resilient against training data integrity attacks. Pattern Recognit. 2019 Dec 1; 96: 106985.
7. Reis I, Baron D, Shahaf S. Probabilistic random forest: A machine learning algorithm for noisy data sets. The Astronomical Journal (AJ). 2018 Dec 20; 157(1): 16.
8. Tabash M, Abd Allah M, Tawfik B. Intrusion detection model using naive bayes and deep learning technique. Int Arab J Inf Technol. 2020 Mar 1; 17(2): 215–24.
9. Bansal M, Goyal A, Choudhary A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decis Anal J. 2022 Jun 1; 3: 100071.
10. Demirović E, Lukina A, Hebrard E, Chan J, Bailey J, Leckie C, Ramamohanarao K, Stuckey PJ. Murtree: Optimal decision trees via dynamic programming and search. J Mach Learn Res. 2022; 23(26): 1–47.

