KEC_AI_MIRACLE_MAKERS@LT-EDI-2024: Stress Identification in Dravidian Languages using Machine Learning Techniques

Kogilavani Shanmugavadivel[1], **Malliga Subramanian**[1], **Monika R J**[1], **Monishaa S**[1], **Rishibalan M B**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{monikarj.22aid, monishaas.22aid}@kongu.edu
rishibalanmb.22aid@kongu.edu

Abstract

Identifying an individual where he/she is stressed or not stressed is our shared task topic. we have used several machine learning models for identifying the stress. This paper presents our system submission for the task 1 and 2 for both Tamil and Telugu dataset, focusing on using supervised approaches. For Tamil dataset, we got highest accuracy for the Support Vector Machine model with macro f1-score of 0.98 and for Telugu dataset, we got highest accuracy for Random Forest algorithm with macro f1-score of 0.99. By using this model, Stress Identification System will be helpful for an individual to improve their mental health in optimistic manner.

1 Introduction

Stress, anxiety, and depression (SAD) are psychological disorders that have a serious negative impact on mental stability. These disorders interfere with an individual's ability to go about their everyday life normally and can sometimes worsen into trauma. The human body releases a variety of chemicals when under stress, despair, or worry, and this results in alterations to nonverbal body language. These psychological diseases can be generically categorised as stress, anxiety, and depression according to the different stages involved in their exploration. Stress is the second stage of mental illness, during which psychological illnesses become more moderate since anxiety is a persistent factor. The third most serious psychological condition that can have a long-term negative impact on a person's physical and mental health is depression.An individual's level of discomfort is a result of stress, and this discomfort manifests as anxiety or depressive episodes. Stress is the culmination of all the things that can make someone feel stressed out. Exercises, additional work, a task overload, shallow breathing, insufficient sleep, questionnaires, etc. are examples of stressors. According to a study, stress can have a beneficial or negative effect depending on the circumstances. The study looked at people's social media posts, where they shared their feelings and emotions, to determine whether or not they were stressed. Social media posts in code-mixed Tamil and Telugu should be classified as either Stressed or Not stressed by the system. Numerous machine learning methods, including the Random Forest, Naive Bayes, and Support Vector Machine (SVC) algorithms, have been employed. This is how the rest of the paper is structured. The literature on work linked to stress identification is briefly discussed in Section 2. Section 3 provides a detailed description of our system, and Section 4 presents the findings and conclusions from the experiments. We wrap off the work by discussing potential implications for further research.

2 Literature Review

In Singh and Kumar (2022), the researchers have used some existing computer vision models for systematic review and used machine learning algorithms to detect SAD, which is more efficient than medical investigations because machine learning is fast and best for computing stress.

The proceedings in Robles et al. (2022),The researchers have used surface electromyography signals (sEMG) for detecting stress with the help of convolutional neural networks, and they got moderate range of the macro f1-score for a bi-class and multi-class classification. But they didn't provide necessary information about the size or diversity of the dataset used for training and testing, and insights into the interpretability of the model.

S and Karthick (2022) have also used the deep learning modal with a convolutional-based network approach and multimodel data with the help of sensors in which the data are collected, such as heartbeat, body temperature, respiration, electromyographic (EMG) data, and additional long and short-term Memory is used. They didn't provide any

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 277–281
March 21, 2024 ©2024 Association for Computational Linguistics

limitations or drawbacks.

In Tahira and Vyas (2023) a hybrid deep learning model that combines bidirectional long short-term memory (BLSTM) and convolutional neural networks (CNN) is presented for exploiting EEG signals to determine stress. Even though the modality got higher accuracy, they didn't explore other potential factors that are responsible for the stress.

In Gowtham et al. (2023), the researchers used the BERT model for text-based research and achieved better range of the f1-score, and they combined stacked transformer encoder layers with stacked bi-directional LSTM. But it did not explore other modalities such as signal- or speech-based analysis, and it is not clear how the model's performance compares to other existing state-of-the-art models in stress analysis.

In Suba Raja et al. (2023), they will send the test data through SMS alerts using a GSM module by extracting facial expression and mapped onto the emotion space and the EEG signal value is evaluated. The accuracy and robustness have been limited for the evaluation of this system and have not been discussed the potential limitations.

Saputra and Nafi'Iyah (2022) used feature extraction techniques including mean, standard deviation, and MAV, were applied to the EEG signals to capture relevant information. They have used several machine learning models to features, but the KNN algorithm achieved the highest accuracy in distinguishing between stressed and normal individuals. But they did not provide information about demographic characteristics and also not investigate the impact of external factors.

Garg et al. (2021) aimed to identify the stress among individuals using machine learning and wearable sensors with a random forest model in both binary and three-class classifications, achieving macro f1-scores of 83.34 and 65.73, respectively.But this paper fails to discuss the ethical considerations and privacy concerns related to the use of the wearable sensors.

Sharma et al. (2021) provides a comprehensive review and analysis of supervised learning (SL) and soft computing (SC) techniques used in diagnosis and the potential use of the hybrid technique gives a more accurate stress diagnosis. Their limitations are due to the factors such as real-time data collection, bias, integrity, multi-dimensional data, and data privacy.

Kul (2021) focuses on predicting and detecting stress in individuals by using IoT technology and body sensors, and that uses deep learning algorithms to analyze this data and suggest sending alerts, messages to the individual's relatives for support. But they didn't compare with any other existing methods and didn't provide any real-world validations of the proposed modal in practical scenarios.

3 Problem and system description

From the given dataset, we have to train the model whether the given sentence is stressed or not stressed. This shared task is to detect the individuals whether he/she affected by stress from their social media postings by analysing their shared feelings and emotions. Given dataset of social media postings consists of both Tamil and Telugu languages with this, we have to classify the given test data with 2 labels namely "stressed" or "not stressed".

3.1 Dataset description

The shared dataset consists of 2 languages namely Tamil and Telugu. In Tamil, the training dataset consists of 1,784 Stressed class labels and 3,720 Non-Stressed class labels out of 5,504 labels and the test dataset consists of 1,020 labels. The Telugu training dataset consists of 1,783 Stressed class labels and 3,314 Non-Stressed class labels out of 5,097 labels and the test dataset consists of 1,050 labels. Additionally, they are provided with the development dataset to check the model.

Dataset	No. of Comments
Train	5,504
Test	1,020

Table 1: Tamil Dataset Description

Dataset	No. of Comments
Train	5,097
Test	1,050

Table 2: Telugu Dataset Description

3.2 Work flow of the proposed system

1.Data pre-processing 2.Encoding module 3.Model description

The above mentioned are the major sub categories in the work flow which is explained below with detailed description.

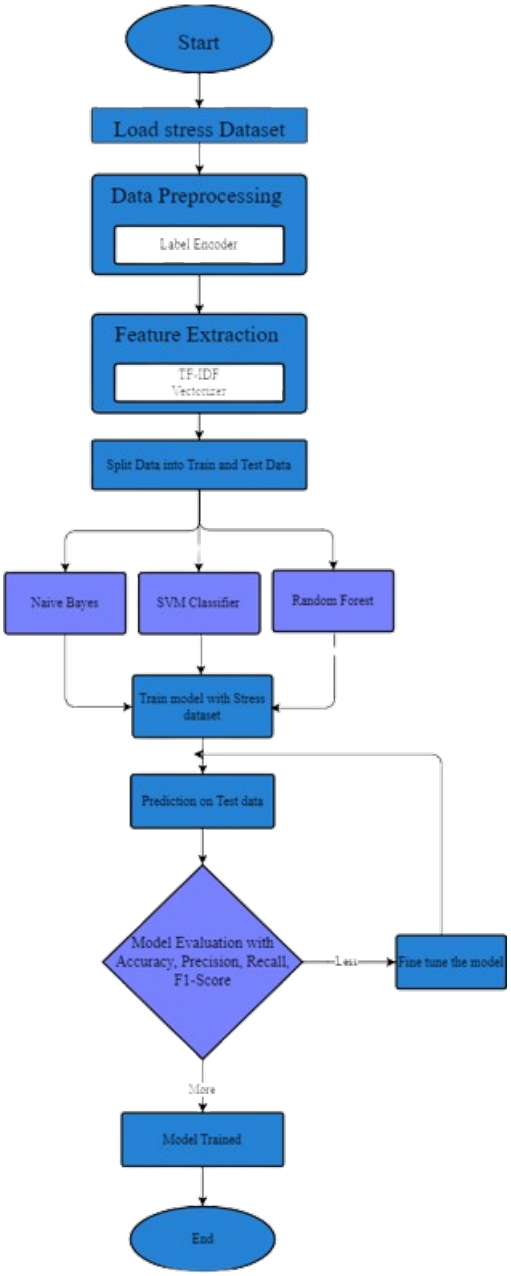

Figure 1: Proposed System Workflow

3.2.1 Data pre-processing

For the given dataset, we have used label encoder which is used to convert the categorical data into the numerical data. It will assign a unique integer to each category which helps the algorithm assume categorical data as numerical data so it makes easier for the models to process the given dataset.

3.2.2 Encoding module

For our dataset, we have used tfidfvectorizer which is imported from sklearn.feature-extraction. The feature extraction is used to makes the dataset in more efficient manner and is very helpful for better predictions by enhancing the model performance and reducing complexity. The TfidfVectorizer accepts the given dataset as input and which transforms the text into matrix where the rows are represented as documents and the columns are represented as unique word and TF-IDF will be calculated to create the matrix. The main use of vectorizer is the conversion of text data into the numerical representations such as matrix to get better model performance.

3.2.3 Model description

To predict where the person is stressed or not stressed by their social media postings, we used

three machine learning models for both the dataset i.e., Tamil and Telugu dataset to find the highest accuracy model. The three machine learning models are namely,

Naive Bayes classifier algorithm works based on the Bayes theorem which gives equal importance to all the features to predict the class label. In training dataset, it calculates the class and feature probabilities. During prediction, it computes the likelihood probability of each class given the features, assigning the highest probability class.

Random forest algorithm is a machine learning method that construction of the multiple decision trees by randomly selecting features and samples and handles the high dimensional data. It excels in accuracy for classification, regression and feature selection tasks. It can be used for finding both classification and regression the given dataset.

Support vector machine(SVM) is an algorithm which is also used for both classification and regression. It has diverse domains like text classification and image detection . It identifies the hyperplane that maximizes the margin between classes and can also handle the non-linear classification. It can enable SVM to learn complex decision boundaries.

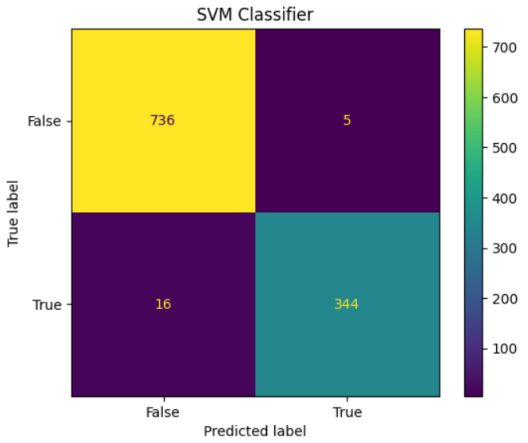

Figure 2: Confusion Matrix Of Support Vector Classifier Model- Tamil Data

4 Experimental Analysis

In this experiment we have used 2 different languages of dataset and 3 machine learning model to predict the class label whether it is “stressed” or “non-stressed”. In Tamil dataset, we have gotten accuracy of 98.09% in SVM classifier,97.27% in random forest algorithm and 89.19% in Naïve Bayes algorithm. As of our accuracy result, all the model will have the high accuracy hence we considered the support vector machine algorithm as the best algorithm among the other algorithm and it also have 0.98 macro f1-score. In Telugu dataset, we have got accuracy of 98.9% in SVM classifier,99.01% in random forest algorithm and 92.9% in naïve Bayes algorithm. As of our accuracy result, all the model will have the high accuracy hence we considered the random forest algorithm as the best algorithm among the other algorithm and it have macro 0.99 f1-score.

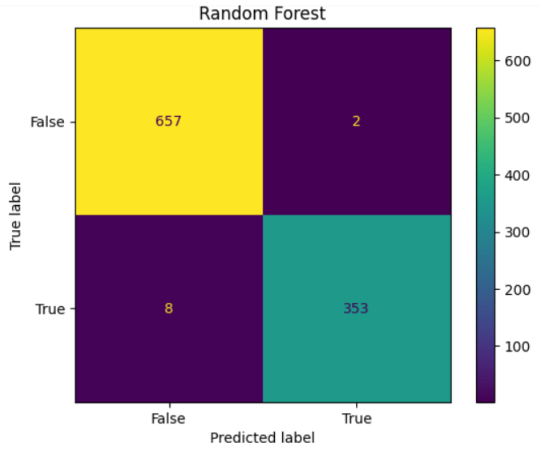

Figure 3: Confusion Matrix Of Random Forest Model- Telugu Data

Model	Macro F1-Score
Support Vector Classifier	0.98
Random Forest	0.97
Naive Bayes	0.89

Table 3: Macro F1-Score Metrics for Tamil Data

Model	Macro F1-Score
Support Vector Classifier	0.98
Random Forest	0.99
Naive Bayes	0.93

Table 4: Macro F1-Score Metrics for Telugu Data

5 Conclusion

Stress Identification is a very sensitive topic where many people around us and we also got stressed now-a-days. Some peoples are handling the things in practical ways but most of 80% of peoples are going to the depression state and they are pushed to take the wrong decision by the surroundings. Hence stress identification system will help an individual to improve their mental health in positive

manner. For both the datasets, we got the highest accuracy points and high macro f1-score. So, we got SVM for Tamil dataset with 98% and random forest algorithm for Telugu dataset with accuracy 99% as best predicting models. Therefore,we got more accuracy rate while comparing with any machine learning model and deep learning model,

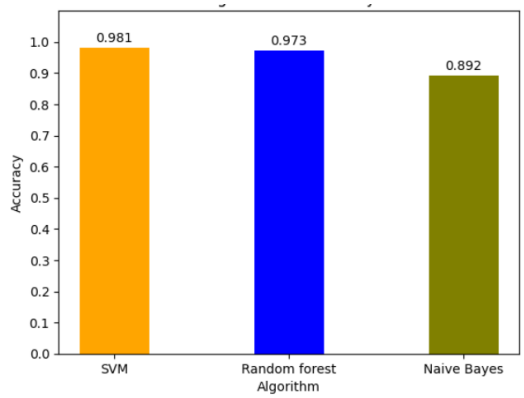

Figure 4: Accuracy - Tamil Data

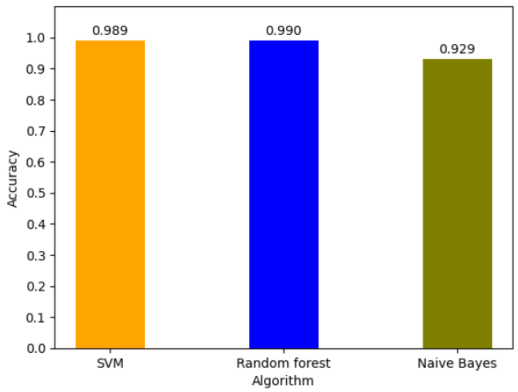

Figure 5: Accuracy - Telugu Data

Model	Accuracy
Support Vector Classifier	0.98
Random Forest	0.97
Naive Bayes	0.89

Table 5: Accuracy for Tamil Dataset

Model	Accuracy
Support Vector Classifier	0.98
Random Forest	0.99
Naive Bayes	0.92

Table 6: Accuracy for Telugu Dataset

References

2021. Stress prediction and detection using iot and deep learning: A comprehensive review. *International Journal for Research in Applied Science and Engineering Technology*, 9(9):1874–1880.

Prerna Garg, Jayasankar Santhosh, Andreas Dengel, and Shoya Ishimaru. 2021. Stress detection by machine learning and wearable sensors. In *26th International Conference on Intelligent User Interfaces - Companion*, IUI '21 Companion, page 43–45, New York, NY, USA. Association for Computing Machinery.

B Gowtham, H Subramani, D Sumathi, and BKSP Kumar Raju Alluri. 2023. Stress analysis using machine learning. In *Applied Computing for Software and Smart Systems: Proceedings of ACSS 2022*, pages 227–234. Springer.

Diego Robles, Mouna Benchekroun, Vincent Zalc, Dan Istrate, and Carla Taramasco. 2022. Stress detection from surface electromyography using convolutional neural networks. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 3235–3238.

Praveenkumar. S and T. Karthick. 2022. Automatic stress recognition system with deep learning using multimodal psychological data. In *2022 International Conference on Electronic Systems and Intelligent Computing (ICESIC)*, pages 122–127.

Nophaz Hanggara Saputra and Nur Nafi'Iyah. 2022. Identification of human stress based on eeg signals using machine learning. In *2022 1st International Conference on Information System Information Technology (ICISIT)*, pages 176–180.

Samriti Sharma, Gurvinder Singh, and Manik Sharma. 2021. A comprehensive review and analysis of supervised-learning and soft computing techniques for stress diagnosis in humans. *Computers in Biology and Medicine*, 134:104450.

Astha Singh and Divya Kumar. 2022. Computer assisted identification of stress, anxiety, depression (sad) in students: A state-of-the-art review. *Medical Engineering Physics*, 110:103900.

S. Kanaga Suba Raja, Durai Arumugam S S L, R. Praveen Kumar, and J. Selvakumar. 2023. Recognition of facial stress system using machine learning with an intelligent alert system. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1–4.

Maryam Tahira and Prerna Vyas. 2023. Eeg based mental stress detection using deep learning techniques. In *2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pages 1–7.