Comparative Analysis of Human Stress Prediction Using Multi-Models Natural Language Processing System

A S M Nasim Khan , MD. Adnan Howlader , Mohammad Nasif Sadique Khan , Farah Binta Haque , Md Fardin Rahman Ami , Md Sabbir Hossain , Ehsanur Rahman Rhythm , and Annajiat Alim Rasel

Department of Computer Science and Engineering
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{*a.s.m.nasim.khan, md.adnan.howlader, mohammad.nasif.sadique.khan, farah.binta.haque, md.fardin.rahman.ami, md.sabbir.hossain1, ehsanur.rahman.rhythm*}*@g.bracu.ac.bd*
annajiat@gmail.com

***Abstract*—**
***Index Terms*—**

I. Introduction

Modern life's rapid pace, constant connectivity, and work-related demands are fueling an increase in stress. The past pandemic further exacerbated stress through health concerns, remote work challenges, and disrupted routines. Working with stress is significant in light of the fact that stress has turned into an unavoidable and possibly crippling issue in present-day culture. The effects of weight on both individual prosperity and by and large general well-being are significant. Constant stress can prompt different physical and psychological well-being issues, going from cardiovascular illnesses to tension and sorrow. By creating models that can predict feelings of anxiety, we can distinguish those in danger and give opportune mediation, empowering people to embrace compelling survival techniques. Moreover, understanding the examples and triggers of stress through NLP can prompt experiences that illuminate fitted ways to deal with stress of the executives, cultivating better ways of life and working on by and large personal satisfaction. Stress presents critical risks, obvious from insights showing its inescapable effect. Around 77% of people experience actual side effects because of stress, with 73% confronting mental impacts, according to the American Establishment of Stress. The American Mental Affiliation features that pressure costs the US economy more than $300 billion every year. Moreover, the World Wellbeing Association recognizes pressure as a main calculated psychological well-being issue, influencing 1 out of 4 individuals internationally. Using techniques like AI becomes basic to anticipate, forestall, and oversee pressure, relieving its destructive results on people and society. Stress and sentiment analysis share a connection in examining emotional expressions within textual data. By applying sentiment analysis to texts related to stress, such as social media texts or online discussions, it becomes possible to understand the emotional tone or feelings and sentiment associated with stress issues. This approach provides a structured method for understanding how people convey their stress-related feelings and allows for insights into common stressors, and potential triggers. Through sentiment analysis, researchers can derive valuable information to tailor interventions, develop support strategies, and gain a deeper comprehension of the emotional nuances surrounding. Sentiment analysis is the technique of extracting subjective information from text. Sentiment analysis can assess the intensity of emotion in a text or speech. It is able to identify people who are stressed by examining their written messages, can create solutions for reducing stress, and determine the causes of stress, is possible to track variations in stress levels over time using sentiment analysis, help people control their stress, and customized stress treatments can be developed via sentiment analysis. Overall, sentiment analysis is an effective approach for locating, measuring, and monitoring stress levels. In addition, it can be implemented to create individualized stress-reduction programmes. Sentiment analysis can be done using several neural network models. RNNs are used for sentiment analysis in order to identify long-term dependencies between texts. While LSTM requires greater computation, both GRU and LSTM networks have proven useful for sentiment analysis applications. To extract features from text, CNNs can also be used. CNNs can grab local patterns and features in text for sentiment analysis. An entire sentence's context is extracted by bidirectional LSTM networks. Sentiment analysis is improved by processes of attention and bidirectional LSTM models. Also, several transformer-based models can be used. BERT and GPT models are powerful models for sentiment analysis due to their bidirectional nature and contextual understanding. For sentiment analysis, attention models might concentrate on relevant words or phrases. For sentiment analysis, hierarchical

979-8-3503-3286-5/23/$31.00 ©2023 IEEE

models may represent both local and global context. The dataset and its use determine which neural network model is most effective for sentiment analysis. Experimentation is required to determine hyperparameters such as learning rate, number of layers, and neurons per layer. Before training the neural network, preprocessing the text data may improve performance.

II. Related Works

This article [1] talks about Natural language processing (NLP) has been used to identify mental illness using text over the past ten years. NLP-driven psychological maladjustment location research is on the ascent, as per a story survey of 399 examinations distributed in the beyond a decade, and profound learning-based techniques have acquired prevalence lately.

The audit distinguished three principal sorts of datasets utilized for psychological sickness location: online entertainment posts, screening overviews, and clinical notes. Depression and suicide were the most frequently studied mental illnesses, followed by stress, anxiety, and schizophrenia.

The survey likewise found that an assortment of NLP techniques have been utilized for psychological sickness discovery, including conventional AI strategies, for example, support vector machines (SVMs) and gullible Bayes classifiers, as well as profound learning strategies like intermittent brain organizations (RNNs) and convolutional brain organizations (CNNs). Profound learning techniques have for the most part been displayed to beat conventional AI strategies.

According to the survey's reasoning, NLP may be an important tool for the early detection of mental illness. Notwithstanding, the creators brought up that there are as yet various obstructions that should be survived. A portion of these deterrents incorporate the shortfall of huge, top notch datasets, the necessity for models that are straightforward, and the need to think about the moral ramifications of utilizing NLP to identify psychological instability.

In conclusion, NLP-driven psychological maladjustment location research has seen significant advancements in the past decade, with deep learning techniques leading the way. By leveraging textual data from online posts, screening surveys, and clinical notes, NLP has the potential to revolutionize mental health care by enabling early detection and personalized interventions. However, addressing challenges related to data quality, model transparency, and ethical considerations is essential for realizing the full potential of NLP in mental illness detection and ensuring its responsible and ethical use in healthcare settings. This study [2] plans to anticipate self-destructive ideation and uplifted mental side effects in grown-ups who have as of late been released from mental ongoing or trauma center settings utilizing AI and normal language handling (NLP) in Madrid, Spain. Collapse is an essential in general success concern, and early unquestionable check of people in danger is key for persuading mediation and balance attempts.

The review gathered information from members who answered organized mental and actual wellbeing instruments at numerous subsequent places. Notwithstanding the organized information, members were asked to answer an unstructured inquiry, "how would you feel today?". NLP-based models were utilized to dissect the text reactions to this genuine inquiry, while calculated relapse expectation models were assembled utilizing the organized information.

The outcomes showed that the NLP-based models, which used the unstructured text information, accomplished moderately high prescient qualities for recognizing people in danger of self-destructive ideation and mental pain. In any case, it is significant that the exhibition of the NLP-based models was somewhat lower contrasted with the organized information based models, which approached more unambiguous and distinct data.

Prediction based on NLP could have significant effects on mental health. The review's discoveries recommend that NLP can be used to distinguish people in danger of self destruction or mental misery quickly. In situations where lengthy structured surveys are impractical, this automated approach may offer a cost-effective screening alternative.

The utilization of NLP for anticipating ailments has shown guarantee in different biomedical applications, and this exploration stretches out its pertinence to emotional well-being evaluation utilizing instant messages, explicitly short message administration (SMS) messages, happening beyond clinical settings. Continuous alarms created from SMS texts hold the possibility to illuminate ideal clinical intercessions, in this manner forestalling self destruction and tending to elevated mental pain.

The study emphasizes the significance of using cutting-edge, cost-effective methods like natural language processing (NLP) for data collection and analysis in suicide prevention efforts. By joining NLP with existing information from Electronic Wellbeing Records (EHRs) and different libraries, specialists can upgrade the distinguishing proof and mediation processes for patients with a high probability of self destruction endeavors. Additionally, the application of natural language processing (NLP) to unstructured clinician notes and other textual data presents opportunities to enhance the identification of various health conditions, mental health issues that resist treatment, and negative health outcomes.

Notwithstanding, the exploration recognizes that moral contemplations, including security and information insurance, should be painstakingly tended to while using NLP for psychological wellness appraisal. Guaranteeing the mindful utilization of these advances and dealing with delicate information properly is critical to acquire the trust of the two patients and clinicians.

The presented research article titled "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification" [3] delves into the domain of text classification and explores the potential of transfer learning models, specifically BERT-based models, for automating text classification tasks. The study focuses on three diverse datasets related to COVID-19, covering fake news, English tweets, and extremist/non-extremist content. The objective is to assess the performance

of transfer learning models on these datasets.

The research employs various advanced transfer learning techniques, including BERT-Base, BERT Large, RoBERTa-Base, RoBERTa-Large, DistilBERT, ALBERT-Base-v2, XLM-RoBERTa-Base, Electra-Small, and BART-Large.

The datasets used in the study vary in size, with the COVID-19 fake news dataset containing over 10,000 instances, the COVID-19 English tweet dataset comprising almost 7,000 instances, and the extremist non-extremist dataset being the largest with over 21,000 instances.

The methodology involves data preprocessing, encoding, model evaluation, and testing. After cleaning the datasets by removing URLs, converting text to lowercase, and lemmatization, the cleaned data is encoded and used to train and evaluate the transfer learning models. The evaluation metrics include accuracy, precision, recall, and F1-score. The results are compared against state-of-the-art techniques to highlight the effectiveness of the proposed approach.

The experimental results show the impressive performance of the models on different datasets. For the COVID-19 fake news dataset, the RoBERTa-base model achieves a remarkable accuracy of 99.71%. On the COVID-19 English tweets dataset, the BART-large model emerges as the top performer with an accuracy of 98.83%. Finally, in the extremist & non-extremist datasets, both BERT-base and BERT-large models achieve an outstanding accuracy of 99.71%.

In conclusion, the research emphasizes the potential of transfer learning models, specifically BERT-based models, in text classification tasks. The findings highlight the exceptional performance of these models across various datasets. The researchers suggest future research directions, including the exploration of larger datasets, multiclass classification, and the incorporation of emoticons to enhance the accuracy and scope of text classification. The study underscores the power of transfer learning in automating text classification tasks and encourages further investigation into real-time sentiment analysis and broader social network analysis.

In a paper named Stress detection using natural language processing and machine learning over social interactions [4], the authors Tanya Nijhawan, Girija Attigeri and T. Ananthakrishna worked on research of detecting stress using sentimental and emotional techniques on a large Twitter dataset that contains individual posts and comments. As a language model, they used BERT and used an unsupervised machine-learning technique that adopted Latent Dirichlet Allocation. This unsupervised method helped to cluster similar genre data and types. LDA is used to calculate the topic structure from the density of the topic. Their deep learning-based BERT model classification helped to classify stress labelling. From this approach, they achieve around 94% of accuracy. They also compared with typical machine learning models but they found the BERT technique more evident. In their future work, they are willing to work on bi-polar sentiments and like to apply on distinct platform datasets to justify the approach. In [5] paper mentioned “Feature-Based Depression Detection from Twitter Data Using Machine Learning Techniques”, researched by Piyush Kumar, Poulomi Samanta, Suchandra Dutta, Moumita Chatterjee4 and Dhrubasish Sarkar, mainly focuses on classifier techniques. They went through comparison and review analysis of multiple classifiers and justified which one gives a better outcome. The implementation went on a real-life dataset extracted from Twitter and evaluated the sentimental analysis on it. They used bigram, trigram and unigram approaches with TF-IDF embedding system and applied through multiple classifiers. They found satisfactory accuracy from XGBoost and SVM models consecutively at 81% and 89%. Comparatively higher accuracy was found from a combination of LDA+Bigram+TF-IDF in SVM model. In their paper, they mentioned their future works on identifying the personalities from data on linguistic style.

Social media is an effective way to detect psychiatric problems, thereby preventing undiagnosed mental disorders.This paper [6], categorized postings from Reddit to determine stressful and non-stressful using machine learning models and multiple embedding methods. A high F1 score of 0.76, a Precision score of 0.71, and a Recall score of 0.74 were obtained for the results. These results can be used to evaluate mental stress among social media users in real-world situations.

The ability to identify and provide support for difficulties with both physical and mental wellness depends increasingly on stress evaluation and sentiment analysis of posts on microblogging websites. To aid in early treatment for depressive disorders, this research provides an understanding of the issues and traits that stress people from all over the world.

The popular social media platform Reddit’s posts have been taken for consideration throughout this paper. This dataset contains 2800 texts for training. Reddit receives about 470,000 comments a single day, and there are 500 million tweets every day. People can express their emotions on social media platforms, and the volume of messages sent via text makes it easier to recognise signs of mental stress. To identify stress on social media, embedding methods and machine learning models are applied.

This paper used the Dreaddit dataset used to identify mental stress in Reddit posts. In order to categorize stressful postings on Reddit, NLP and machine learning models like BERT, TF-IDF, and Word2Vec are utilized. The Dreaddit dataset, containing important metadata and text content, can be useful for NLP exercises, social media activity analysis, and stress analysis. The dataset is a text corpus containing the body, community name, label field, and other fields from a Reddit post. Preprocessing entails tokenization and text cleanup using NLP methods. For smaller datasets, conventional machine learning techniques are used.

This paper used the following methodologies: Removal of noise to improve text classification, Keywords combined to facilitate BERT and ELMo tokenization, ELMo and BERT used to create word embeddings and tokenization, ELMo model trained on cleaned text to form vectors as word embeddings, Machine learning algorithms trained on vectors to compare results.

RAKE (Rapid Automatic Keyword Extraction) is an algo-

rithm that uses stoplists and scoring systems to extract relevant phrases and words from a target text. Elmo vectors generate an array of forms with 1024 parameters from contextualized encodings of word strings in each post. ML models optimized by the use of multiple methods to extract features. In comparison to other test classifiers, LR, SVM, and XGBoost provide better results, with Logistic Regression having the greatest F1 score. The highest F1 score is produced using bag-of-words embeddings with weighted TF-IDF vectors.

The foundational mathematics and stochastic nature of each method determine the accuracy, precision, and recall of a model. While the LR model employs hyperparameters to change performance measures, XGBoost builds new models using decision trees and other mathematical methods. The best F1 score that XGBoost has received is 0.70. SVMs use hyperplanes to differentiate between dataset classes while LSTM and BERT models extract distinctive characteristics to boost accuracy, giving them a combined F1 score of 0.76. The best results were from LR and SVM, with ELMo vectors and BERT embeddings outperforming the Bag-of-Words model in terms of performance.

Authors' mental states can be better understood by examining posts they have published across different platforms. Using PRAW API data, this paper successfully acknowledged stressful posts and produced significant results using conventional criteria for assessment. Important NLP research areas include sentiment analysis, text categorization, opinion mining, Word2vec, GloVe, and tf-idf vectors, which can enhance translation, author identification, spam detection, and language identification. It is possible to perform multimodal sentiment analysis by fusing text and visuals from different sources.

This study proposes using machine learning and natural language processing (NLP) to identify indicators of mental stress in social media posts. Using innovative techniques like ELMo embeddings along with established models like SVM can enhance the analysis of mental stress in social media. Future research exploring neural network-based models and pre-trained language models can build on the findings of this study. Finally, this proposed method has the potential of identifying symptoms and offering assistance to lessen mental health problems.

In this research paper [7], titled "Dreaddit: A Reddit Dataset for Stress Analysis in Social Media," the authors address the prevalent issue of stress in the online world, particularly in social media platforms. Stress is a nearly universal human experience, and while it can serve as a motivator, excessive stress has been associated with negative health outcomes. Recognizing the importance of stress identification across various domains, the authors introduce the Dreaddit dataset, a comprehensive text corpus comprising over 190,000 posts from five distinct categories of Reddit communities.

The dataset's objective is to facilitate stress detection and analysis, with potential applications in fields such as diagnosing physical and mental illnesses, monitoring public sentiments in politics and economics, and assessing the impact of disasters. Unlike existing computational research that predominantly focuses on stress analysis in speech or short texts like Twitter, Dreaddit offers lengthy multi-domain social media data, allowing for a more comprehensive understanding of stress expression.

The research team carefully selects ten subreddits from five domains, covering topics such as interpersonal conflict, mental illness (anxiety and PTSD), and financial need. These subreddits serve as platforms for users to share personal experiences, seek advice, and provide support related to stressful situations. The average post length in the dataset is 420 tokens, significantly longer than microblog data, enabling in-depth examination of stress expressions.

To train supervised models for stress detection, the authors label 3,553 segments from the dataset using human annotators recruited via Amazon Mechanical Turk. The labeled data ensures a balanced representation of stressful and non-stressful content, with 52.3% of the data labeled as stressful. Furthermore, the research team conducts a thorough data analysis, examining vocabulary patterns, lexical diversity, and syntactic complexity in each domain. The analysis reveals significant distinctions among domains, influencing the classification system's performance.

The authors experiment with various supervised models, including Support Vector Machines (SVMs), logistic regression, Naive Bayes, Perceptron, and decision trees. Input representations such as bag-of-n-grams, pre-trained Word2Vec embeddings, Word2Vec embeddings trained on the dataset, and BERT embeddings are tested to identify the most effective model. Additionally, neural models like bidirectional Gated Recurrent Neural Network (GRNN) and Convolutional Neural Network (CNN) are trained and compared to traditional models and BERT.

The best-performing model is a logistic regression classifier utilizing domain-specific Word2Vec embeddings, high-correlation features, and high-agreement data. This model achieves an impressive F1-score of 79.80 on the test set, comparable to the state-of-the-art BERT-based model. Although neural models demonstrate potential, their performance is hindered by the relatively small dataset size.

An error analysis of the models reveals that both tend to overclassify stress, particularly with low-agreement data and less explicit stress expressions. To further enhance stress detection accuracy, future work will focus on incorporating the context and intentions of the writers.

In conclusion, the Dreaddit dataset provides valuable insights into stress expressions in social media. The research's significant contributions lie in the development of accurate supervised models and the provision of a resourceful dataset for further research in stress analysis. The findings emphasize the importance of domain knowledge and lexical features in stress detection, and the potential for utilizing large unlabeled datasets in neural models to improve performance.

III. Methodology

In our dataset, we are focusing on the text part, which contains stress and non-stress texts. As the target or prediction

result, we took the label column. The label column contains either stress text or non-stress text. Our motive is to focus on a comparison analysis between classical machine learning models and deep learning models. At the very beginning, all features except text and label columns were dropped. As the text is natural, it is not well formatted. The texts contain HTML tags, punctuation, different symbols, and stopwords. These things do not carry any meaning in the sentences. Even they do not have that much effect on stress and non-stress classification. So after normalization and removing unnecessary symbols, we turned the text into a straightforward text that contains only words. To get the root forms of words, stemmers were used. These things were done in the preprocessing phase. To incorporate this data into the model, we must vectorize the words so that they can be understood. There are several word embedding options, including word2vec, TF-IDF, countervectorizer, and rapid text, among others. We attempted to use the word2vec and countervectorizer tools. Due to the fact that word2vector converts vectors to negative values, a few models were also challenging to adapt. Consequently, the word2vec embedding did not produce a great deal of satisfactory outcomes. As a result, a countervectorizer was chosen for the word embedding assignments to facilitate further comparison.

To start our study, we carefully chose several traditional machine learning models known for their distinct abilities. The initial models we picked were Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), Multinomial Naïve Bayes, Random Forest, Decision Tree, Multi-layer Perceptron, and Adaboost. This selection was intended to enable a solid comparison and analysis at a fundamental level, helping us uncover the complexities of the data.

We had multiple reasons for choosing these non-neural models. Firstly, they represent a wide range of machine learning techniques, covering both linear and non-linear methods. This allowed us to explore different ways of modeling stress prediction, giving us a thorough perspective of the problem. Additionally, these models are well-established in the machine learning field and have a successful track record across various tasks, making them suitable candidates for our initial investigation.

Our aim in using these models was two-fold. Initially, we expected these models to provide useful baseline results that could act as reference points for assessing more advanced models. Secondly, we aimed to identify any inherent patterns or tendencies in the data that could guide us toward the most effective modeling approach.

While the initial non-neural models provided valuable insights, we saw room for improvement. One notable area for enhancement was incorporating neural models, specifically those employing attention mechanisms. Neural models, particularly those with attention mechanisms, have shown a remarkable ability in capturing complex patterns and relationships in intricate datasets. The unique feature of attention mechanisms, allowing the model to focus on specific parts of the input data, could be crucial in identifying subtle cues related to human stress levels.

Introducing neural models with attention mechanisms could potentially enhance our stress prediction performance. The adaptable nature of attention allows the model to dynamically assign importance to different features, helping it recognize and prioritize relevant information. This aligns well with the complex nature of human stress, which often involves intricate interactions between various factors. Moreover, neural models can learn hierarchical representations, potentially revealing hidden features that conventional machine learning models might miss.

In summary, our research paper conducts a comparative study of predicting human stress. We begin by examining various traditional non-neural machine learning models, which provide a solid foundation and prepare the groundwork for introducing neural models with attention mechanisms. This approach is expected to lead to improved performance and a deeper understanding of the factors influencing human stress levels.

IV. DATA ANALYSIS

As our motive is to work with stress and not stress text classification, so we chose a dataset from Kaggle that is scrapped and labelled. It contained texts from social media from different users, and according to their texts the stress and not stress is defined. This dataset was also used on a published paper, which inspired us to explore it again. The dimension of the dataset is (3553, 116). We chose all the rows and two columns for our project. The rest of the columns were dropped during pre-processing.

Before doing preprocessing, the duplicate values were checked. From the whole dataset, there were no duplicates in the text column. So all our 3553 values were unique and balanced. After the duplication test, our text column went through a preprocessing phase. As it is real world data, it has many unnecessary symbols, characters, and marks that do not carry any meaning. Those were removed during the preprocessing phase. Besides, there are multiple forms of words, which makes for a scattering of words. To unite them to root form of words, stemming was done. (i.e. looking → look, looks → look etc)

	text	label
0	Its like that, if you want or not.“ ME: I have...	No Stress
1	I man the front desk and my title is HR Custom...	No Stress
2	We'd be saving so much money with this new hou...	Stress
3	My ex used to shoot back with "Do you want me ...	Stress
4	I haven't said anything to him yet because I'm...	No Stress
5	Thanks. Edit 1 - Fuel Receipt As Requested. <u...	No Stress

Fig. 1: Before Preprocessing

	text	label
0	like want not“ problem take longer ask friend ...	No Stress
1	man front desk titl hr custom servic repres jo...	No Stress
2	wed save much money new housrit expens citi go...	Stress
3	ex use shoot back want go time matter almost w...	Stress
4	haven’t said anyth yet i’m sure someon would t...	No Stress

Fig. 2: After Preprocessing

After doing so we split the training and testing data in an 80:20 ratio. In total, we checked the stress and not stress total data were almost equal. Though the not stress data was 200 more. As the data split was randomly shuffled so it probably won’t affect that much to the training phase.

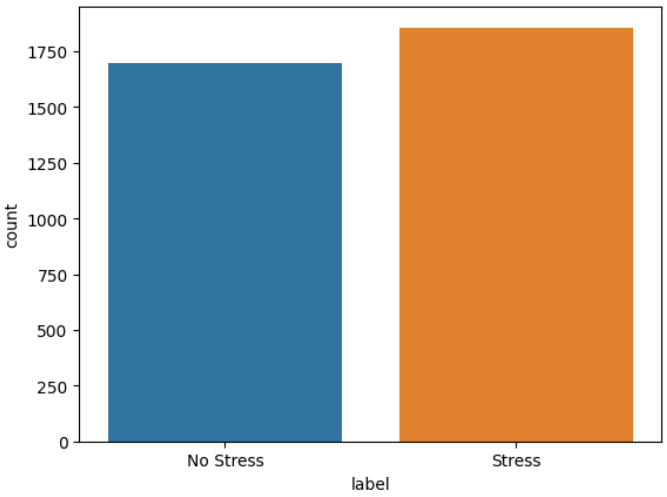

Fig. 3: Stress and not stress ratio

A. *WordCloud*

WordCloud is a visualization tool by which we can picturize words according to their importance or frequency. From our first sample, we can see the word ‘feel’ is used mostly in the context. As the dataset was taken from “Stress Analysis in Social Media”, we can understand that the users mostly used words like ‘feel’,’know’,’time’,’even’,’want’,’really’,’help’. The size of the words are the induction of their frequency. More the words are found in the corpus, the larger it appears in the word cloud. This word cloud was generated based on the label.

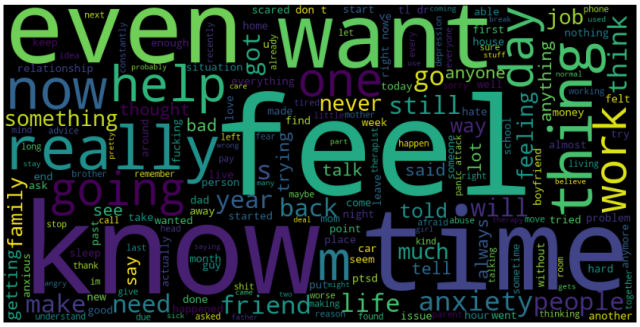

Fig. 4: Word Cloud Base on Label

If we generate the word cloud based on the sentiment of the corpus the word sizes and colors look different.Sentiment wise,the word ’know’ gets the most highlights in this corpus After that, rest of the words like ’want’, ‘feel’,’time also key aspects of this word cloud indicating the importance of these word to understand the sentiment of this corpus.This is and effective content prioritization. From this, we can tell which words are associated with the sentiment of the users.

Fig. 5: Word Cloud Base on Sentiment

This following pie chart shows the cause of people’s stress. From the chart we can see that most people’s stress is caused by ptsd. Around 20% of people are having stress because of ptsd. In second place,approximately, 19% of the people are stressing over relationship problems. Another leading cause of people’s stress is anxiety. Anxiety is responsible for around 17% of people’s trees which makes it the third leading cause of stress. Besides these three, we can observe that there are other factors that are causing stress in people’s lives. Some such causes are domestic violence, assistance,survivors of abuse etc. We can visualize the ratio of these factors with respect to the three leading factors from our pie chart. In short, this pie chart presents a comparison between the factors of stress and gives the insight of the leading causes of stress.

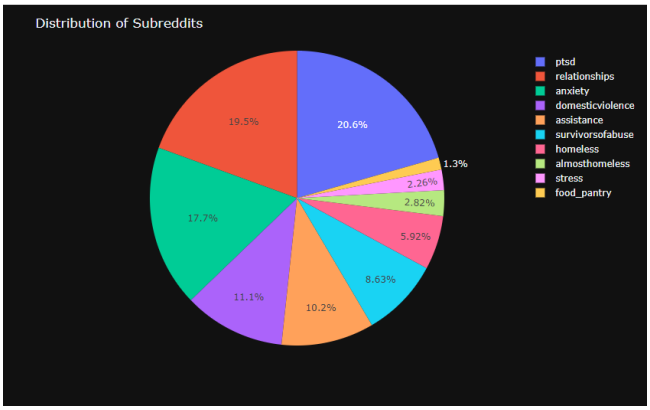

Fig. 6: Distribution of stress types

V. Result Analysis

All the confusion Matrix

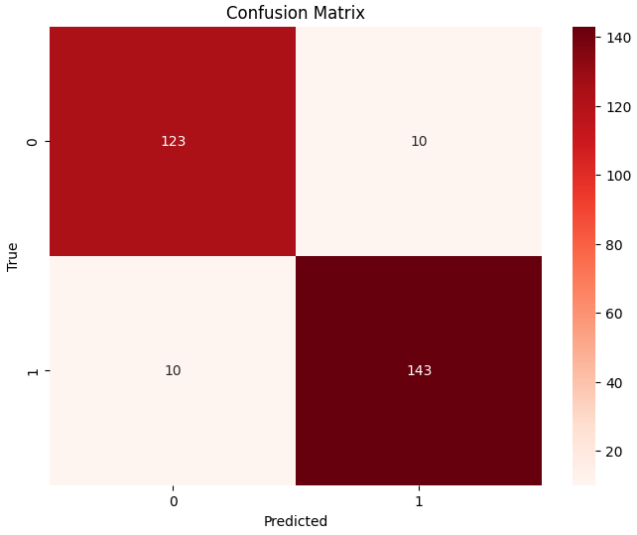

Fig. 7: Confusion Matrix report for the Decision Tree classifier model

From this Confusion Matrix report for the Decision Tree classifier model, we got an accuracy rate 60.3%. That indicates that the model was able to classify 536 instances correctly out of the total 889 instances. The precision for “No Stress” we got from this model is 59% which means 59 instances out of each 100 instances predicted as ”No Stress” are actually correct. For “Stress” it is 61%. On the other hand, the recall measures the ability of the model to correctly identify. The recall value for “No Stress” we got 57% and 63% for “Stress” prediction. The F1-score of “Stress” and “No Stress” is 0.62 and 0.58 respectively. In summary, this model shows a decent accuracy of about 60.3%, with similar performance for both classes. The precision, recall, and F1-score are not extremely high but we can expect a balanced performance in distinguishing between ”No Stress” and ”Stress” instances.

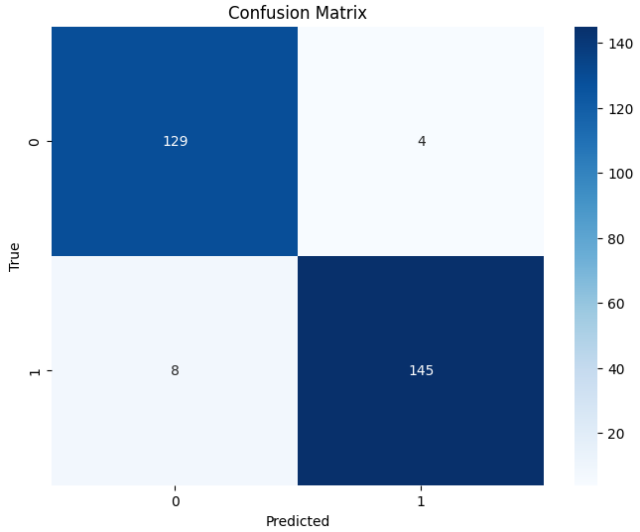

Fig. 8: Confusion Matrix report for the MLP classifier model

The next model we examined is MLP classifier. From this model we got an overall accuracy of 69.4%. This means the model is able to correctly classify about 617 instances out of 889 instances.The precision for “No Stress” of this model is 70% which means 70 instances out of 100 instances predicted as ”No Stress” are actually correct. For “Stress” the model was able to perform with 69% precision. So the precision of this model is higher than our previous model. The recall of this model is also above our previous model. We got 75% and 63% for two classes ”No Stress” and ”Stress”. Similarly, this model again topped the Decision Tree classifier model in terms of the F1-score. It has an F1-score of 0.67 for the ”No Stress” class , and for the ”Stress” class, it is 0.72. In a nutshell, the model shows a more balanced performance in distinguishing between ”No Stress” and ”Stress” instances.. The precision, recall, and F1-score values also suggest that the model is moderately effective for the task.

Another reputed classifier we used is the AdaBoost classifier. The AdaBoost classifier showed an overall accuracy of approximately 67.4%. The model classified the ”No Stress” class with the precision of 66%. For the ”Stress” class, it was 68%, meaning that 68% of instances predicted as ”Stress” were indeed correct. We got 66% and 69% as the recall from this model for the ”No Stress” and ”Stress” class. This model provides an F1-score 0.66 for the ”No Stress” class, and 0.69 for the ”Stress” class. In short, this model performed better than many models but yet MLP classifier works better for this particular task.

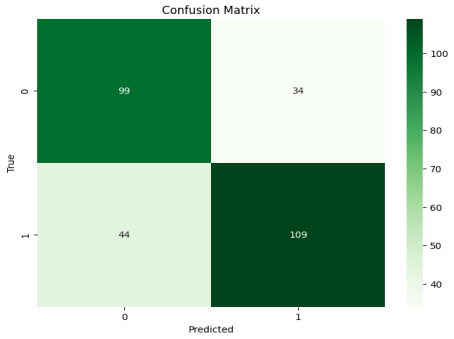

Fig. 9: Confusion Matrix of the AdaBoost classifier model

We use another non-parametric, supervised learning algorithm K-nearest neighbors (KNN). As the overall accuracy is 0.5639, 56.39% of the data points are correctly identified. The accuracy is the percentage of positive data points that are actually positive when they were classified. The recall is the percentage of accurately classified positive data items. The recall value for No Stress is 0.54, indicating that 54% of the positive data points were identified correctly. The precision for class Stress is 0.65, indicating that 65% of the data points that were classified as positive were actually positive. The ”No Stress” class has a f1-score of 0.59. The precision for the ”Stress” class is 0.60, that means that 60% of the data points that were classified as “Stress“ were actually “Stress“. Recall for the ”Stress” class is 0.48, which indicates that 48% of the ”Stress” data points were correctly identified. The ”No Stress” class has a f1-score of 0.53. The model fails to do a good job of classifying data points into an appropriate class, evidenced by the low precision and recall for each class.

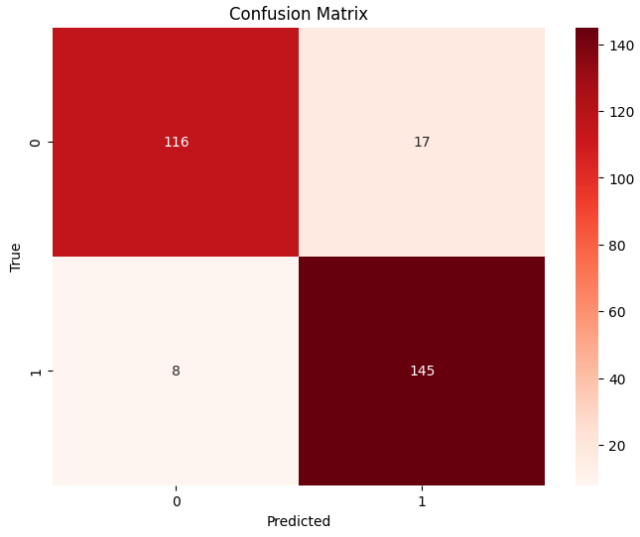

Fig. 10: Confusion Matrix report for the K-nearest neighbors (KNN) classifier model

We use a supervised machine learning algorithm named Support Vector Machine(SVM). As the overall accuracy is 0.673699, 67.37% of the data points are correctly identified. The accuracy is the percentage of positive data points that are actually positive when they were classified. The recall is the percentage of accurately classified positive data items. The recall value for No Stress is 0.65, indicating that 65% of the positive data points were identified correctly. The precision for class Stress is 0.63, indicating that 63% of the data points that were classified as positive were actually positive. The ”No Stress” class has a f1-score of 0.65. The precision for the ”Stress” class is 0.68, that means that 68% of the data points that were classified as “Stress“ were actually “Stress“. Recall for the ”Stress” class is 0.71, which indicates that 71% of the ”Stress” data points were correctly identified. The ”No Stress” class has a f1-score of 0.69. The model is effectively classifying data points into the appropriate classes, as indicated by the relatively good precision and recall for each class.

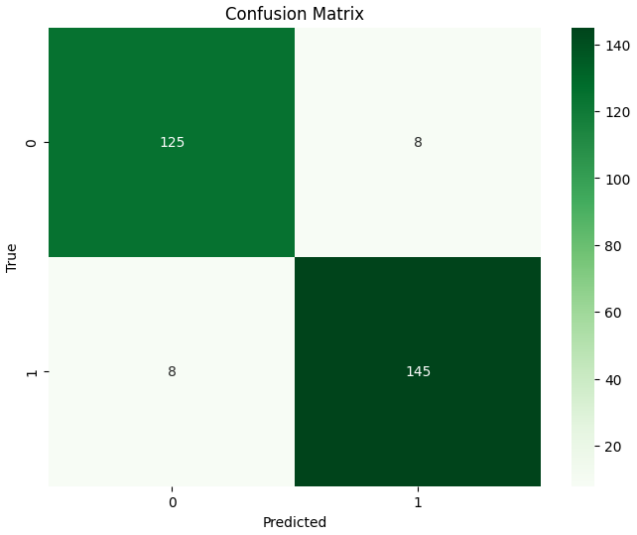

Fig. 11: Confusion Matrix report for the SVM classifier model

Random forest uses multiple decision tress and shows the average result.Our implementation random forest classifier model shows accuracy overall 0.700421 , in other words 70% accuracy overall. No stress Precision value of 0.74 means that out of all the instances that the model classified as not stressed ,74% of them were actually not stressed. In similar ways stress precision 0.67 means that out of all classified as stressed, 67% of them actually stressed. Here recall value is 0.58 for no stress classification and 0.81 for stress classification. Recall value is lower since it try to capture all relevant instances while precision focus on true positives.F1 score takes into account both false positives and false negatives and it provides a single value that reflects the overall effectiveness of the model. F1 score is 0.65 for no stress classification, which means model

has achieved a balanced performance in terms of precision and recall. F1 score is 0.74 for stress classification which is a desirable score which indicates strong overall performance.

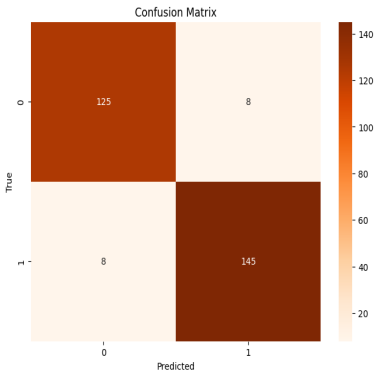

Fig. 12: Confusion Matrix report for the Random Forest classifier model

Naive bayes is popular machine learning algorithm used for classification task. It is based on bayes theorem. It is called naive because it assumes that all features are independent of each other. Multionomial Naive bayes is special variant which is suited for text classification task. Our implementation of naive bayes model shows overall accuracy of 0.7158 or 71.58%. Precision value for no stress is 0.78 which means 78 percent people are not stressed out of all no stressed classified. Recall value for no stress is 0.58 which is significantly less than precision value. However f1 score for no stress is slightly higher than recall value which is 0.66. Similarly precision value for stress is 0.68 which a lot lesser than no stress classification. But stress recall value is 0.85 which is a great score and significantly higher than no stress recall value. Similarly f1 score for stress classification is 0.75 which is higher than no stress f1 score. 0.75 means that the model displays strong performance.

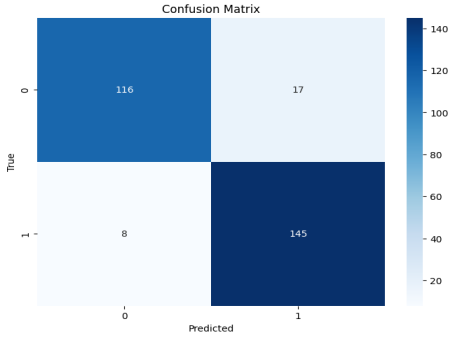

Fig. 13: Confusion Matrix of the Multinomial Naive bayes model

In our analysis, we evaluated the performance of several non-neural machine-learning models for predicting human stress levels. We measured accuracy as well as F1 scores for stress and no-stress classes to provide a comprehensive understanding of each model's capabilities.

Among the models, Multinomial Naïve Bayes exhibited the highest accuracy with a score of 0.716, closely followed by Random Forest with an accuracy of 0.700. Support Vector Machines (SVM) and Multi-layer Perceptron (MLP) achieved accuracies of 0.674 and 0.682, respectively. AdaBoost demonstrated a slightly lower accuracy of 0.681. However, K-Nearest Neighbors (KNN) and Decision Tree lagged behind with accuracies of 0.564 and 0.613, respectively.

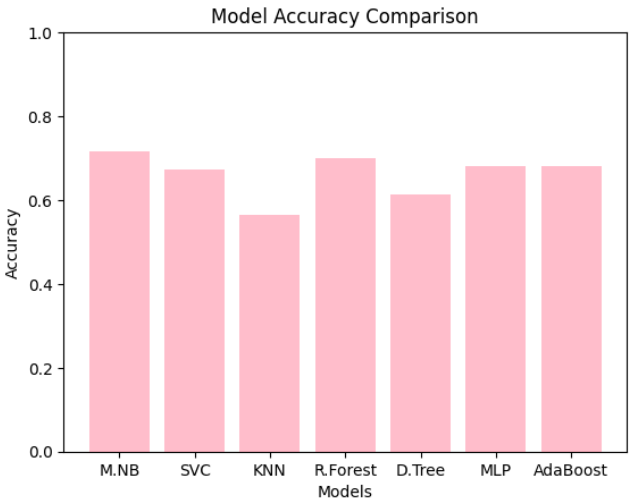

Fig. 14: Accuracy Comparison

When focusing on F1 scores, a metric that considers both precision and recall, the models' performances differed across stress and no-stress classes. Multinomial Naïve Bayes attained the highest F1 score for stress prediction at 0.75, indicating a good balance between precision and recall. Random Forest and MLP also performed well in this category, achieving F1 scores of 0.74 and 0.71, respectively. However, Decision Tree and K-Nearest Neighbors yielded lower F1 scores for stress, at 0.64 and 0.53, respectively.

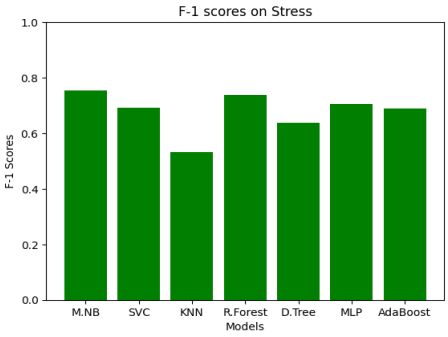

Fig. 15: F1-Score comparison for Stress

For the no-stress class, Multinomial Naïve Bayes and SVM achieved F1 scores of 0.66 and 0.65, respectively. AdaBoost displayed a relatively higher F1 score of 0.67, indicating its proficiency in identifying instances without stress. On the other hand, K-Nearest Neighbors and Decision Tree produced F1 scores of 0.59 for the no-stress class, indicating challenges in effectively distinguishing such cases.

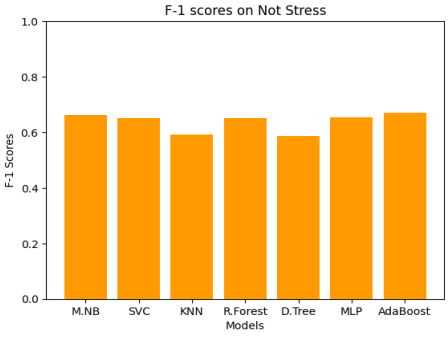

Fig. 16: F1-Score comparison for No Stress

Comparing these results, it becomes apparent that certain models excel in stress prediction while struggling with no-stress instances, and vice versa. The Multinomial Naïve Bayes model demonstrates strong performance across multiple metrics, particularly in stress prediction. Random Forest and MLP also exhibit noteworthy abilities in this context. On the other hand, while SVM and AdaBoost perform relatively consistently, K-Nearest Neighbors and Decision Tree show limitations, especially in capturing the complexities of stress and no-stress instances.

In conclusion, the comparative analysis of non-neural machine learning models for human stress prediction reveals nuanced strengths and weaknesses. As discussed earlier, these results lay the groundwork for exploring the potential benefits of incorporating neural models with attention mechanisms, which may further enhance the accuracy and predictive power of stress prediction systems.

REFERENCES

[1] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *NPJ digital medicine*, vol. 5, no. 1, p. 46, 2022.

[2] B. L. Cook, A. M. Progovac, P. Chen, B. Mullin, S. Hou, E. Baca-Garcia, *et al.*, "Novel use of natural language processing (nlp) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in madrid," *Computational and mathematical methods in medicine*, vol. 2016, 2016.

[3] R. Qasim, W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi, "A fine-tuned bert-based transfer learning approach for text classification," *Journal of Healthcare Engineering*, vol. 2022, p. 3498123, Jan 2022. [Online]. Available: https://doi.org/10.1155/2022/3498123

[4] T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions," *Journal of Big Data*, vol. 9, no. 1, pp. 1–24, 2022.

[5] P. Kumar, P. Samanta, S. Dutta, M. Chatterjee, and D. Sarkar, "Feature based depression detection from twitter data using machine learning techniques," *Journal of Scientific Research*, vol. 66, no. 2, pp. 220–228, 2022.

[6] S. Inamdar, R. Chapekar, S. Gite, and B. Pradhan, "Machine learning driven mental stress detection on reddit posts using natural language processing," *Human-Centric Intelligent Systems*, vol. 3, no. 2, pp. 80–91, 2023.

[7] E. Turcan and K. McKeown, "Dreaddit: A Reddit dataset for stress analysis in social media," in *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 97–107. [Online]. Available: https://aclanthology.org/D19-6213