Stress Analysis in Social Media

1st Nadeen Mohamed Farid
Nile University
N.farid@nu.edu.eg

2st Nouran Hady Shaaban
Nile University
N.Hady@nu.edu.eg

3st Sama Ahmed Okasha
Nile University
S.Okasha@nu.edu.eg

4st Dr. Walaa M.Medhat Asal
Nile University
WMedhat@nu.edu.eg

***Abstract*—In order to do stress analysis, we tried to collect data and insights from various social media platforms. We were able to draw diverse inferences from the data on various social media sites thanks to various data mining techniques. A method for determining the stresses and strains that force on materials and structures create is called stress analysis. We discovered a dataset with information about the person's identity, his present state of mind, and input values for every stress-related consequence that might exist. to ascertain whether or not this person is stressed. To ascertain the degree of accuracy we might attain, we used a variety of models, including Logistic Regression, Decision Tree Classifier, Linear SVC, SGD Classifier, and others. The study provides comprehensive results.**

I. Introduction

The rapid advancement of technology in today's society is causing a variety of stress-related symptoms in people. Stress-related illnesses like anxiety, depression, and ADHD are quite common in younger people. All of this is a result of a number of variables, including dietary preferences, economic and societal pressure, and low self-esteem. Patients frequently struggle to distinguish between situations that are stressful and those that are not related to stress. Methodologies for stress detection can help people and reduce stress. It is necessary to raise awareness of the different stress-related symptoms that would otherwise go unnoticed. Therefore, we made the decision to select a dataset that depicts this issue and analyse it to determine whether or not the majority of people exhibit signs of stress.

We decided to use various machine learning and data mining models on our dataset. Based on the text description of Offord in our dataset, we believe that utilising a pre-trained model with a high accuracy rate will produce accurate and trustworthy results. As By pre-training on the enormous unlabeled corpus to avoid overfitting on small-size data, the pre-trained language models have the advantage of learning universal language. As a result, we compared the results of several to determine which model had the best accuracy.

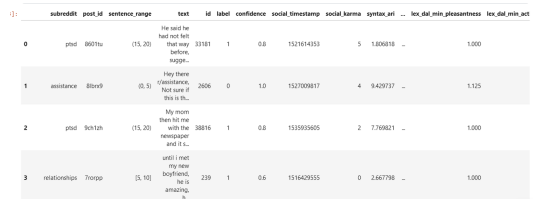
II. Background and literature review

In the last several years, the discipline of stress analysis has seen many incredible advancements that have produced results. Thus, in order to determine a new accuracy, we want to compare our findings to theirs. As a result, we decided to use multiple models in an effort to improve our accuracy and outcome. Due to our labelled data, we decided to apply supervised learning approaches. In addition to the Naive Bayes classifier, we also used linear regression and decision trees. To determine which model best fits our data, which includes the person's anxiety, confidence, social time stamp, and a text in which he expresses how he feels. And these models determine whether or not this person is stressed.

graphicx

A. Dataset

- This dataset was made to determine whether or not this person was under any stress. The dataset is divided into the train and test folders. Each folder contains a CSV file with the information of every person, including their name, text (a description of how they are feeling), confidence level, social time stamp, social karma, and analysis of their attitude—whether they are angry, sad, or socially awkward—time spent with their families, how they are feeling, body language, health, and accomplishments, among other things. There are 117 total columns and 2839 total rows. a specific class of dataset that may be fitted to various machine learning and data mining models.



	subreddit	post_id	sentence_range	text	id	label	confidence	social_timestamp	social_karma	syntax_ari	...	lex_dal_min_pleasantness	lex_dal_min_act
0	ptsd	8601tu	(15, 20)	He said he had not felt that way before, sugge...	33181	1	0.8	1521614353	5	1.806818	...	1.000	
1	assistance	8ibrw9	(0, 5)	Hey there r/assistance, Not sure if this is th...	2606	0	1.0	1527009817	4	9.429737	...	1.125	
2	ptsd	9ch1zh	(15, 20)	My mom then hit me with the newspaper and it s...	38816	1	0.8	1535935605	2	7.769821	...	1.000	
3	relationships	7vorpp	[5, 10]	until i met my new boyfriend, he is amazing, h...	239	1	0.6	1516429555	0	2.667798	...	1.000	

Fig. 1. An image of a dataset

B. Tools and Libraries

The necessary procedures for data prepossessing and post-processing are performed using Python, Pandas, PIL, wordcloud, nltk, sklearn. metrics, and numpy. to properly develop and assess models.

We also used gradio for the deployment part. The User can enter any sentence and the interface will tell whether they are stressed or not.

III. Methodology

The purpose of data pre-processing is to carry out data cleaning, integration, reduction, and transformation. We didn't need to do much because the data has already been cleaned up. We added a new column named mood to help in the modelling process and cleaned the text from any unnecessary quotations that will affect the accuracy of our models, dropped unnecessary columns, and finally encoded the texts to enter them into the model. After looking for any potential missing values in the dataset and visualizing the data, we were able to enter our data into the models of our choice. We divided the training dataset into train and val, fed it into the model, started receiving the results, visualising and contrasting them, and so on.

As we previously mentioned, we tested various models to identify the one that best fits our dataset. We choose to utilise the Naive Bayes model as our initial option. As a result, we prepared our data for it, divided it, and fed it into the model, yielding an accuracy that was satisfactory. Since our issue is a classification problem, we, therefore, choose to apply a different algorithm, the linear regression algorithm. Finally, after visualising the results and attempting to use the decision tree model. We implemented the fed it our processed data, used the prediction, sped up the output, and calculated the accuracy. The accuracy was the least among all the models we tried so far. We used the accuracy score to calculate the accuracy for each model. We compared the outcomes of all the models after visualising all of their outputs. Every model has a unique equation that produces results that are similar in only one way, that they all solve classification problems. Because of this, we decided to apply them to our data.

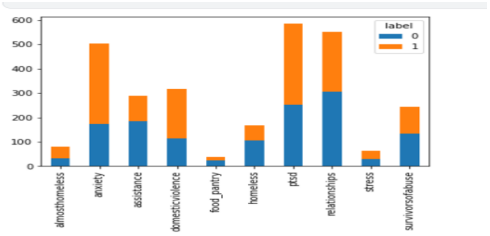


Fig. 2. Visualization stress probability

IV. Results and Analysis

The outcomes of every model we've worked on demonstrated that they can all predict whether a person is stressed or not. However, some models are superior to others. The Naive Bayes Model, which has the best accuracy with our data with an accuracy of 73.79, and Linaer regression, which has an accuracy of 72.44, are the two best models. While the Desicion tree had an accuracy of 62.77 which is the least accurate accuracy so far. However, the results of all the models are similar

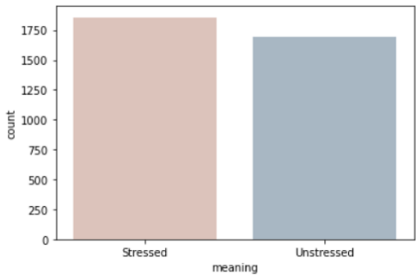


Fig. 3. Checked Preprocessing data is balanced

and have a similar range of predictions. However, the results of all the models are similar and have a similar range of predictions. Therefore, greater prepossessing of our data is required in order to aid in improving accuracy. Additionally, the deployment outcomes were excellent and produced correct findings.

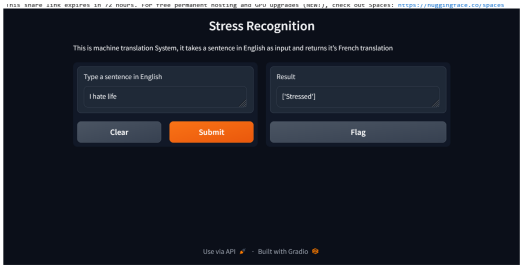


Fig. 4. Results of deployment

V. Conclusion

The goal of this research was to examine the studies that used machine learning approaches to detect and analyse human stress. These studies have demonstrated that it is possible to quickly and effectively provide results to identify the sources of stress and its impact on health with the aid of machine learning algorithms. The conclusion that can be drawn from this is that machine learning techniques are necessary for precise and effective stress analysis and prediction in the modern world. When used on different datasets, different machine learning approaches perform in different ways.

VI. Future Work

Our long-term goal is to apply machine learning techniques to diverse datasets other than this one in order to achieve high accuracy. constructing algorithms that take advantage of hybrid machine learning methods. Application of machine learning algorithms on global datasets and the use of additional machine learning algorithms to evaluate the effectiveness of the model projected. Additionally, enhance the data-cleaning process. experimenting with various neural network models on additional datasets.

REFERENCES

[1] R. Bhatia, Stress analysis in social media, https://www.kaggle.com/datasets/ruchi798/stress-analysis-in-social-media/discussion?select=dreaddit-train.csv

[2] T. Nijhawan, G. Attigeri, and T. Ananthakrishna, "Stress detection using natural language processing and machine learning over social interactions - Journal of Big Data," SpringerOpen, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00575-6

[3] U. Bhushan and S. Maji, "Prediction and analysis of stress using Machine Learning: A Review," SpringerLink, https://link.springer.com/chapter/10.1007/978-981-19-3148-2$_3$5