DREADDIT: A REDDIT DATASET FOR STRESS ANALYSIS IN SOCIAL MEDIA

BrainStation Data Science Diploma Capstone Project

Sarah Lipoff

BrainStation | Data Science

September 29, 2022

Problem Statement & Background:

In today's social media world, users send and receive messages numerous times a day. This way of living makes stress more obvious and easier to observe than ever before. A tremendous amount of data on stress is easily accessible through platforms like Facebook, Twitter, and Reddit. While some stress can be motivating, too much stress is associated with many unhealthy outcomes. Recent studies have shown that overall stress levels have increased tremendously since the dawn of social media. This report intends to model the relationship between certain words, and stress levels to better detect stress before it worsens into problems such as depression and anxiety.

About the Data:

In the Dreaddit datasets, social media text was collected for detecting the occurrence of stress. There are different categories of social media posts in the dataset, each including text detection on which does and doesn't predict stress from the person posting. The dataset is publicly available at https://paperswithcode.com/paper/dreaddit-a-reddit-dataset-for-stress-analysis-1. The attached notebooks go through the cleaning, preprocessing, analyzing, visualizing, and modeling of the data.

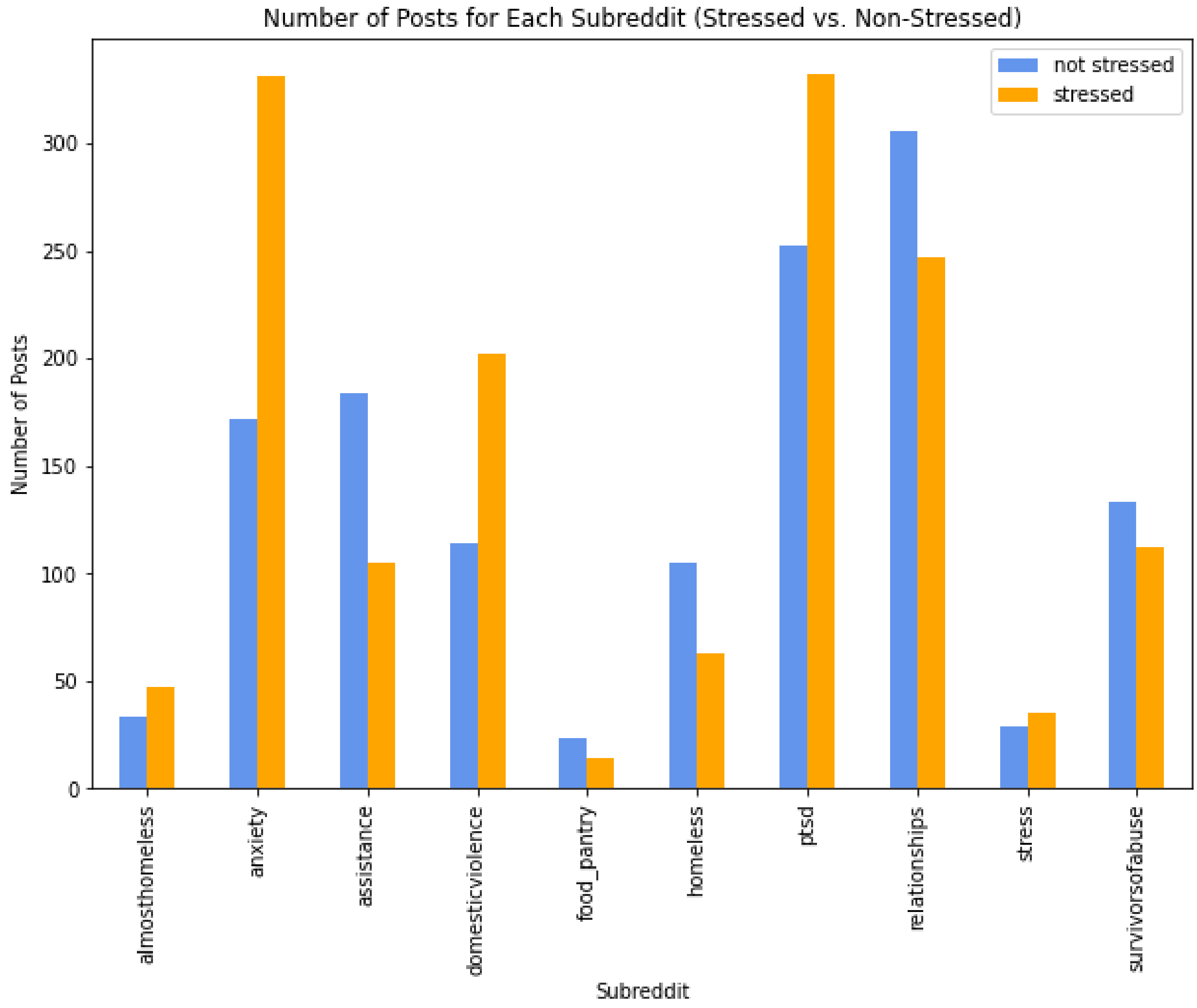
Elsbeth Turcan and Kathleen McKeown write about the topic and data. The dataset consists of 190K posts from five different categories of Reddit communities. Additionally, 3.5K total segments taken from 3K posts using Amazon Mechanical Turk have been labeled.

The original datasets came already split into train and test subsets. The training data started out with 2,838 rows and 116 columns, while the test data had 715 rows and 116 columns. The focus of this report is on the post ID, subreddit, text, and label columns.

Column Name	Column Contents	Data Types
Subreddit	The Subreddit topic	String
Post ID	Unique ID number for each post	Integer
Text	The actual text from posts	String
Label	If stress is present or not	Integer

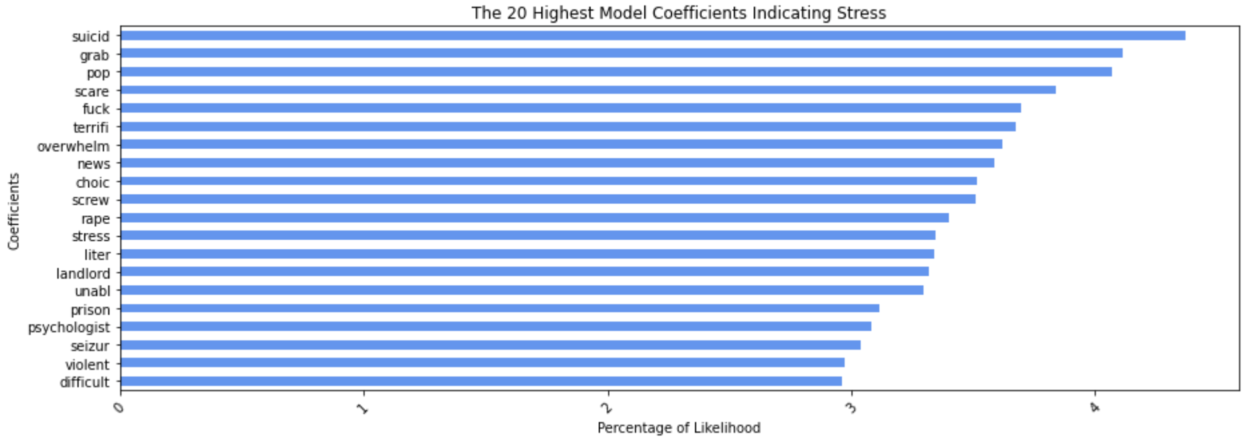
Data Cleaning, Exploratory Data Analysis (EDA), & Preprocessing:

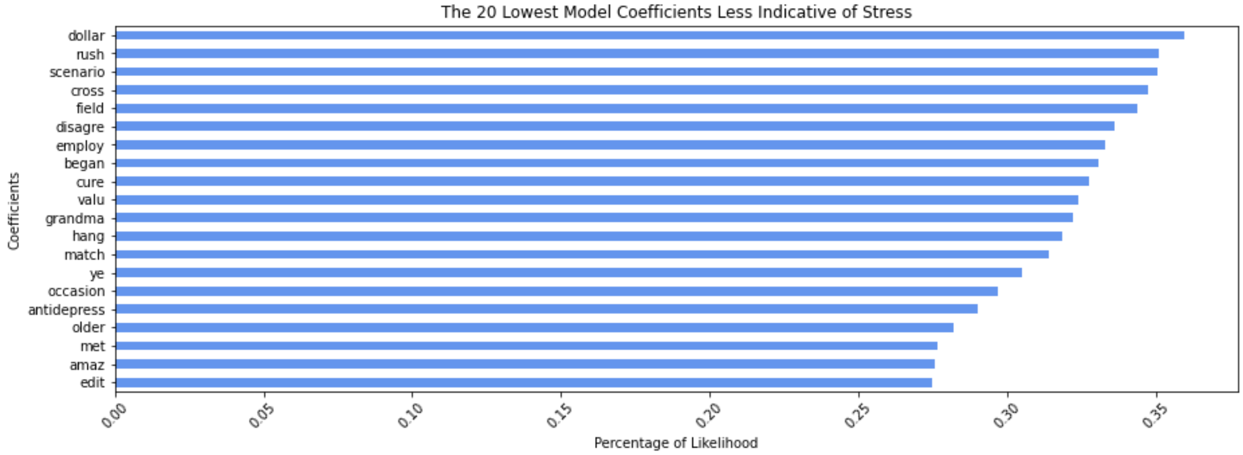
Most of the posts had approximately 65-90 words per post, with 350-400 words being the most common character count per post. Looking at the different subreddits, ‘ptsd’, ‘relationships’, and anxiety’ have by far the most posts. I also looked at each subreddit with its associated number of posts. Only a little less than half of the posts under the ‘anxiety’ subreddit were marked as experiencing stress which seemed surprising to me considering the subreddit (see below chart).



Looking at the sparse matrices, I was able to reduce the number of features from 11,516 to 2,356! The feature reduction enables the model to learn on fewer words. Doing so will reduce overfitting and training time (by removing any unnecessary noise from the data), hopefully improving the model's performance and accuracy.

The plots below show the top and bottom 20 coefficients, or words that are most and least likely to be associated with stress. For example, someone who writes text using the word “suicide” is more than four times more likely to be experiencing stress than a word with a coefficient of about 1. On the other hand, someone who writes text using the word 'dollar' is less likely to be experiencing stress. Since these are the words that have the lowest coefficients, the percentage is a fraction. I found it very interesting that the words 'antidepressant' and “hang” were on the least likely to be experiencing stress list. I would’ve expected them to be under the most likely to be experiencing stress list. It’s the same for the word “employ”. I think it was originally “employer” or “employment”, which would likely be associated with stress as well. Although there are words here that seem surprising, there are also words that make sense to me. The word “cure” seems like it belongs on this list of less likely to be experiencing stress.



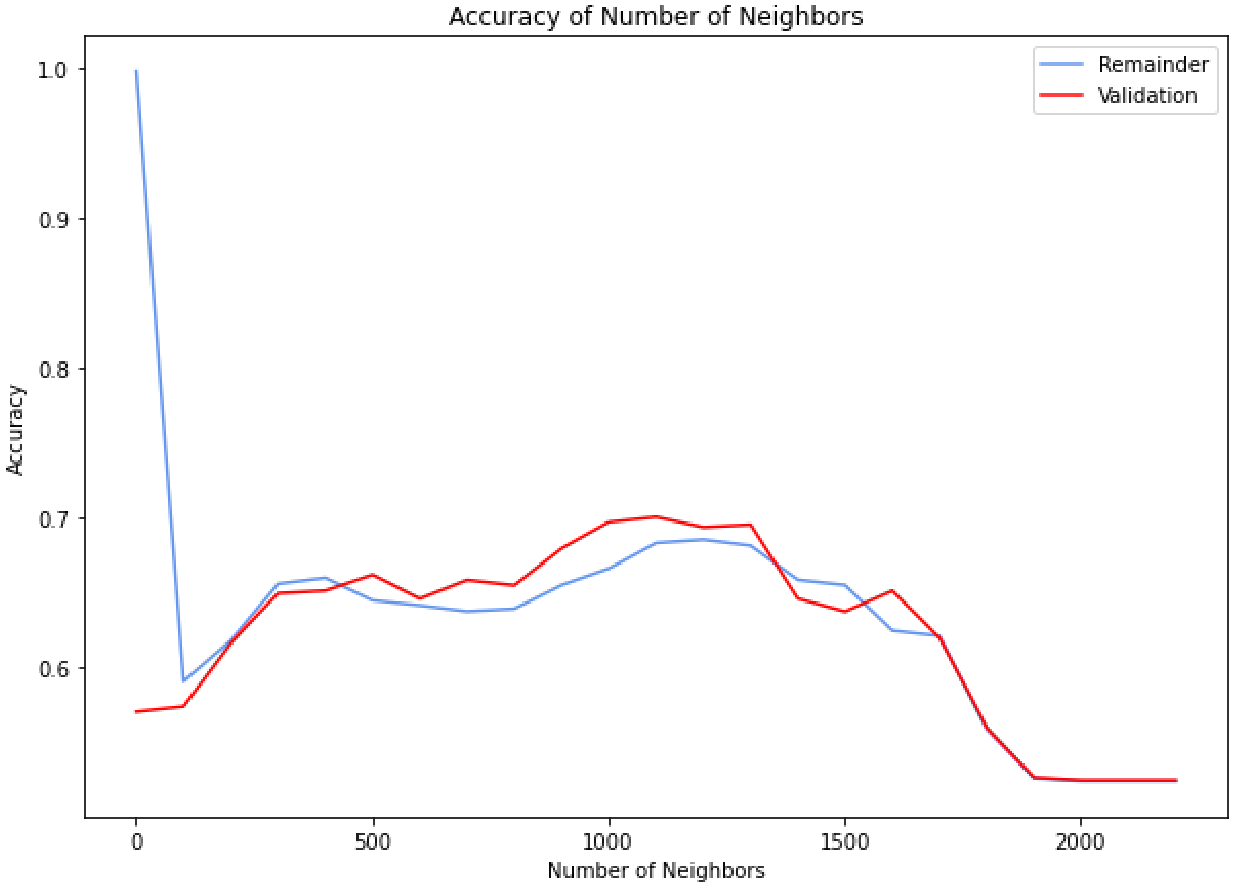


Insights, Modeling, & Results:

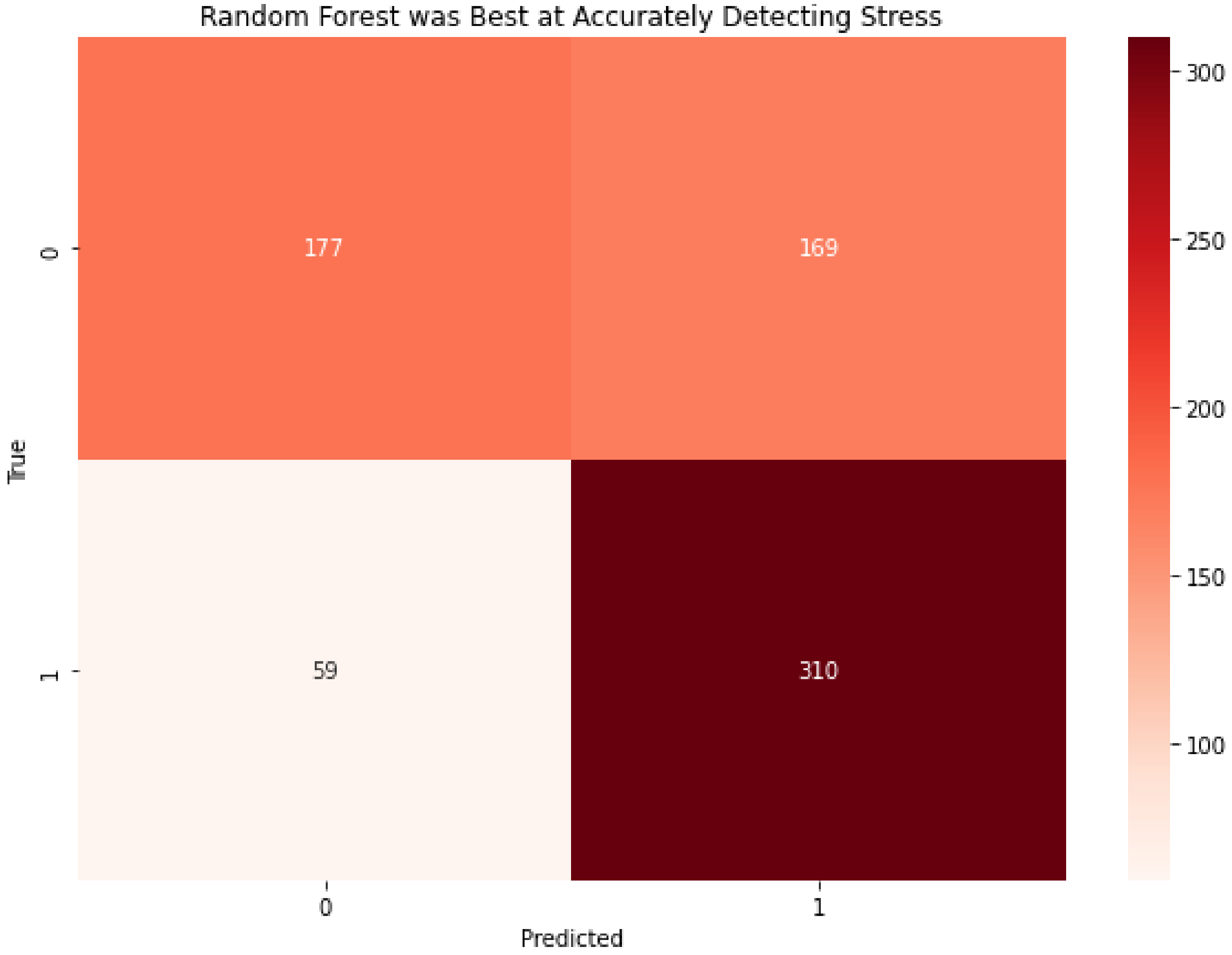
For the modeling, I started with logistic regression and tested for the best value for C. The optimal C value turned out to be low, which helped with some of the overfitting. It also enabled the model to better generalize the new data coming in. From the logistic regression, 79% of the 369 stressed individuals were correctly identified as stressed.

The following step was testing a decision tree, looking at the optimal max depth parameter. Only 57% of stressed individuals were accurately predicted as stressed.

Next was a KNN model. Considering that an increased number of neighbors increases the computational cost, I looked for the optimal number of neighbors. I also included weights and the number of jobs in the pipeline as well. I first took a small subsample of the data and slowly increased the sample size. I then updated the KNN with the optimal parameters. The kernel kept on crashing while trying to score the KNN. After hours of work, I realized that the sample size wasn't the issue, but rather the number of jobs was causing the crashing. I then commented out the number of jobs and the cells ran. After all that work, I saw that the score only improved slightly and there was still a lot of overfitting. Looking at the below visual, I saw that the accuracy dropped at 1 neighbor and then increases and fluctuates at about 100 neighbors. It begins to drop dramatically again after the approximate 1,100 neighbors' peak. The recall on the KNN was very low, accurately predicting only 34% of stressed people as stressed. This was not a good model to predict stress, and one might as well just flip a coin to see if they are stressed.



Finally, I tested out a random forest, optimizing the number of jobs, max depth, and the minimum number of samples required at a lead node. The model accurately predicted 84% of stressed individuals as stressed! This can be seen by looking at the below heatmap. The "1" means that it's the positive (or stressed) class, while the "0" refers to the negative (or not-stressed) class. The darker colors are for the higher numbers and you can see that 310/(59+310)=84.01, or 84%. This seems to be the best of the bunch of models tested.



Overall, the models that performed best (in order) were the random forest, logistic regression, decision tree, and KNN. See below for the percentages of the recall on the positive class for each model. These are the percentages of the stressed individuals accurately predicted as stressed.

Model	Recall on the positive class
Logistic Regression	79%
Decision Tree	57%

KNN	34%
Random Forest	84%

Findings and conclusion:

Following the hyperparameter optimization and scoring of each model, I looked at the classification reports and confusion matrix. Since knowing one is stressed is crucial for dealing with the stress, I decided to focus on the recall for each model. I ran a logistic regression, decision tree, KNN, and random forest. Of those four, the random forest has the highest recall. It accurately predicts 84% of stressed individuals as stressed!

Using the recall percentages, I may be able to help people detect their stress early on, preventing others from dealing with worse physical and/or mental health problems. There are a few potential next steps that can be taken. Firstly, I can go back and collect more data by contacting the authors and figuring out a way to do additional annotation. I can also potentially manually label more data and build a semi-supervised learning model that would label the text for me. Another idea would be to build a simple app (with some big disclaimers) to see if people are stressed. Being cognizant of the fact that people's mental health can be complicated, I would have to investigate the liabilities of doing so first though. Lastly, I could try neural networks as a model to see if we can get higher recall scores. Google's latest is the GTP-3 (Generative Pre-Trained Transformer) model. Google is also planning to release a new NLP GTP-4 model shortly. The new GTP-4 is "expected to have around 100 trillion parameters and will be five hundred times larger than GPT-3". By taking these next steps, I can have the potential to help prevent many people from dealing with avoidable physical and mental health problems.

References:

https://www.analyticsinsight.net/will-gpt-4-be-a-massive-language-model-that-can-imitate-human-brain/

https://paperswithcode.com/paper/dreaddit-a-reddit-dataset-for-stress-analysis-1