Brook Zeleke
Anooj Pai
Applied Analytics & Predictive Modeling
Due: 14 March 2024

Project Description Report 1

Data Breakdown:

The dataset used consisted of a main title column that contained Reddit data from submissions, or posts, in the opiates subreddit. The ‘submission’ file contains the main posts made in the subreddit and the ‘comments’ file consists of replies in threads to the submission posts. Other key features of the dataset include the author, score, and created_utc. The score column is a calculated column based on upvotes and downvotes for the post, with an upvote being plus one and down vote being minus one. The created_utc column represents the time of the post in Unix form. The remaining columns are attributes extracted from the text in the post or comment. These attributes come from the LIWC dictionary. They include features such as different linguistic dimensions, grammatical dimensions, psychological processes, time orientation, and more.

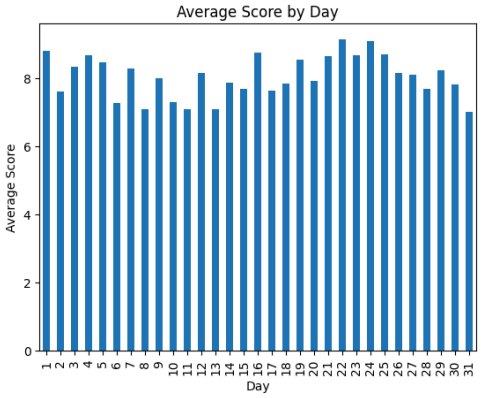
Preprocessing:

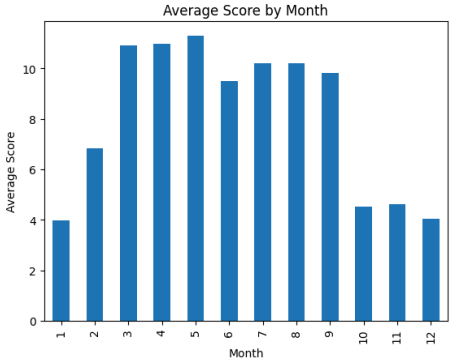
The first step in preprocessing the data was to observe all the features and their definitions to determine their usefulness for analysis. Upon inspection, the first edit that we decided to make was to convert the Unix time column to more interpretable data. We decided to convert this column into four new columns. These columns were hour, day, month, and year. Once this change was made we decided to create four additional columns which were one-hot encoded from the hour column. These columns were morning, afternoon, night, and overnight. The morning column was defined by the hours of 6-12, afternoon from 12-18, night from 18-24, and overnight from 0-6. This allowed us to observe changes in the values of the different features throughout the day. An example of this is shown in the second chart to the right. As a test, we wanted to see what the average score of a post would be during the four subsections of the day. In further analysis, we would like to analyze different sentiments, such as positive or negative

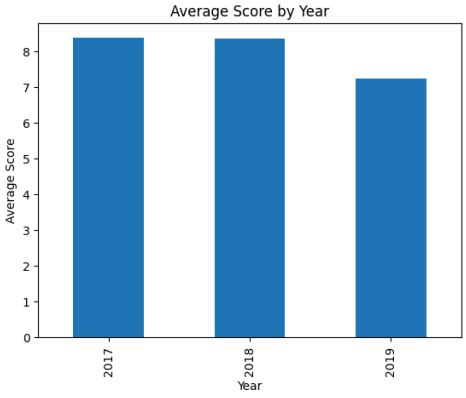
emotional indicators, at different points of the day. This would open the door to exploring how human emotion around this subject changes throughout the day.

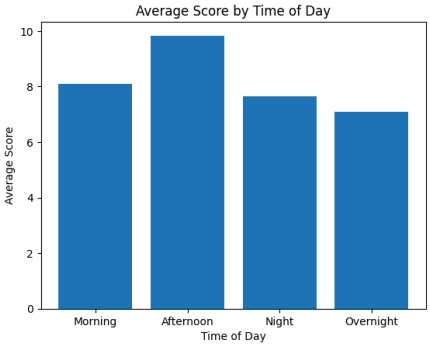
The remaining three charts show how the average score for a post changes with the different time metrics created in the previous preprocessing steps. Some patterns derived from these charts include average scores decreasing from 2017 to 2018 and decreasing more significantly from 2018 to 2019. They also show that the peak for the highest average score comes between March and May with the lowest average score coming in the colder months from October through January. Another interesting pattern is that throughout each month, the highest average scores are early in the month as well as just before the end of the month. This could possibly indicate people's increased inclination to seek help for an issue or make changes at the beginning of a month.

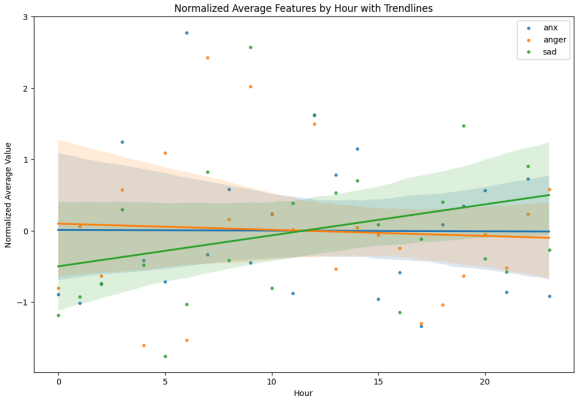
The time aspect of this data was the feature that intrigued us the most. Studying human behavior in this dataset over the course of time, whether it be over the course of a day, month, or year, can be very helpful in predicting the best way and time to help these individuals, who are mostly seeking help for their addictions.

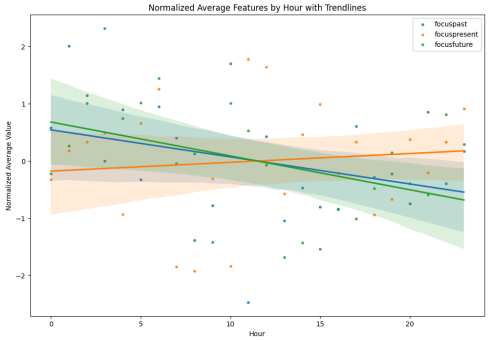












Task 1:

For the prediction task in this report, we will be analyzing the human emotional features over the course of each time measurement (hour, time of day, month, and year), in order to study how and when certain emotions are present during different times. This prediction model would be important because by predicting when individuals may be expressing certain emotions or seeking help for their issues, researchers can intervene or provide offers for help to those who need it more effectively.

We will build several models including a decision tree classifier, random forest classifier, and XGBoost to perform these predictions. The data to be used is still being determined but will include time orientation, psychological processes, and cognitive processes. Time orientation will allow us to gather data on whether the poster is being reminiscent or forward-thinking in their communication. Psychological processes will allow us to gather data on the overall sentiment that the poster is expressing. This includes emotional indicators such as sadness, anger, anxiety, etc. as well as overall positive and negative emotional indicators. Cognitive processes allow us to gather data on the six sub-scores of the metric, including insight, causation, discrepancy, tentativeness, certainty, and differentiation.

Task 2:

For the second task, we are building a classification model to understand who is writing the posts in regard to the topic as well as the gender split. To complete this we will use the pronoun, social-related, writing style, and word usage features. This model is key to understanding not only who is posting in this subreddit but also how opioids are affecting both genders. This will allow us to see who the most vulnerable groups are to these drugs and how to help them.

Model Process:

Preprocessing and Setup:

- To complete this task we will start by extracting the needed features from the data set.
- Then we will normalize the data to ensure that each feature has equal weighting in the model.

Model Building:

- Using Python and the Scikit Learn library we will use a Random Forest Classification model as the core.
    - We chose to use Random Forest because it uses multiple trees instead of just a single tree. This amount can be chosen by us and allows for more accurate predictions. As a result of using more trees, we can reduce overfitting as well as ensure better performance which is key with such a large data set.

Model Testing:

- Once the model has been built and we have created a multitude of train and test data sets, we will run rigorous testing on the model to make it as accurate as possible.
- Using the Sklearn accuracy_score, confusion_matrix, and classification_report, we will assess the scores and modify the model accordingly.

After the model has been refined to the best of our abilities, we will run our final test with all of the data to get the analysis that we initially wanted. This will be the basis for our understanding of the authorship and gender impact.