Stress Analysis on Reddit

R for Data Science Project 1

Melanie McCord

Background

- Stress is a common aspect of modern lives and can result in negative health outcomes.
- Being able to predict stress on social media can be beneficial.
- Reddit is a social media platform where users post questions and can get advice.
- Mining from stress related subreddits can be beneficial to understanding the problem and being able to predict it.

Research Questions

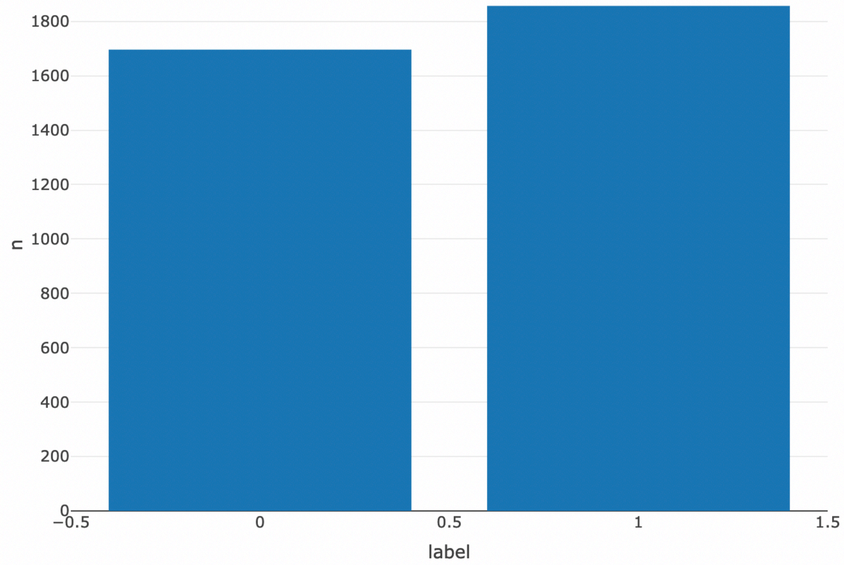
- How does stress levels differ among subreddits?
- How can we predict stress level given words in the data and lexical features?
- Is there an association between stress-related data and subreddits?

Dataset Overview

- Dataset consists of training data with 2838 posts, test data with 715 posts
- Sampled from 11 subreddits

Getting the Labels Distribution

```
label_counts <- reddit_stress_data %>%
  group_by(label) %>%
  count()
plot_ly(label_counts, x = ~label, y = ~n, type = "bar")
```



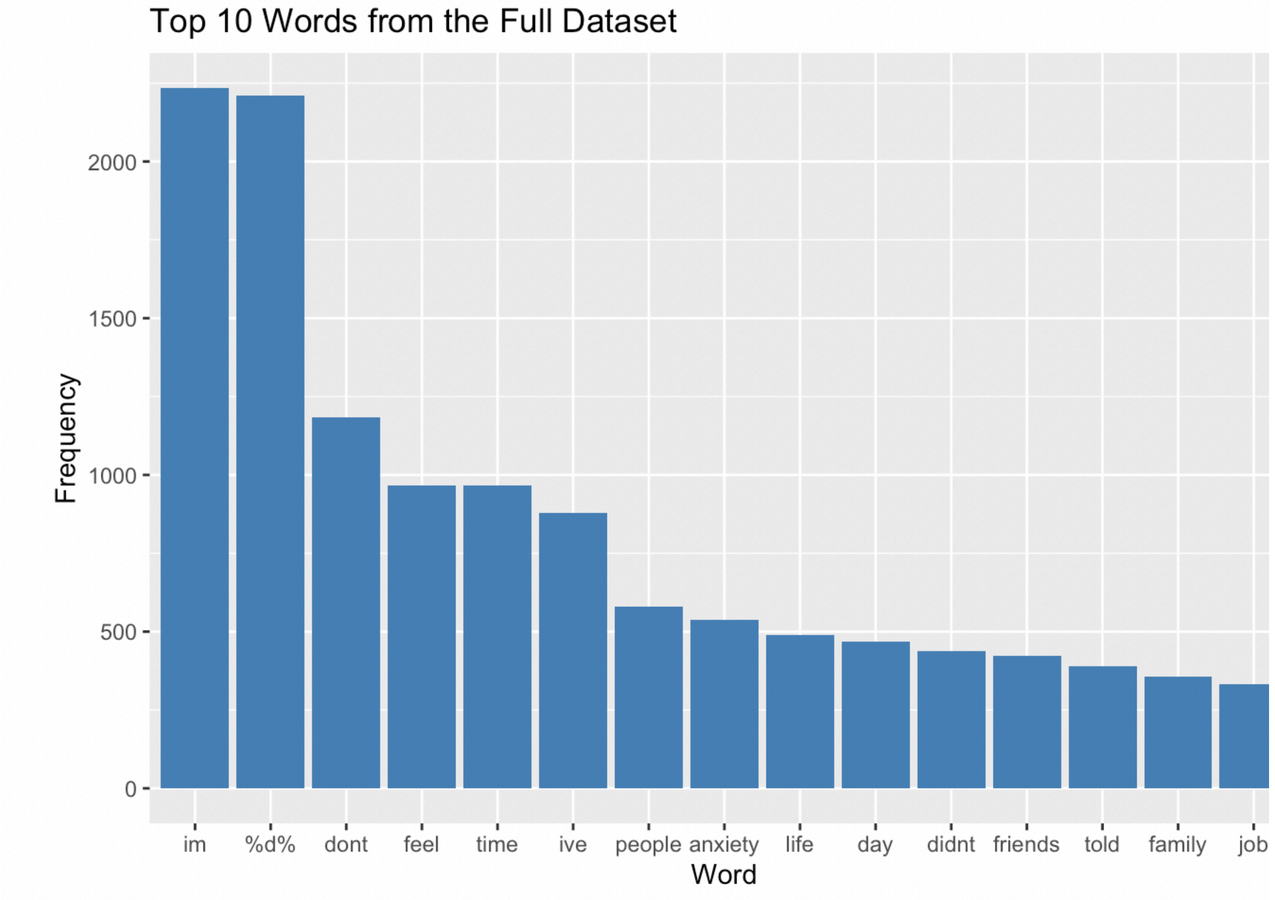
Special Issues with Text-Related Data

- Text data is inherently messy.
- Firstly, computers can only process numbers, meaning that any text data needs some way to be converted to numbers.
- Some words that commonly appear in text have very little influence on the text. We call these **stop words**. (Such as “and”, “so”, “to”, etc.)
- A computer treats “cold”, “Cold”, “cold;” as separate objects, even though they are the same word.
- Additionally, there are rare words that may not appear very often but may have an influence on the text data.

Data Wrangling Completed

- Converting all words to lowercase
- Removing punctuation
- Counting the number of times a word appears and adding it as a feature to each post
- Removing rare words
- Adding an indicator to clarify the difference between “subreddit” as the feature and “subreddit” in the post
- Joining the training and test data
- Adding the word columns to the original data

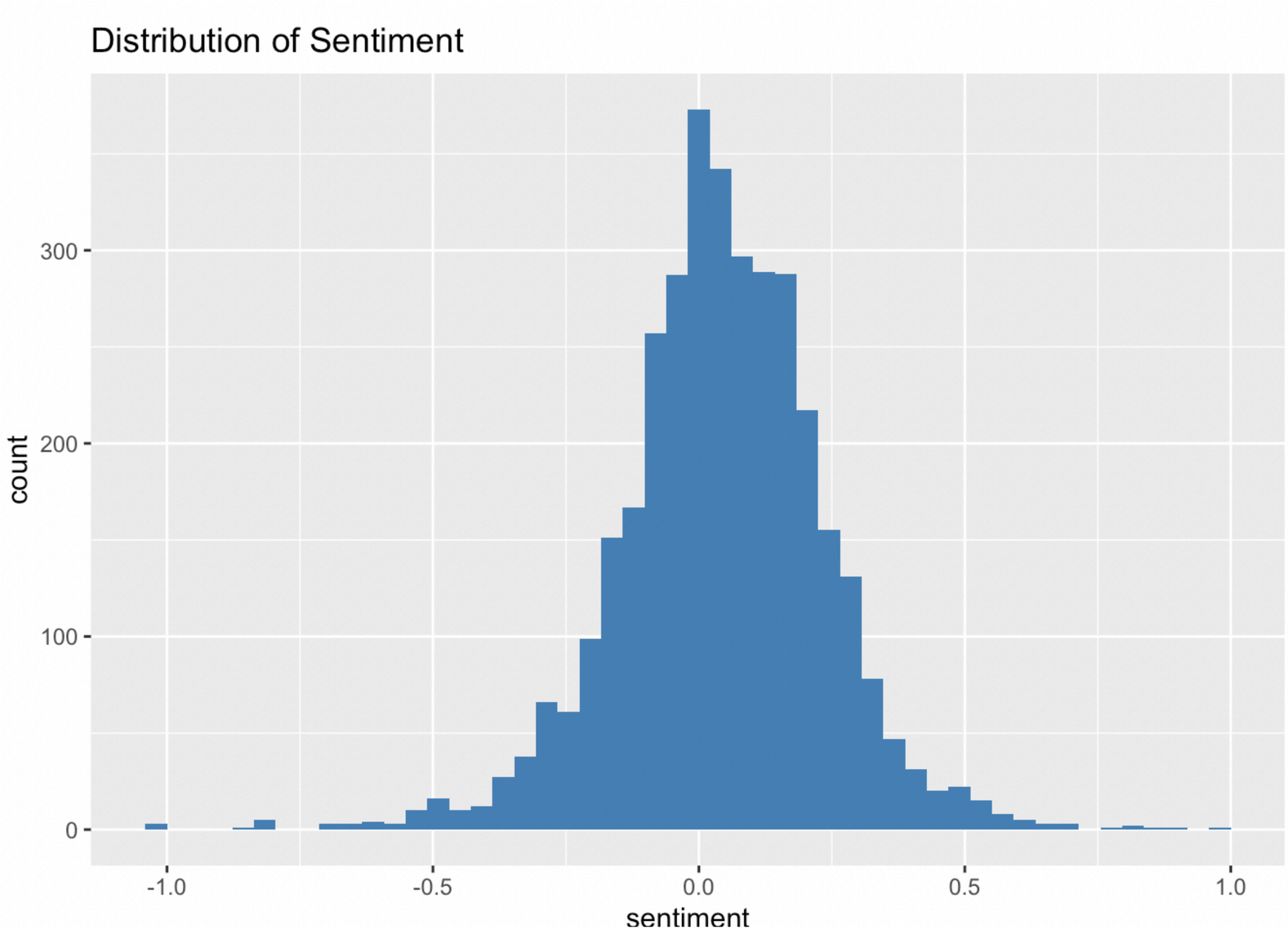
Top Words By Category



Results

- Completed:
  - Visualizations of unknown words
  - Visualizations of stress distribution and subreddit distribution
- To be done:
  - Decision tree model of the words
  - Chi-squared test of association between subreddit and label

Distribution of Sentiment



Work to be done

- RFE feature selection of the data
- Chi-squared test of the differences of label by subreddit and sentiment
- Decision tree model in order to predict label