R for Data Science Project 1: Stress Analysis On Reddit

Melanie McCord

Contents

Introduction

Background

Stress is a common aspect of modern lives that can cause negative health outcomes, such as anxiety, depression, insomnia, and a weakened immune system. If we can predict stress on social media, we can help understand the extent of the problem and perhaps gain insights on how to address it. For this specific project, the focus is on Reddit. Reddit is a social media platform where users post questions and can get advice. Importantly, Reddit data is concentrated on specific subreddits, divided by topic. This makes it easier to filter by specific topic and analyze the data from each specific topic. If we mine from stress-related subreddits, we can get a more detailed overview of the problem.

Data Structure and Source

This dataset was collected by mining all posts from 10 subreddits between January 1, 2017 and November 19, 2018. Then, the data was annotated by human annotators as being either "stressed", "non-stressed", or "unclear". The posts that were unclear as to whether or not they were stressed were discarded. For each post, the features are the text, the label, and some syntactic features, lexical features, and social media features, as well as the sentiment (how overly positive or negative the posts are).

For more information about the dataset, see this paper: Dreaddit: A Reddit Dataset for Stress Analysis On Social Media.

Major Variables I Will Be Exploring

- Label: stressed (1) or non-stressed (0)
- Sentiment: the intensity of positive or negative a particular post is (-1 being completely negative, 0 being completely neutral, and 1 being completely negative)
- Subreddit: the source of the particular post
- The words that appear in each post (from the text)

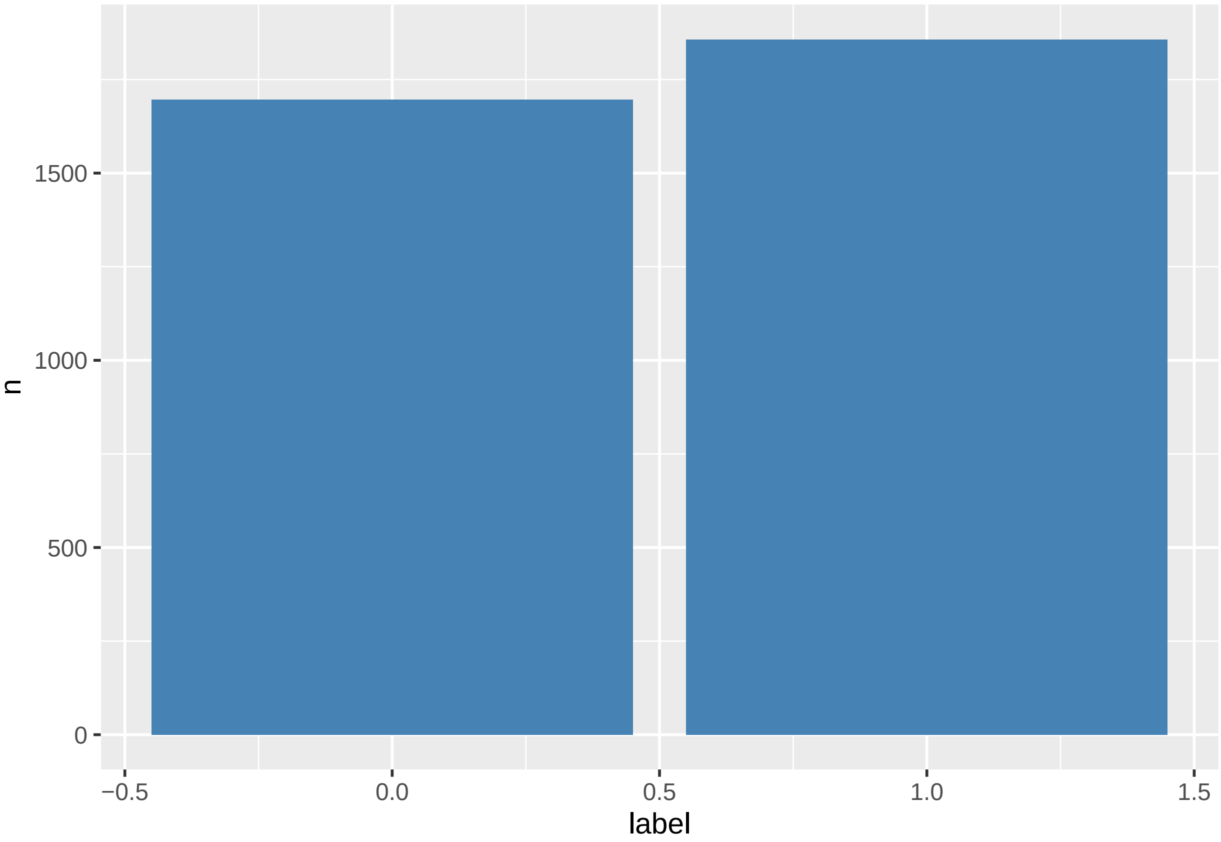
Future work may include exploring the other variables, which include social media features, lexical features, and syntactic features.

Research Questions

The topics I am exploring in this paper are: How does the stress label differ among subreddits? How can we predict stress given words in the data and sentiment? Is there an association between stress-related data and subreddits?

Distribution of Major Features Among the Dataset

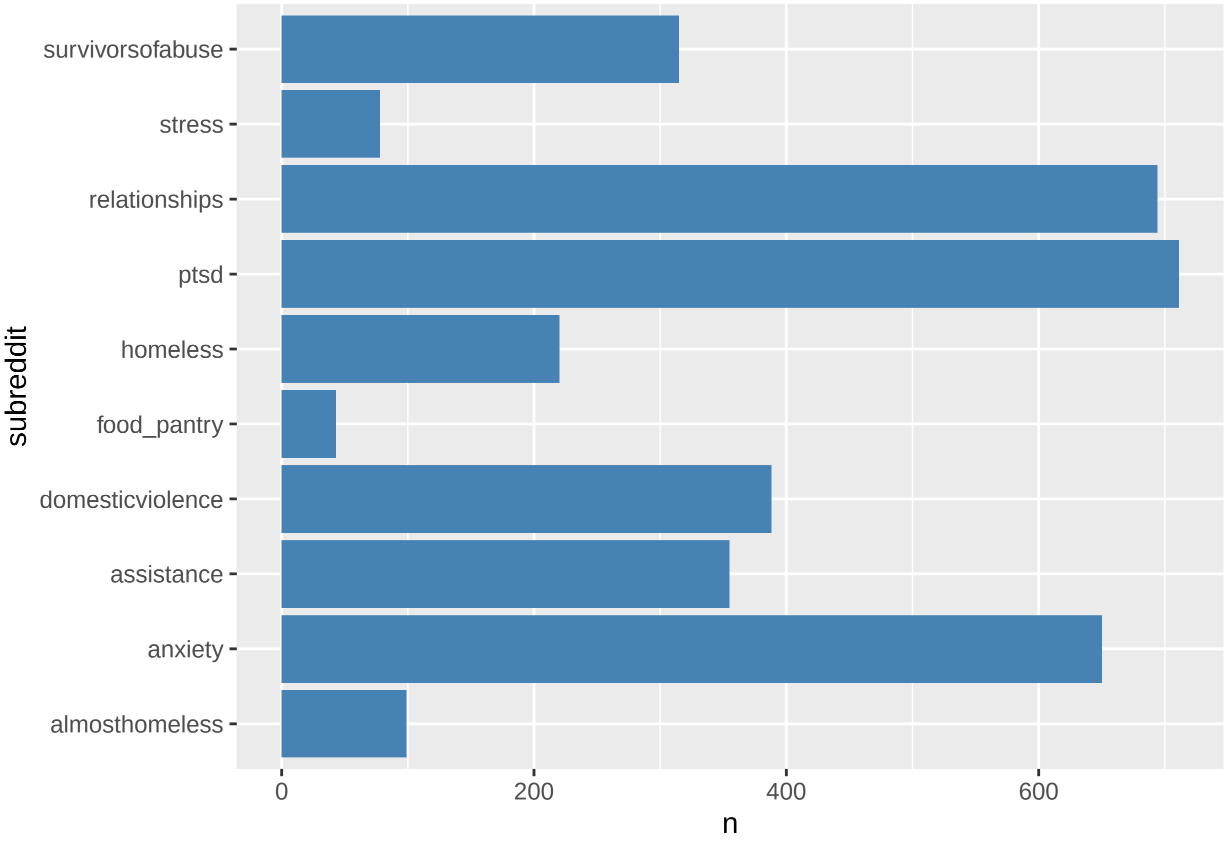
Dataset Label Distribution



```
##
##         0         1
## 0.4773431 0.5226569
```

The data is slightly biased in favor of stress-related posts, but not overly so (52% stressed versus 48% non-stressed).

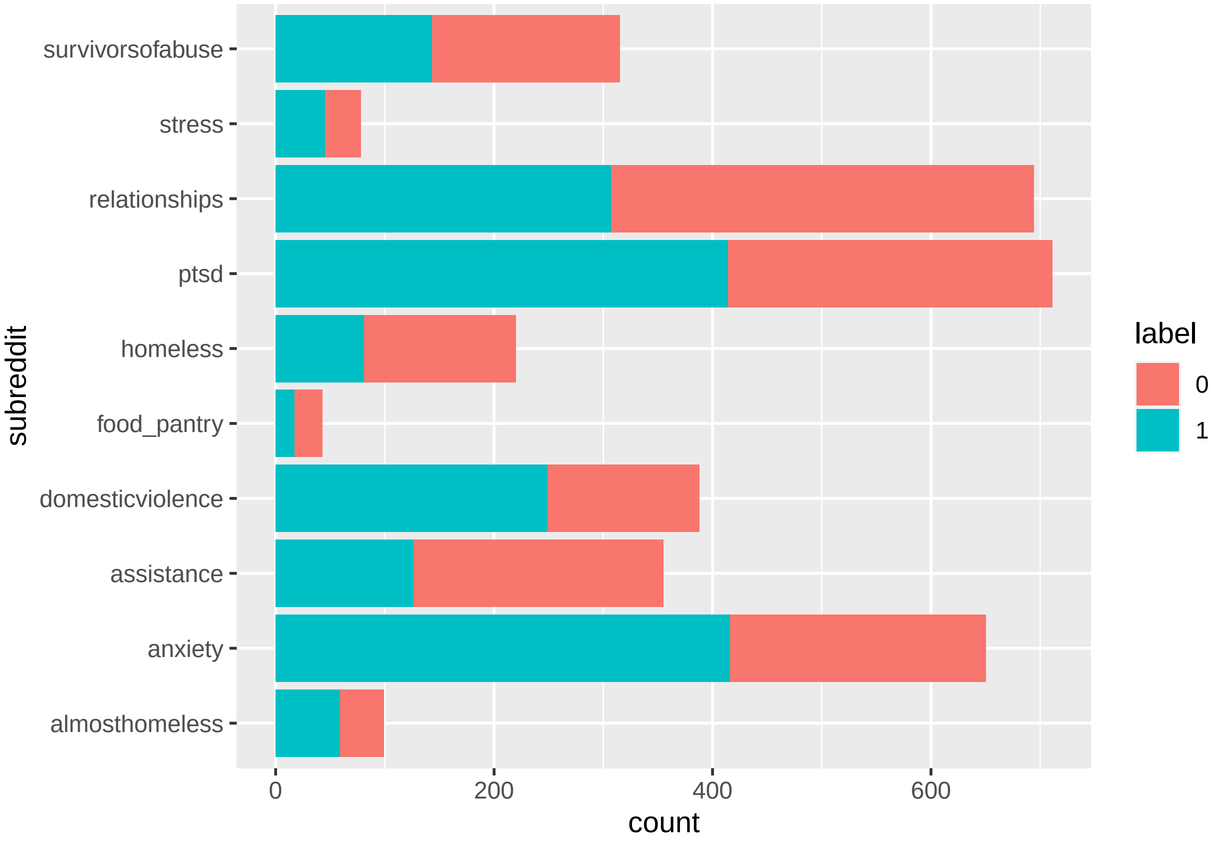
Distribution of Data By Subreddit



```
##
##   almosthomeless          anxiety       assistance domesticviolence
##       0.02786378       0.18294399       0.09991556       0.10920349
##      food_pantry         homeless             ptsd    relationships
##       0.01210245       0.06191950       0.20011258       0.19532789
##           stress survivorsofabuse
##       0.02195328       0.08865747
```

The dataset distribution is heavily imbalanced. 20% of the posts sampled are from r/ptsd, whereas only 1% of the posts sampled are from r/food_pantry. This may be because certain subreddits are less active than other subreddits, or certain subreddits are harder to label as being stressed or not stressed. Regardless, since the data is heavily biased toward certain subreddits, any classification model we will fit to this dataset is likely going to be able to predict stress labels from r/relationships significantly better than labels from r/food_pantry.

Distribution of Posts By Label and Subreddit



Additionally, for each particular subreddit, there are uneven distributions in the percentage of posts that are stress or not stress related. Assistance appears to be imbalanced towards non-stress related posts, whereas domesticviolence is biased towards stress-related posts. One possible explanation could be that although people are seeking advice from these subreddits, assistance may simply be asking what resources are available, whereas domesticviolence or anxiety may be more focused on stressful incidents or needing support.

Methodology

Data Preprocessing

Another detail about the statistical analysis is that since I was primarily focusing on text mining, there were a number of things that I had to consider that are unique to text 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. There are a lot of different ways to do this, but the model I chose to focus on is a bag of words model, which counts the number of times a particular word appears and adds each word as a feature. However, in doing so there were some things I needed to clean. Stopwords, such as “and”, “so”, and “to”, appear frequently but don’t matter much when understanding the meaning of text. Additionally, punctuation, and capitalization causes issues. A computer considers “Don’t”, “don’t,” and “DONT” to be separate words. As part of my data wrangling, I needed to deal with this. Also, some words only appear in one post. For example, one post may ask about resources available in Louisiana, but no other post mentions Louisiana. We need a way to deal with these issues. It’s possible that rare words may indicate some information about the text, so we can’t just remove rare words. Additionally, the rare words cause the number of words that appear in the post to be significantly higher.

```
## [1] 12059
```

Without removing any of the rare words, only removing the stopwords, there are 12059 words in the text.

Let's look at the distribution of the count of these words.

```
##  min Q1 median Q3  max     mean       sd     n missing
##    1  1      2  5 2235 8.049092 41.04796 12059       0
```

```
## [1] 3553
```

Clearly, there is a large disparity among the most common words and the least common words. The median distribution is 2, and the third quartile is 5 words. There are 3553 posts, and 5 words is still quite rare. If we adjust to remove the number of words, we will be significantly benefitted from this.

Let's set the definition of rare words to be 15 and compare the result.

```
## [1] 1723
```

Since we have eliminated a large percentage of words, let's look at the overall distribution of the rare words and compare it to the words that did not get removed.

```
##  min Q1 median Q3 max     mean       sd     n missing
##    1  1      1  3  14 2.738706 2.791588 10869       0
```

```
##  min Q1 median Q3  max    mean       sd    n missing
##   15 20     30 51 2235 56.5521 120.0162 1190       0
```

Now, the min is significantly larger. There are significantly more rare words than common words. But among the non-rare words, there is significantly more variation and you can see the overall spread of words better.

Another related issue I ran into was dealing with the scenario where "id" and "subreddit" both appeared in the dataset of words and in the original dataset. I dealt with this by adding "text_" to each word after the removal of stopwords and punctuation.

I also joined the training and test data in order to reduce potential bias between the data selected as training and the data selected as test data and added the bag of words column to the original data.

Data Exploration

I was interested in exploring the differences between common words among each of the subreddits, and the differences between the most common words by label and by subreddit. A future data exploration may include exploring the differences among the most common words by both label and subreddit, but since this visualization was difficult to read, I chose not to include it.

Additionally, I chose to explore the distribution of sentiment by subreddit, label, and both. This can shed a light on important patterns between the data and the subreddit column.

Statistical Modeling and Analysis

For the statistical analysis portion, I am doing a chi-square test of statistical significance of the differences among label and subreddit, and an analysis of variance between subreddit and sentiment and label and sentiment.

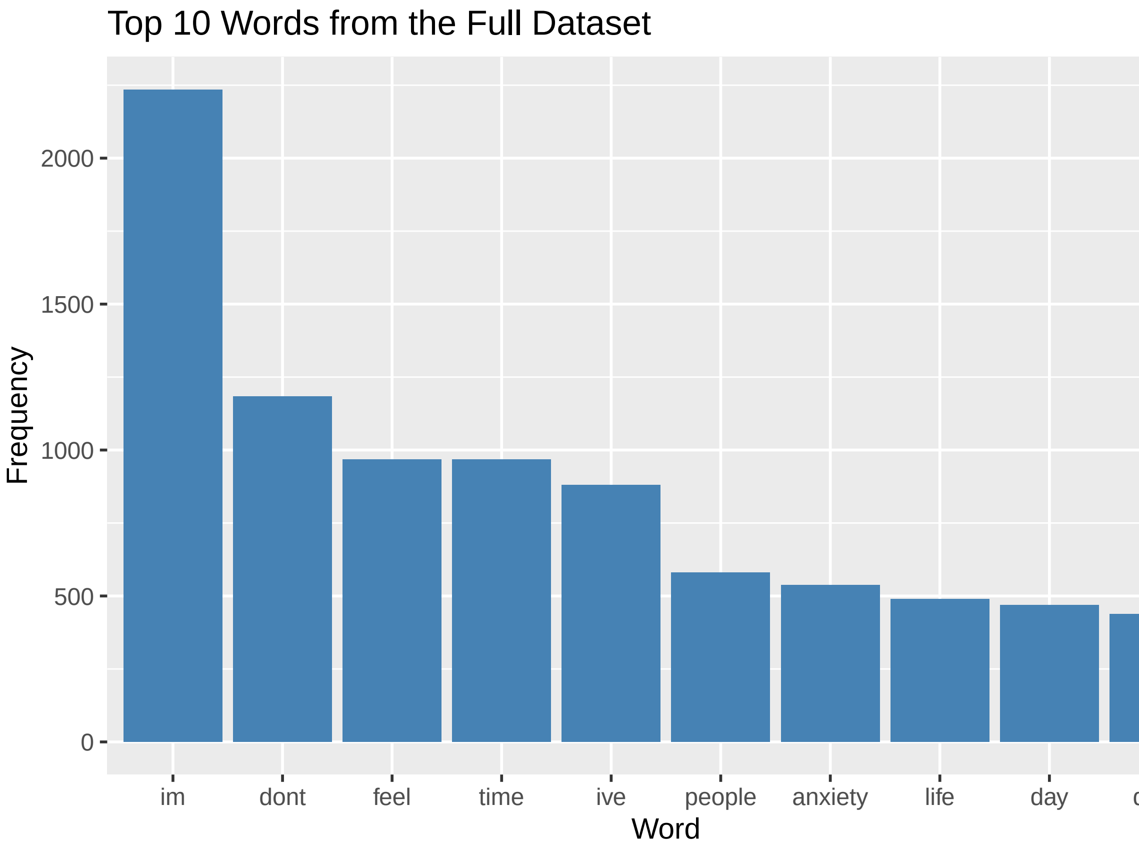
I will also run a decision tree model on the dataset in order to see what the strongest predictors of label are from the words and sentiment and compare its performance on stressed versus non-stressed data.

Results

Data Exploration

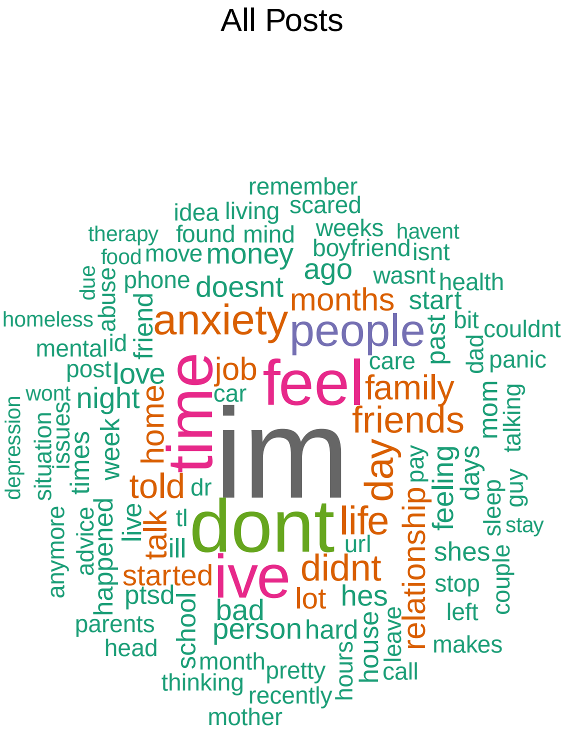
Top 10 Words

Since I'm interested in exploring the Reddit Stress dataset by counting the words that appear, my first step is to do some visualizations of the top 10 most common words.



Top 15 Words Overall

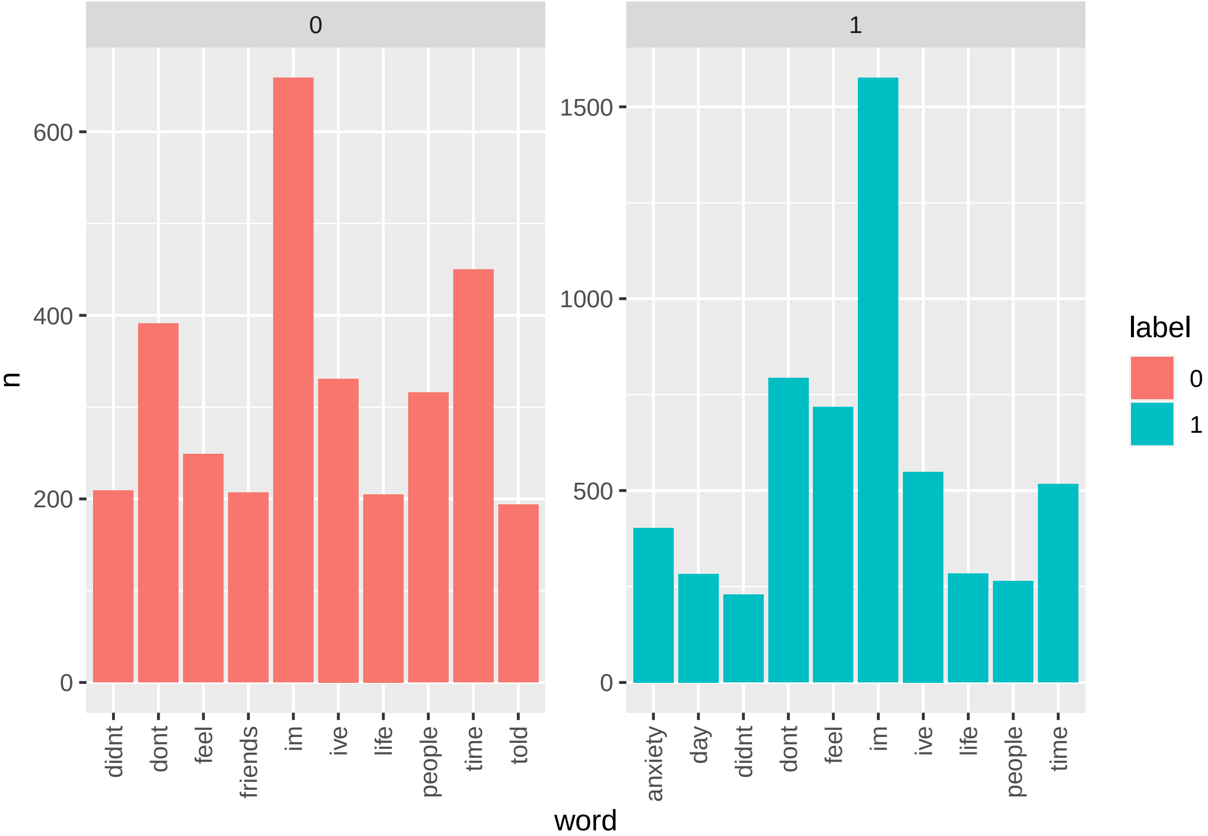
For the whole dataset, the most common words appear to be neutral, possibly linked to the diversity of posts reflected. Anxiety feel and people appear frequently, which is understandable considering the focus is stress-related subreddits. Another way to visualize this is to use a wordcloud, which shows the words in the dataset, with larger words representing more common words in the dataset.



In this visualization, we can get a better idea of the diversity of topics covered in this dataset. Feel, anxiety, friends, relationship, and family all appear very frequently, and the words that appear less frequently can be attributed to the diversity among topics, sych as boyfriend, therapy, homeless, school, etc.

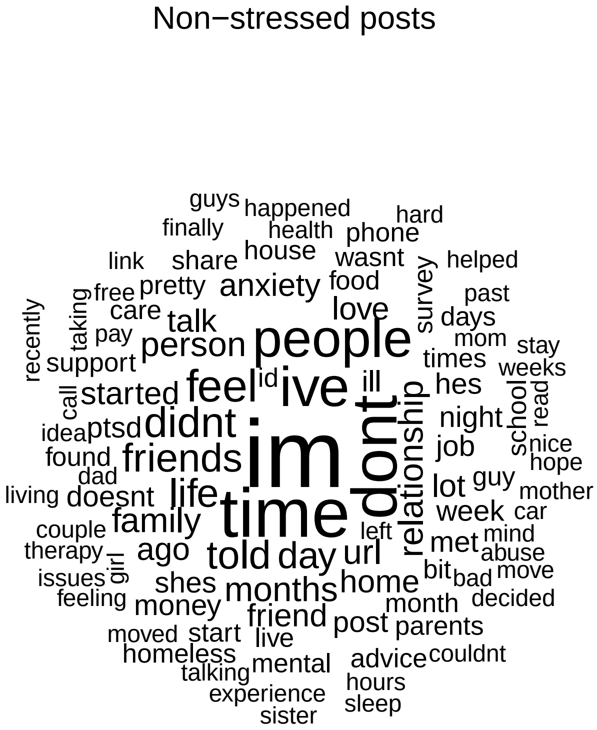
Top 10 Words By Label

```
## Joining, by = "word"
```

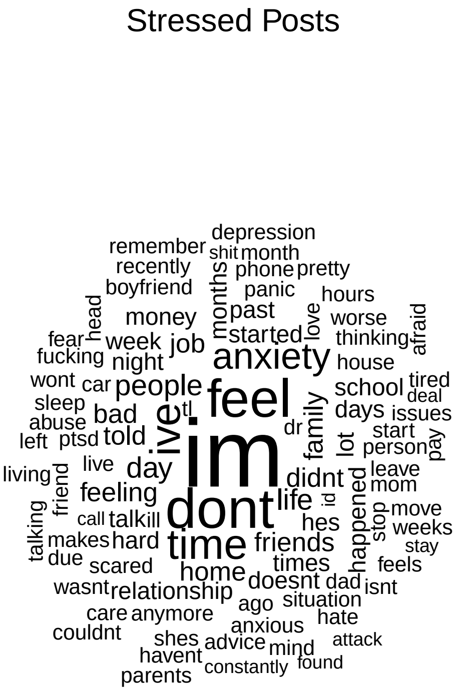


The most common words by label are significantly different. Stress related posts are most likely to talk about anxiety and day. Non-stress related posts are less likely to talk about feel and more likely to talk about friends and people.

Now, let's look at a word cloud of the words by label for a better look at some of the most common words among each category.



From this visualization, you can see that stress free posts have to do with a diversity of topics. There are a significant amount of negative words that appear fairly frequently, such as “homeless”, “ptsd” and anxiety, but the most common words are “I’ve” “don’t”, “time”, and “people”. Since the non-stress-related posts are still picked from subreddits related to stressful topics, this may be people simply talking about how they dealt with these issues or calmly asking for advice, rather than being obviously stressed while writing their posts.

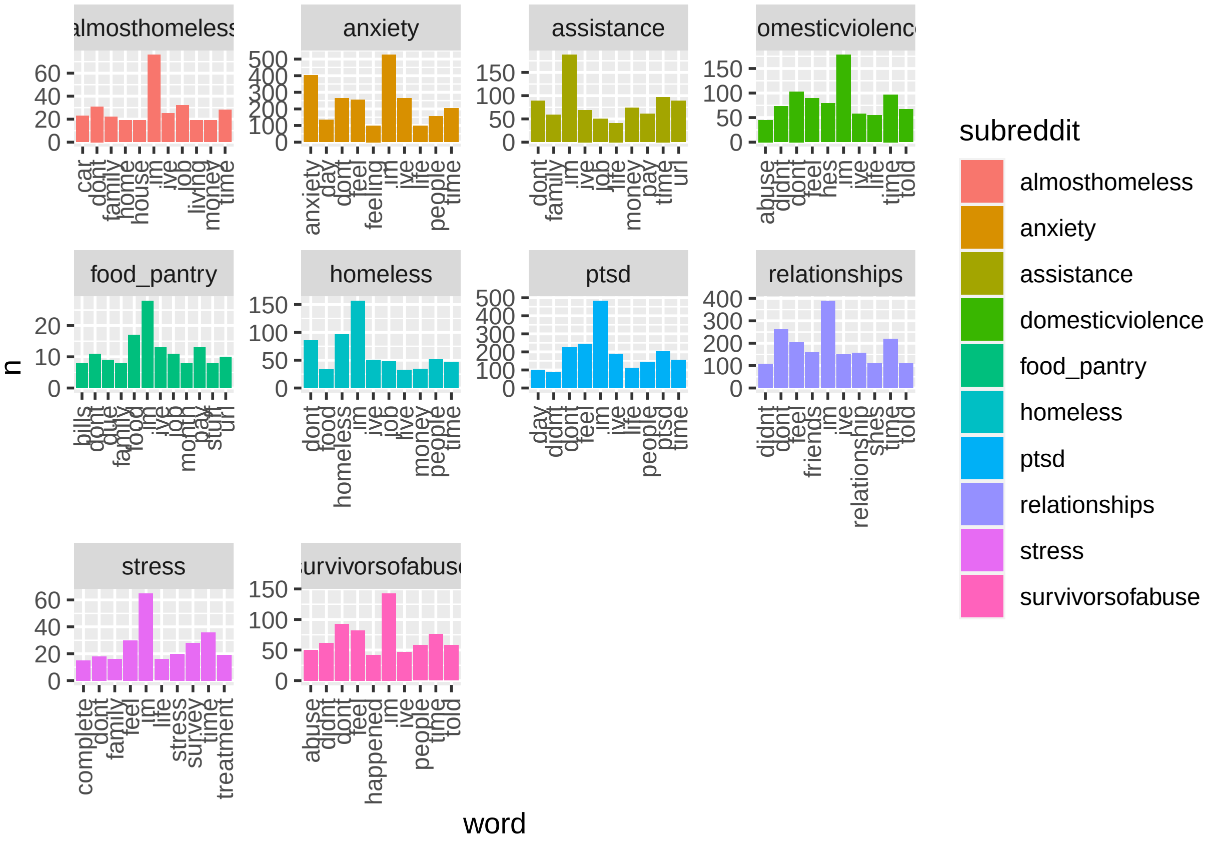


Among the stress-related posts, there is significantly more variation in the words and the only words that appear very frequently are “i’m”, “don’t”, “feel”, “time” and anxiety. There appear to be a lot of posts that seem related to the diversity of posts. Some words that commonly appear are “fucking”, “abuse”, “panic”, and “depression”. Time appears frequently in both the stress-related and non-stress-related subreddits, but proportionately much more frequently in the non-stress-related.

Overall, the stressed posts' word distributions and the non-stressed posts' word distributions appear similar, but the stressed posts appear to be more diverse, and the non-stressed posts appear to be more similar and have less very negative words among the top 100 words.

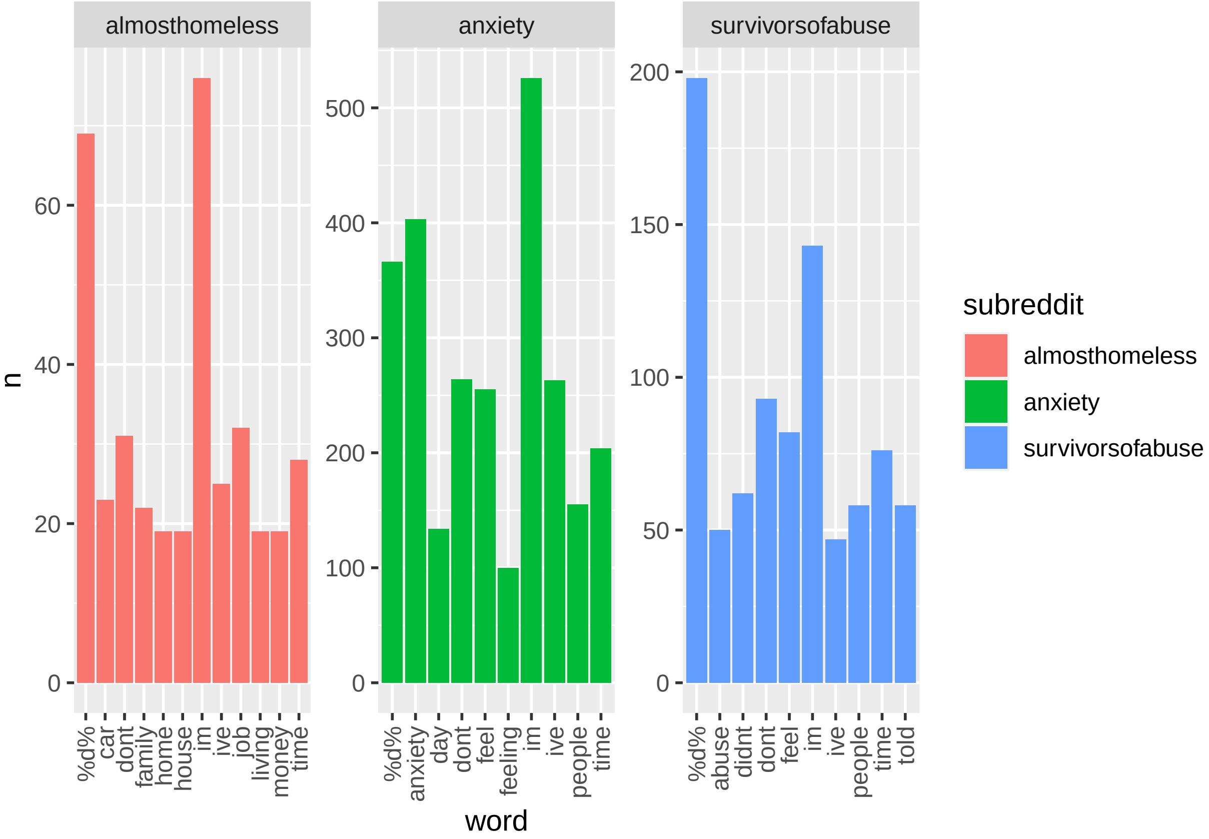
Top 10 Words By Subreddit

```
## Joining, by = "word"
```

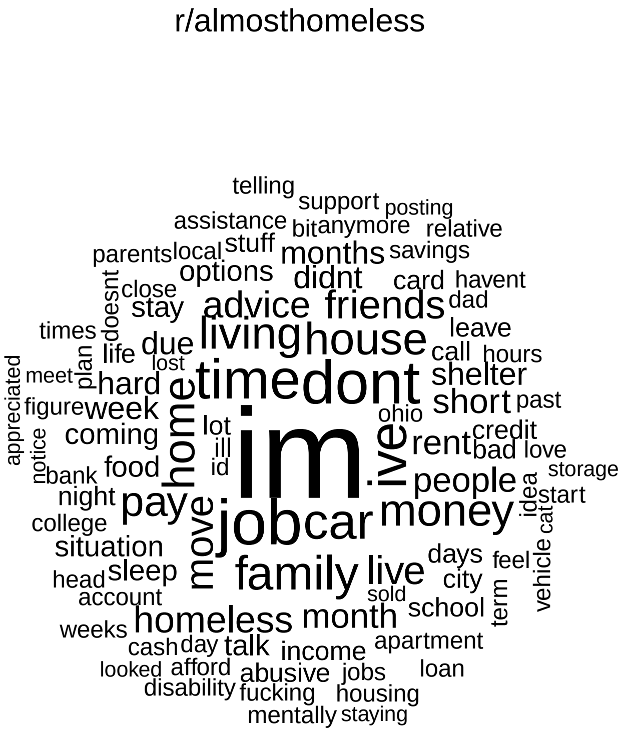


From this visualization, you can see that the most common words differ significantly by subreddit and seem to be related to the major topics of each subreddit. However, from this visualization, it is a little bit difficult to specifically see the differences among each of the 10 subreddits. For a more detailed overview of the differences, I'm going to look at specifically 3 of the subreddits: r/almosthomeless, r/anxiety, and r/survivorsofabuse.

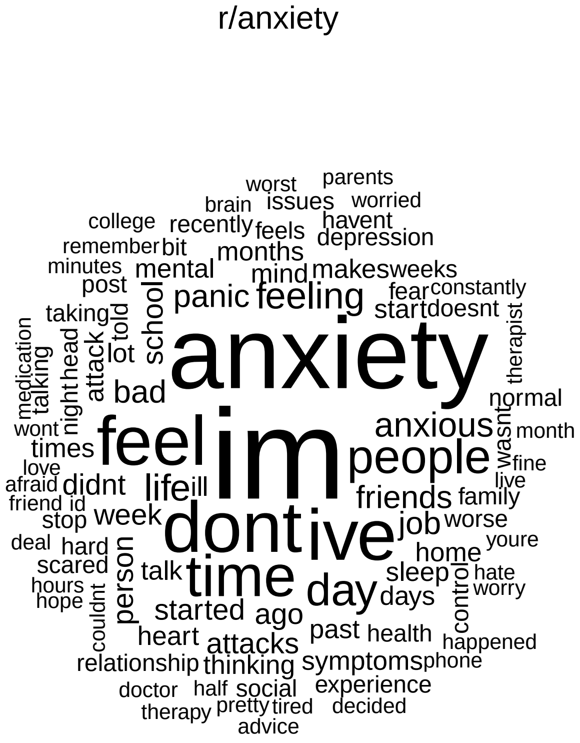
```
## Joining, by = "word"
```



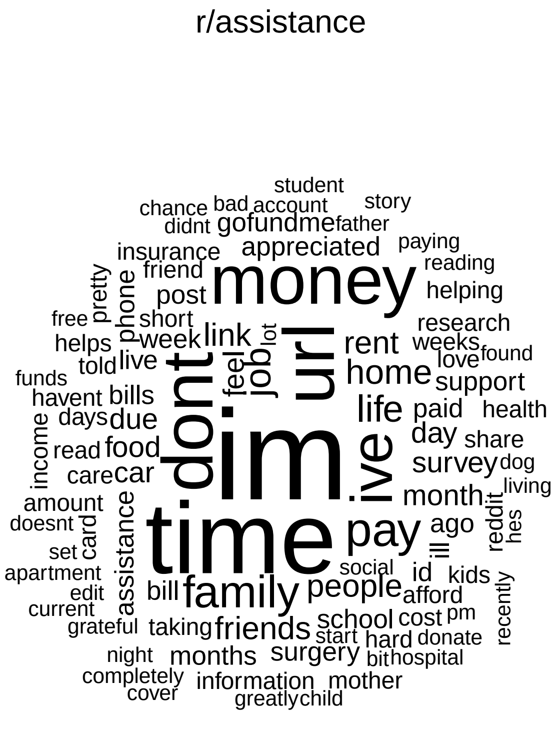
For the word clouds, I will not look at every subreddit's word cloud, but here are the first 5 and a detailed comparison of each.



The posts in r/almosthomeless are clearly very stress-related. Although many of the posts are related to the expected “home”, “car”, “job”, and “house”, there are also significant amounts of very negative words, such as “abusive”, “suicide”, and “mentally”.

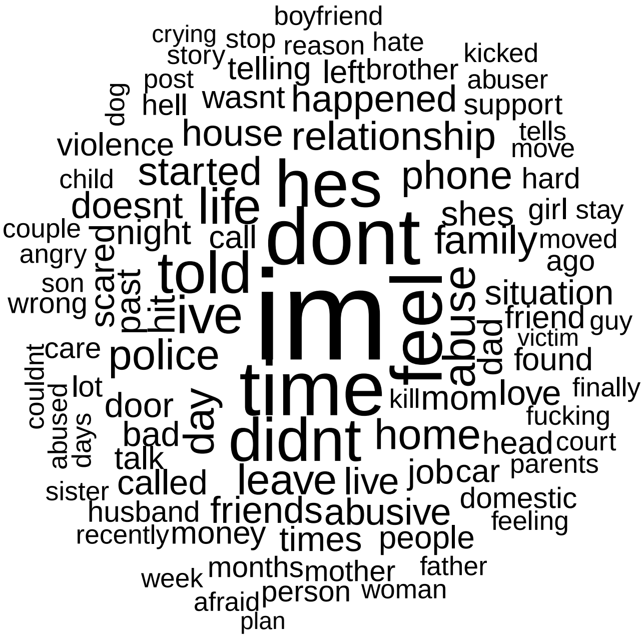

r/anxiety

Anxiety posts appear to be concentrated very closely around anxiety-related issues, such as “feel”, “anxiety”, “panic”, and “anxious”. Some of the top words appear to be related to common triggers for anxiety, such as “relationship”, “college”, “social”, but overall, the majority of the common words are simply anxiety and feeling related.

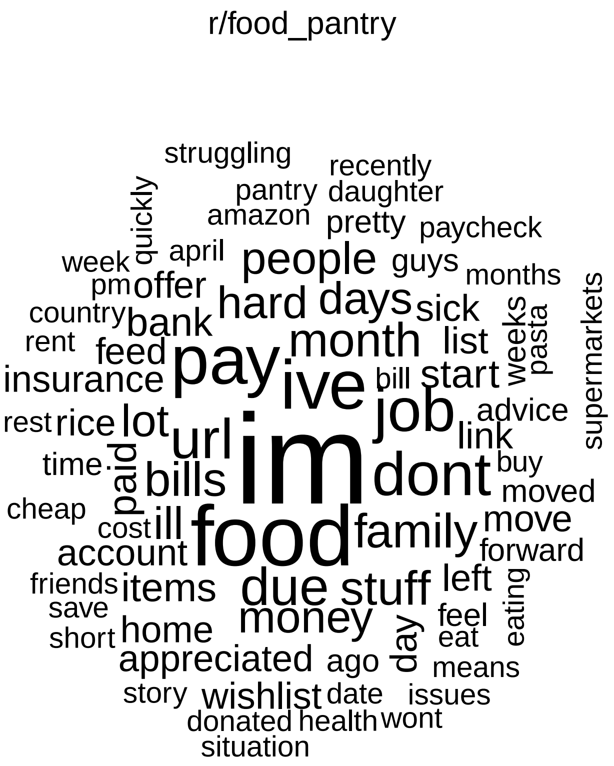

r/assistance

The assistance posts are notably much different from the other two categories discussed above. Many of the assistance posts are money-related. Common words include possible reasons people may need assistance, such as “kids”, “surgery”, and “college”, but most of the words focus on financial-related things. GoFundMe appears somewhat frequently, which may indicate people setting up GoFundMes for people who need assistance or requesting aid for themselves.

r/domesticviolence



The posts from r/domesticviolence seem to be highly focused on abuse and domestic violence related issues and very stressed. The most common words include words like “feel”, “abusive”, “police”, and “situation”. From this visualization, it appears that most posts from r/domesticviolence are scared, tense, and detailing very bad situations.



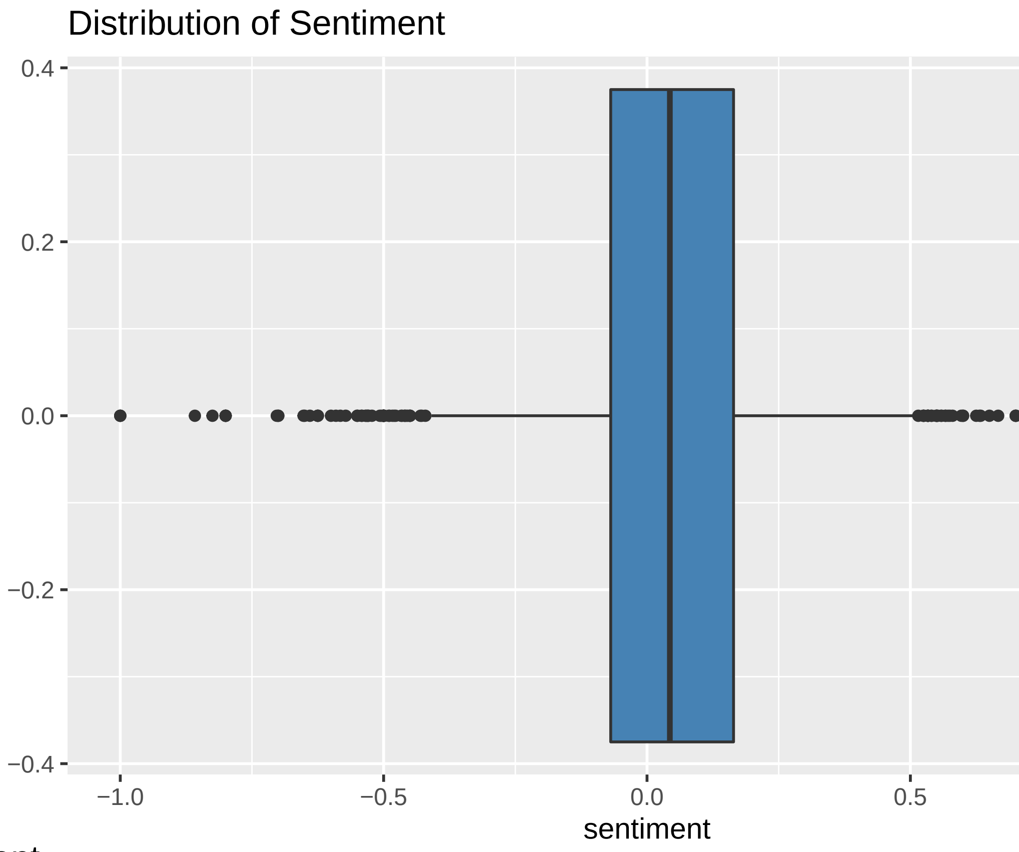
Posts from r/foodpantry are primarily focused on financial- and food-related words. However, overall the words appear to be considerably less clustered around highly negative and tense words and more clustered around things like needing support, for example “supermarkets”, “wishlist”, “bills”, and “rice.”

Overall, from the posts, you can see the patterns among stress-related and non-stress related posts. The stressful topics are reflected among the top most common words among the posts. The words differ by subreddit, and you can see the patterns among them.

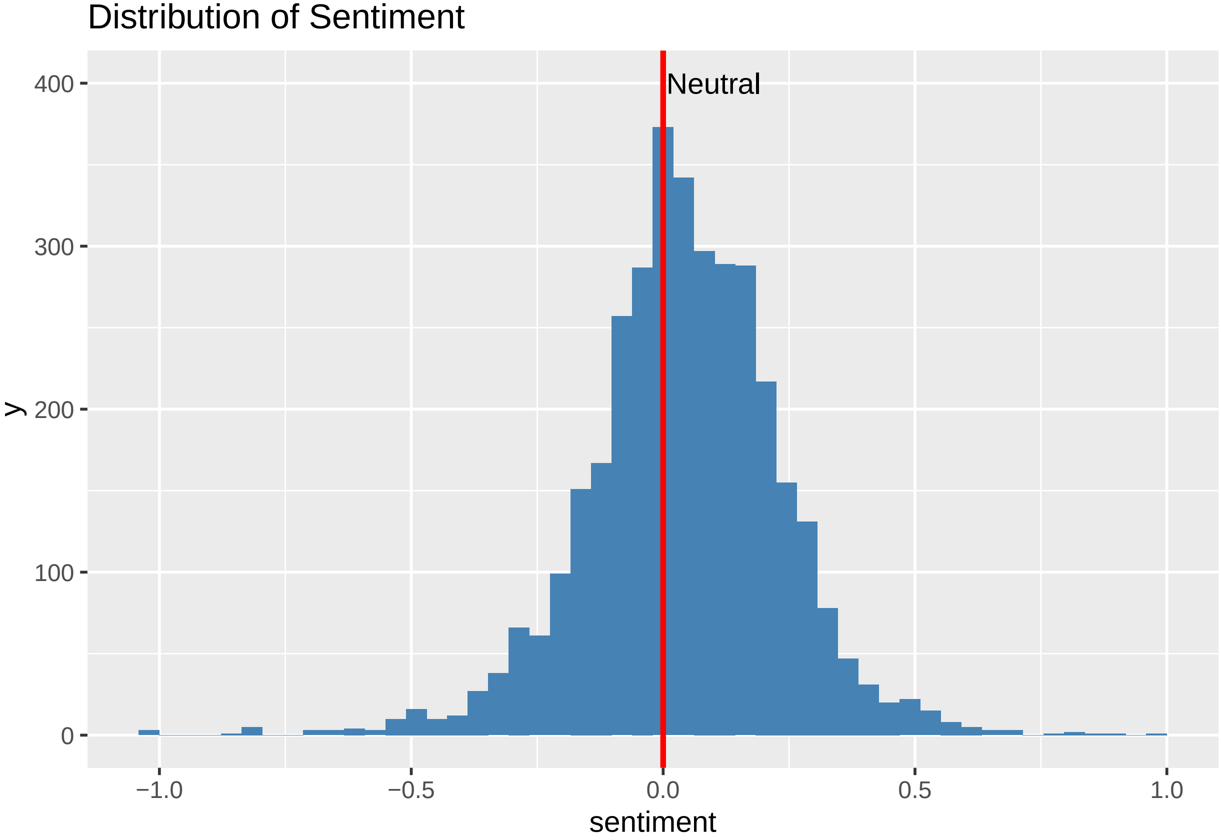
Sentiment Distribution

When we use sentiment to describe posts, we mean a score to determine how negative/neutral or positive some text is. 1 means completely positive, 0 means completely neutral, and -1 means completely negative. This score is calculated using a metric that takes into account what words were used, whether they were positive or negative, and what the ratio is of positive versus negative words.

In order to understand how the sentiment differs by label and subreddit, I am going to look at the sentiment distribution of the overall dataset, the differences by label, and the differences by subreddit.

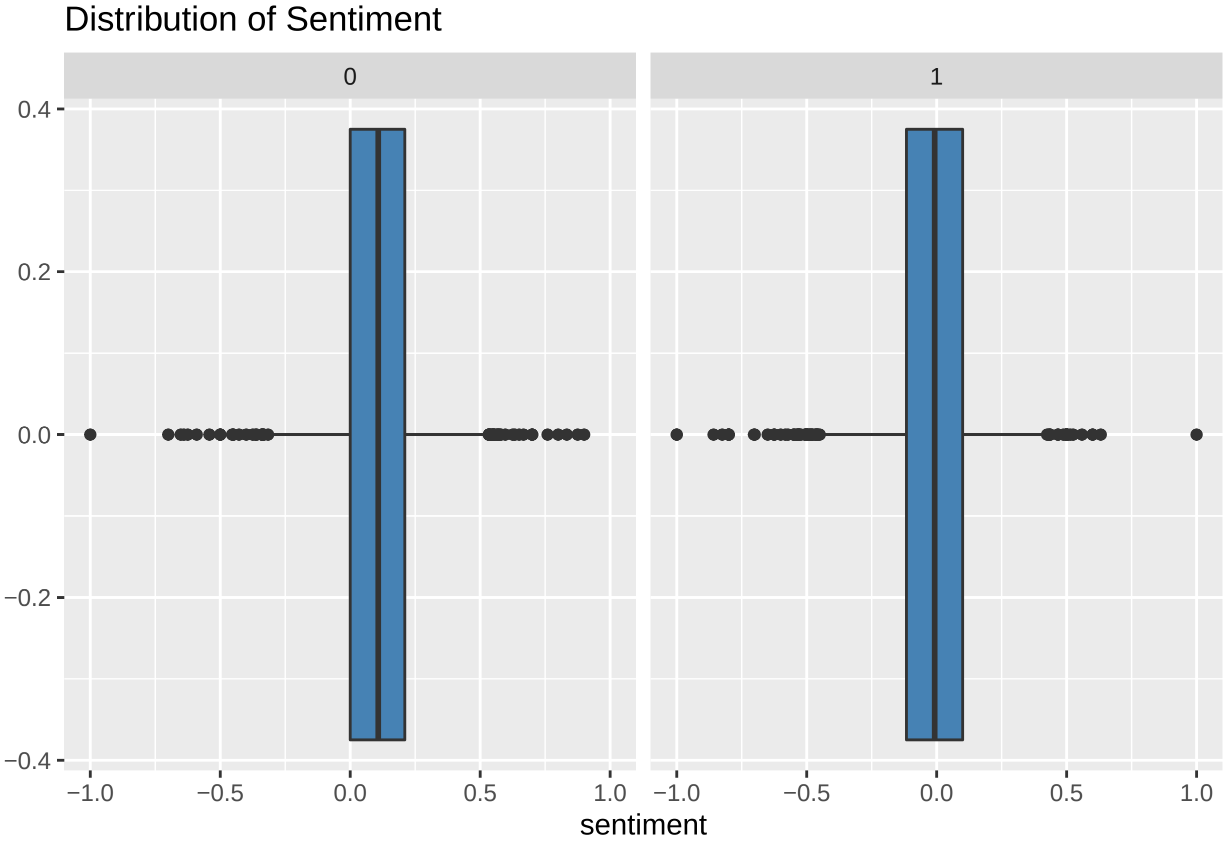


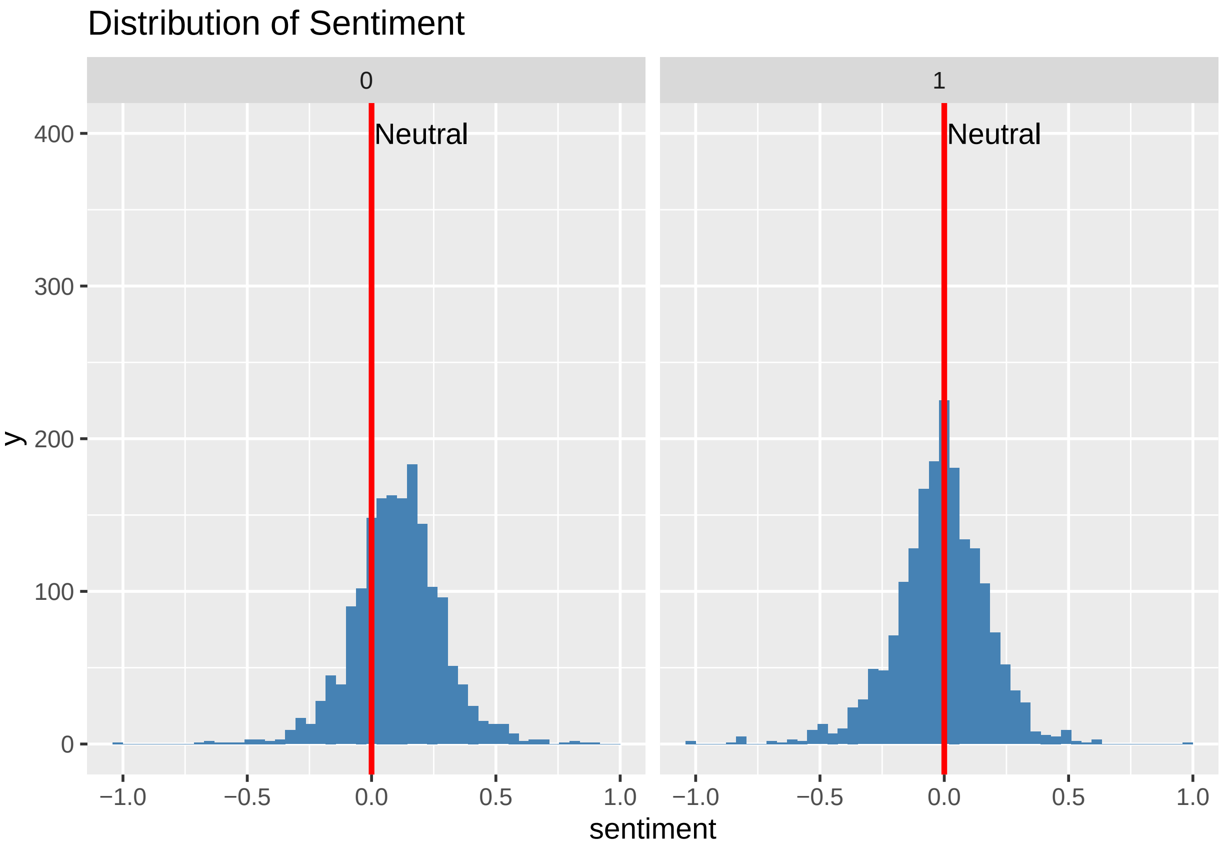
Sentiment Distribution Overall



The overall sentiment distribution of the posts is neutral, leaning slightly positive, but there are many outliers. We can see that the data is approximately normally distributed, but with more negative than positive outliers. #### Sentiment Distribution By Label

```
## # A tibble: 2 x 2
##   label `mean(sentiment)`
##   <dbl>             <dbl>
## 1     0            0.105
## 2     1           -0.0157
```



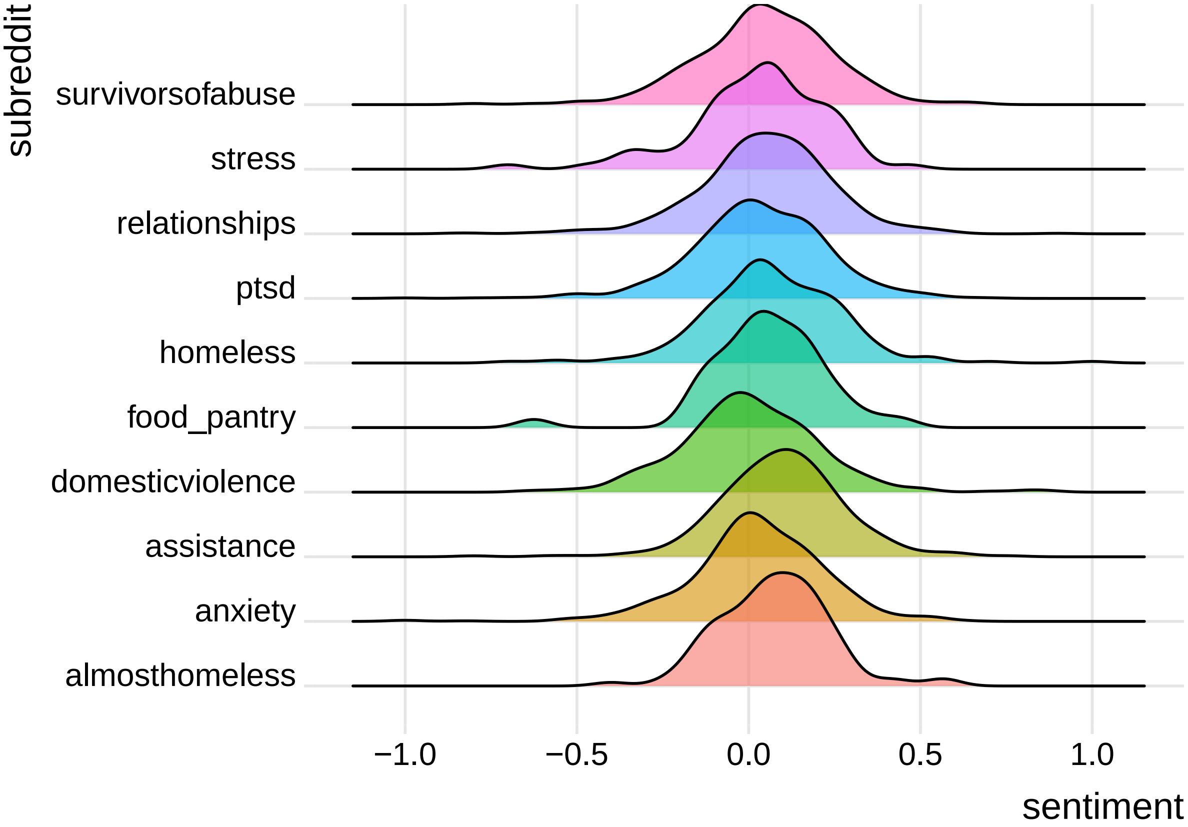


The stressed data is centered close to 0, meaning that the overall distribution of posts is neutral, leaning slightly negative. However, there are a lot of outliers on either end, though there are more negative outliers than positive ones. The non-stressed data is centered slightly above 0, with a significant amount of both positive and negative outliers. Overall, the distributions are similar, but stressed data is significantly more negative, as to be expected.

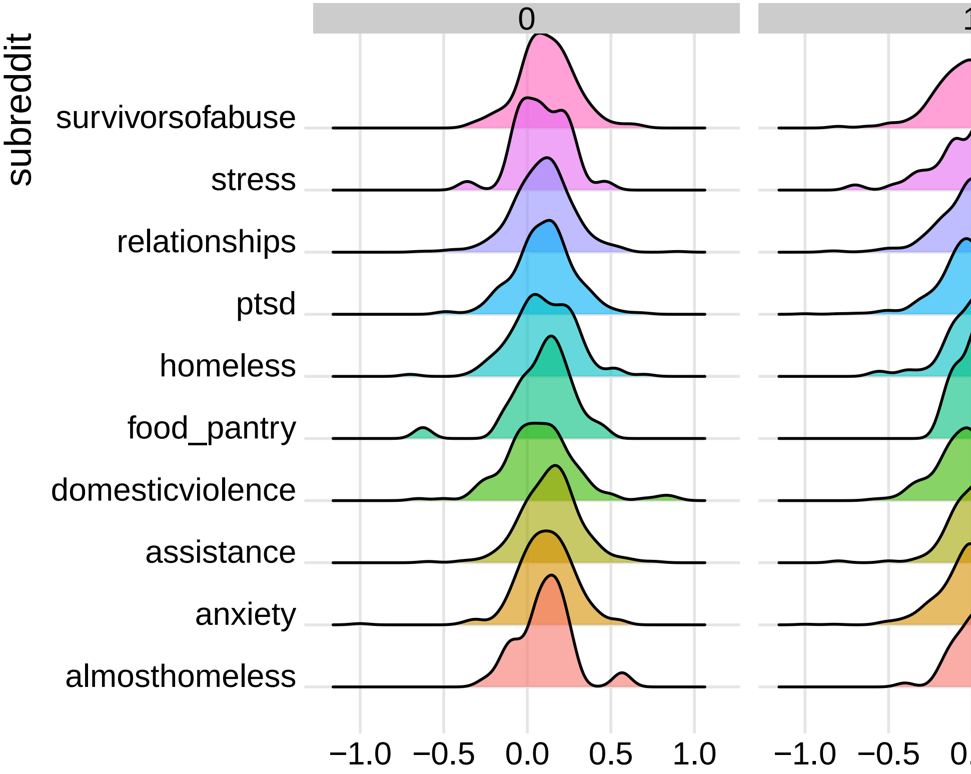
Sentiment Distribution By Subreddit

```
##           subreddit        min           Q1     median        Q3       max
## 1    almosthomeless -0.4012500 -0.033125000 0.07272727 0.1683333 0.5666667
## 2           anxiety -1.0000000 -0.075491071 0.02311508 0.1449026 0.6312500
## 3        assistance -0.8000000 -0.009732143 0.09761905 0.1972171 0.7593750
## 4  domesticviolence -0.6520833 -0.119423077 0.00000000 0.1227652 0.8750000
## 5       food_pantry -0.6250000 -0.033333333 0.04863636 0.1625361 0.4452381
## 6          homeless -0.7000000 -0.049315138 0.05550595 0.1948295 1.0000000
## 7              ptsd -1.0000000 -0.083333333 0.02916667 0.1561111 0.7000000
## 8     relationships -0.8583333 -0.054301949 0.05446429 0.1666667 0.9000000
## 9            stress -0.7000000 -0.071484375 0.05019841 0.1620164 0.4652443
## 10 survivorsofabuse -0.8000000 -0.074743929 0.05000000 0.1743056 0.6666667
##           mean        sd   n missing
## 1  0.075386229 0.1603693  99       0
## 2  0.025978260 0.1928993 650       0
## 3  0.095990995 0.1822430 355       0
## 4  0.003535541 0.2041274 388       0
## 5  0.052999432 0.1749701  43       0
## 6  0.068570269 0.2023446 220       0
## 7  0.029087630 0.1965557 711       0
## 8  0.049170520 0.1930229 694       0
## 9  0.023944684 0.1958851  78       0
## 10 0.045100626 0.1962821 315       0
```

Among the different subreddits, the overall distribution is similar, but the minimum, median, and maximum values are significantly different. The most positive by median is r/assistance, with a median sentiment value of 0.098, whereas the most negative by median is r/domesticviolence, with a median sentiment value of 0. Both r/anxiety and r/ptsd have posts that are completely negative, with a sentiment score of -1. r/homeless and r/relationships have a post with a sentiment value greater than or equal to 0.9.



Between the subreddits, there are significant differences in the overall distribution. r/relationships appears to be the closest to a normal distribution. All of them have significant amounts of skew in their distribution. r/food_pantry looks like it is primarily distributed along the positive sentiment, but has significant negative outliers. Overall, the most negative overall is once again, r/domesticviolence, whereas the most positive is r/assistance and the one with the most spread overall is r/survivorsofabuse.



Sentiment By Label and Subreddit
If we look at the differences by label, we can see further differences along the distribution. r/almosthomeless is normally distributed for stressed posts, whereas right skewed for the non-stressed posts. r/domesticviolence is right skewed for non-stressed posts and significantly more normally distributed. r/relationships looks like it is approximately normally distributed for both the stressed posts and the non-stressed posts. r/food_pantry is extremely skewed, which may be influenced by the smaller number of posts overall. Some of this may be affected by the overall number of posts in each category: for example, r/domesticviolence tends to have a higher proportion of stressed posts, whereas r/relationships has the most posts and tends to be more non-stressed than stressed.

Statistical Analysis

Chi-Square Test

Let's test to see if the stress data by subreddit and label are associated, and set the p-value to be 0.05.

```
## 
##  Pearson's Chi-squared test
## 
## data:  reddit_stress_data$subreddit and reddit_stress_data$label
## X-squared = 158.87, df = 9, p-value < 2.2e-16
```

Since our p-value is less than 0.05, we can reject the null hypothesis that the subreddits do not differ significantly in label by subreddit, and determine that there are significant differences between the label distribution by subreddit.

ANOVA Tests

Now, let's compare the differences in sentiment by label using the analysis of variance test, and again set our p-value to be 0.05.

```
##                             Df Sum Sq Mean Sq F value Pr(>F)
## reddit_stress_data$label    1  12.82  12.820   371.6 <2e-16 ***
## Residuals                3551 122.52   0.035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is less than 0.05, we can reject the null hypothesis that the differences between sentiment by label are not significant and say that the differences between sentiment by label are significant.

```
##                                Df Sum Sq Mean Sq F value  Pr(>F)
## reddit_stress_data$subreddit    9   2.23   0.248   7.212 1.9e-10 ***
## reddit_stress_data$label        1  11.48  11.476 334.179 < 2e-16 ***
## Residuals                    3542 121.63   0.034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

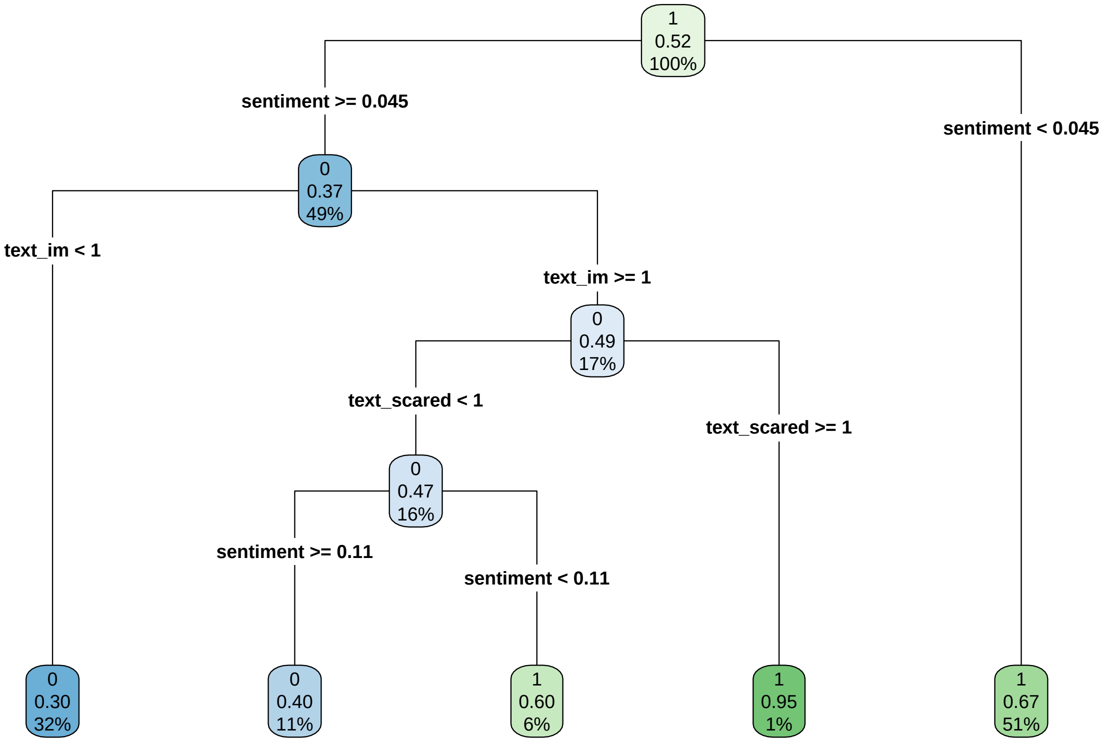
Decision Tree Model

Finally, I will run a decision tree model on the dataset and see what variables are the strongest in predicting whether or not a post is stressed.
A decision tree model is a machine learning model that determines classes by similar groups in the data and then uses them to determine rules for the classification. For example, if $(x_1 > 0.5 \wedge x_2 < 0.5 \implies TRUE)$
For simplicity, I am only going to use the following to predict the result: bag of words model of the posts and sentiment in order to predict the label.
I will read in the data. Since the full dataset has over 1000 variables, I will not display all of them, but just a subset so that you can see some of the variables that were selected.

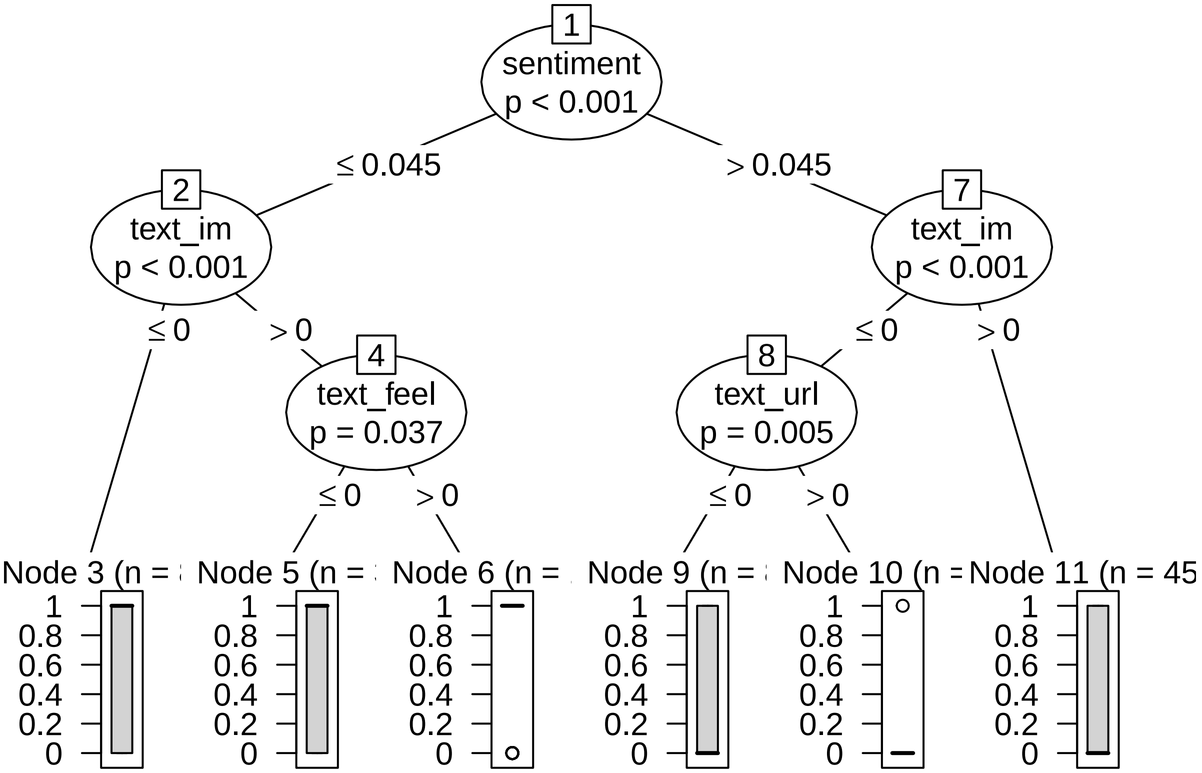
```
## # A tibble: 2 x 5
##   sentiment label text_cancel text_cancer text_constantly
##       <dbl> <dbl>       <dbl>       <dbl>           <dbl>
## 1  -0.00274     1           0           0               0
## 2   0.293       0           0           0               0
```

```
##  logi [1:1724] TRUE FALSE TRUE FALSE FALSE TRUE ...
```

Now, let's run the decision tree model on the dataset and see what the most significant features are for determining whether a post is stressed or not.



According to this model, the most significant features are the sentiment, using “scared”, and using “im”. If sentiment < 0.045, it is likely going to be a stressed label. Otherwise, depending on other features, including whether “scared” appears, whether “im” appears, it may or may not be stressed.



```
##  label
##   0.30 when sentiment >=          0.045 & text_im <  1
##   0.40 when sentiment >=          0.114 & text_im >= 1 & text_scared <  1
##   0.60 when sentiment is 0.045 to 0.114 & text_im >= 1 & text_scared <  1
##   0.67 when sentiment <  0.045
##   0.95 when sentiment >=          0.045 & text_im >= 1 & text_scared >= 1
```

First, let’s see how well my model performs on the training data.

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction    0    1
##          0  769  373
##          1  513 1008
##                                           
##                Accuracy : 0.6673          
##                  95% CI : (0.649, 0.6852)
##     No Information Rate : 0.5186          
##     P-Value [Acc > NIR] : < 2.2e-16       
##                                           
##                   Kappa : 0.331           
##                                           
##  Mcnemar's Test P-Value : 3.015e-06       
##                                           
##             Sensitivity : 0.5998          
##             Specificity : 0.7299          
##          Pos Pred Value : 0.6734          
##          Neg Pred Value : 0.6627          
##              Prevalence : 0.4814          
##          Detection Rate : 0.2888          
##    Detection Prevalence : 0.4288          
##       Balanced Accuracy : 0.6649          
##                                           
##        'Positive' Class : 0               
## 
```

Now, let’s check its performance on the test data.

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   0   1
##          0 250 138
##          1 164 338
##                                           
##                Accuracy : 0.6607          
##                  95% CI : (0.6285, 0.6918)
##     No Information Rate : 0.5348          
##     P-Value [Acc > NIR] : 1.755e-14       
##                                           
##                   Kappa : 0.3152          
##                                           
##  Mcnemar's Test P-Value : 0.1503          
##                                           
##             Sensitivity : 0.6039          
##             Specificity : 0.7101          
##          Pos Pred Value : 0.6443          
##          Neg Pred Value : 0.6733          
##              Prevalence : 0.4652
##          Detection Rate : 0.2809
##    Detection Prevalence : 0.4360
##       Balanced Accuracy : 0.6570
##
##        'Positive' Class : 0
##
```

For both the training and test data, the classification is accurate around 66% of the time. This means that the model does not predict the training data very well nor the test data.

Conclusions and Future Work

Overall, there are significant differences in the data depending on the sentiment, words, and the subreddits by label. The data was imbalanced, resulting in a lower classification.

Future work will include testing the decision tree with more parameters, adding in the other variables, and testing multiple different classification methods.