STAT230–Final Project

Noah Solomon, Oliver Baldwin Edwards, and Martin Glusker

Abstract

The goal of this study was to determine which online comments on the social media website Reddit.com were defined as controversial, based on sentiment analysis and other variables. Multiple logistic regression was used to predict controversiality on a ~400,000 observation data set, in addition to a smaller, balanced dataset with an even balance of controversial and non-controversial comments. No remarkable results were found using the original dataset as none of our models differed meaningfully from the intercept only model. Using the balanced dataset the best model had an accuracy of 57.430%, relative the intercept only model accuracy of 50%. The best model was *Controversiality ~ WordCount + subreddit_scaled_total_mean + subreddit_diff_QDAP + subreddit_diff_LM + relative_sentiment_diff*. Sentiment predictors were found to be associated with controversy, and significantly so due to the large data set, but they were not found to be meaningful as there remained a lot of randomness not explained by the sentiment predictors.

Background and Significance

This project examined how people interact on online forums and social media, an area of particularly relevance in an age increasingly defined by how people interact with one another online. Given the anonymous nature of many internet forums, there has been much speculation about the toxicity of these forums, asking in particular whether these forums incentivize mean of derisive comments? Does the average sentiment of a particular community (in the case of this study, a subreddit) influence which comments are rewarded and popular? For example, do toxic communities reward toxic comments? Or is there another relationship at play?

The primary goal of this study was to see what sort of comments garner attention in the online forums Reddit.com. Researchers were interested in examining the relationship between the sentiment of a comment, and that comment's controversiality (a binary output produced by Reddit). Researchers also looked at the whether the overall average sentiment of a subreddit (a sub-community within Reddit) effects whether positive or negative comments on that subreddit garner attention. For example researchers hypothesized that a subreddit such as 'r/aww', dedicated to sharing photos of cute animals and which was expected to have a very positive average sentiment, would reward comments with positive sentiment much more than negative sentiment. This trend was also hypothesized to apply to toxic subreddits on Reddit, and the relationship between negative and positive comments and controversiality in the context of more negative communities.

Methods

Data collection

These data were collected via a census of comments posted on reddit.com on 01/01/18, 14/01/18, 01/02/18, 17/02/18 yielding approximately 8 million comments. For usability, 100,000 of the comments from each day were then selected at random and the rest were discarded. The data set was then cleaned by removing duplicate rows and those with formatting errors, resulting in around 399,000 samples in the final data set. After modeling with this initial dataset, the researchers determined that the percentage of non-controversial comments was too small, as only ~2% of comments in the dataset were controversial. A balanced dataset was created, where all controversial comments were included, and a equal number of randomly selected non-controversial comments were selected.

Variable creation

Response variable examined was score (equal to upvotes-downvotes), predicted using Stefan Feuerriegel & Nicolas Proellochs' Sentiment Analysis package using the QDAP dictionary compiled by Tyler Rinker. In particular, sentiment analysis was run on the body of the comment, and SentimentQDAP, NegativityQDAP, and PositivityQDAP were used as predictor variables. NegativityQDAP is a quantitative value of how negative the comment was, based on the particular dictionary used, in this case QDAP. In a similar vein, PositivityQDAP is a value of how positive the value is. SentimentQDAP is simply positivity - negativity.

Two new categories of variables were also created, one associated with the comment's sentiment relative to its subreddit's average sentiment. This variable, called 'subreddit_scaled_total_mean' represents the mean of all four different sentiment variables for each subreddit, which were then scaled, and the mean taken of the four scaled mean sentiment variables for each subreddit. This represents an average sentiment score (in units of standard deviation, as it's scaled) for each subreddit. Difference variables were then created, which represent the difference between a specific comment's sentiment and its subreddit's average sentiment. This was done in nominal terms for each of the four sentiment libraries, in addition to a scaled version that represents all four variables.

Analytic Methods

Multiple logistic regression was used to study the association between controversy and SentimentQDAP, NegativityQDAP, PositivityQDAP, and Moderator as well as related factors. The researchers conducted drop in deviance tests to examine which sets of variables produce an effective model. Accuracy on a testing subset of the data was also used as a metric for determining model quality. Accuracy was defined as (correct predictions/total sample size). Finally, randomization tests were then used to increase the robustness of the conclusions.

Results

Data Frame

The researchers' initial data frame was obtained from http://files.pushshift.io/reddit/comments/daily/ and contained all of the comments from the entirety Reddit for one day. They collected four days worth of data (two weekends and two weekdays) and took a random sample of 100,000 comments from each day to end up with a total of roughly 400,000 comments across four days of Reddit content. With this data frame the researchers then added columns that used the R package SentimentAnalysis to keep track of the sentiment of each comment according to different sentiment dictionaries. They also created metrics that kept track of the average sentiment of each comments Subreddit as well as the difference between a comment's sentiment and the average sentiment of their Subreddit. The final dataframe then contained 397,998 observations with 18 variables. The variable of interest that the researchers hoped to predict was controversiality, a binary variable where 1 indicates a comment is controversial, and 0 indicates a comment is not controversial. In the entire dataset, 2.28% of comments were controversial (9,064 controversial comments).

```
## Controversiality
##      0      1
## 388934   9064
```

The researchers then decided to split the dataframe into a training set and testing set so that they could determine a model based on half of the data and then test the accuracy of that model on the second half of the data.

Initial Modeling

For variable selection, the researchers decided to start with a model that used all of the newly created variables (as well as all of the original variables from the initial dataframe) and use likelihood ratio tests to determine which variables were the best at predicting controversiality. The initial model showed that only the variables *Gilded* and *Moderator* were not significant at the 0.05 level. A series of likelihood ratio tests were thus preformed to determine whether or not these variables should be included in our final model. The researchers determined that the variables *Gilded* and *Moderator* both were not significant predictors and did not add anything to the model. From the first series of models, the best model that was found thus included 7 predictor variables and can be seen below in *Table 1.1*.

Table 1.1

```
## Coefficients:
##                               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                 -3.980e+00  2.620e-02 -151.890  < 2e-16 ***
## WordCount                    2.552e-03  4.414e-04    5.781 7.41e-09 ***
## WKNDTRUE                     8.505e-02  3.036e-02    2.802 0.005085 **
## comments_in_subreddit       -6.117e-06  2.879e-06   -2.124 0.033631 *
## subreddit_scaled_total_mean -1.016e+00  6.524e-02  -15.575  < 2e-16 ***
## subreddit_diff_QDAP         -7.723e-01  1.230e-01   -6.280 3.38e-10 ***
## subreddit_diff_LM           -7.616e-01  2.017e-01   -3.776 0.000159 ***
## relative_sentiment_diff      1.528e-01  5.241e-02    2.916 0.003549 **
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 42616  on 196192  degrees of freedom
## Residual deviance: 42253  on 196185  degrees of freedom
##   (2806 observations deleted due to missingness)
## AIC: 42269
##
## Number of Fisher Scoring iterations: 6
```

Randomization Tests

In the final model only three variables had p-values that were greater than 0.001: *comments_in_subreddit*, *WKND*, and *relative_sentiment_diff* (with p-values of 0.0336, 0.0051, and 0.0035 respectively). To confirm that these variables were significant, the researchers decided to perform a randomization test on each to determine if the observed coefficients were statistically significant or not. The results of these randomization tests can be seen in Figures *2.1*, *2.2*, and *2.3* below where the red line indicates the observed coefficient and the density plot shows the randomization distribution. If the red line falls far outside of the density plot, this means that the observed coefficient would not be expected to found by chance and is thus statistically significant. While the researchers were only able to run 100 randomization tests (as opposed to the 10,000 that they hoped to) for each variable, this was due to the limit on computing power that they had access to. This explains the imperfect normal distribution of the plots below, but as the observed coefficients for each predictor fell outside of each randomization distribution, the researchers were able to conclude that each of the three predictors were indeed statistically significant, as the initial p-values for each suggested.

Figure 2.1

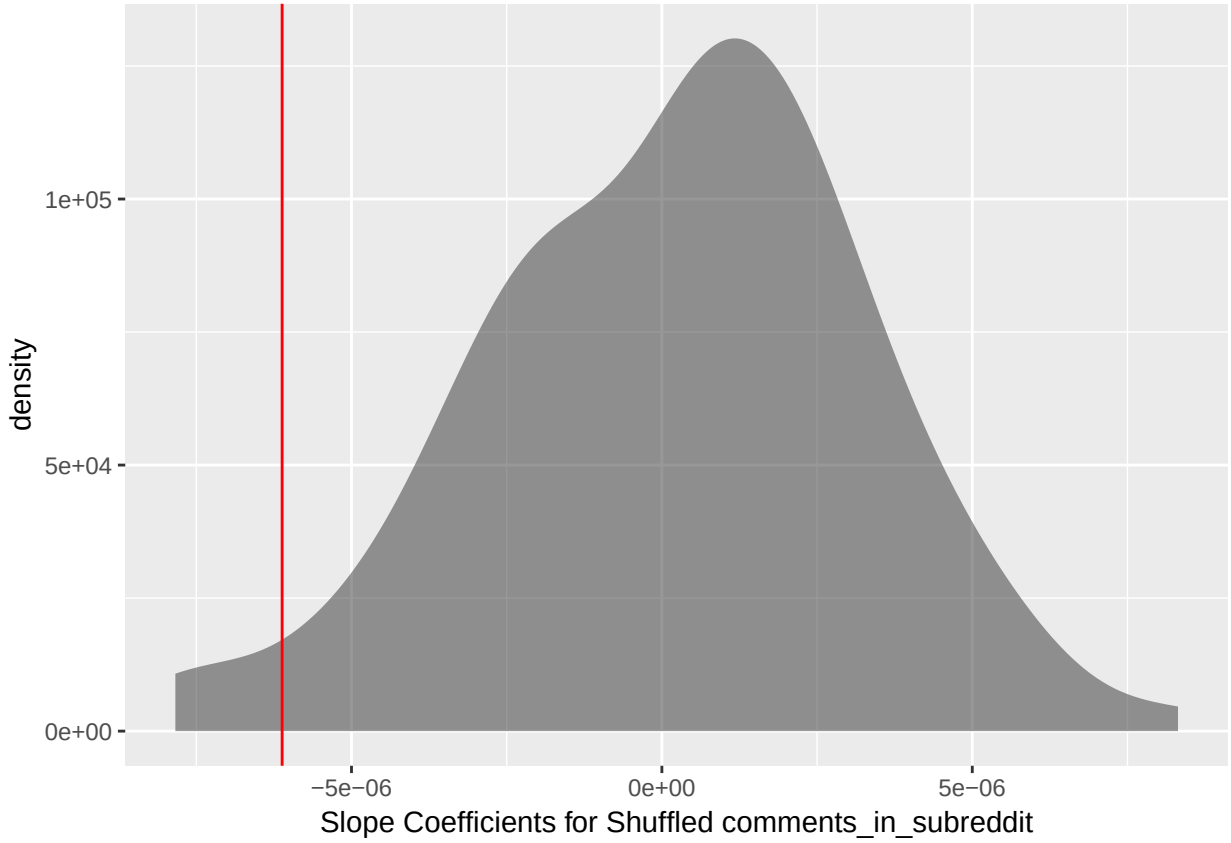

density
1e+05
5e+04
0e+00
−5e−06
0e+00
5e−06
Slope Coefficients for Shuffled comments_in_subreddit

Figure 2.2

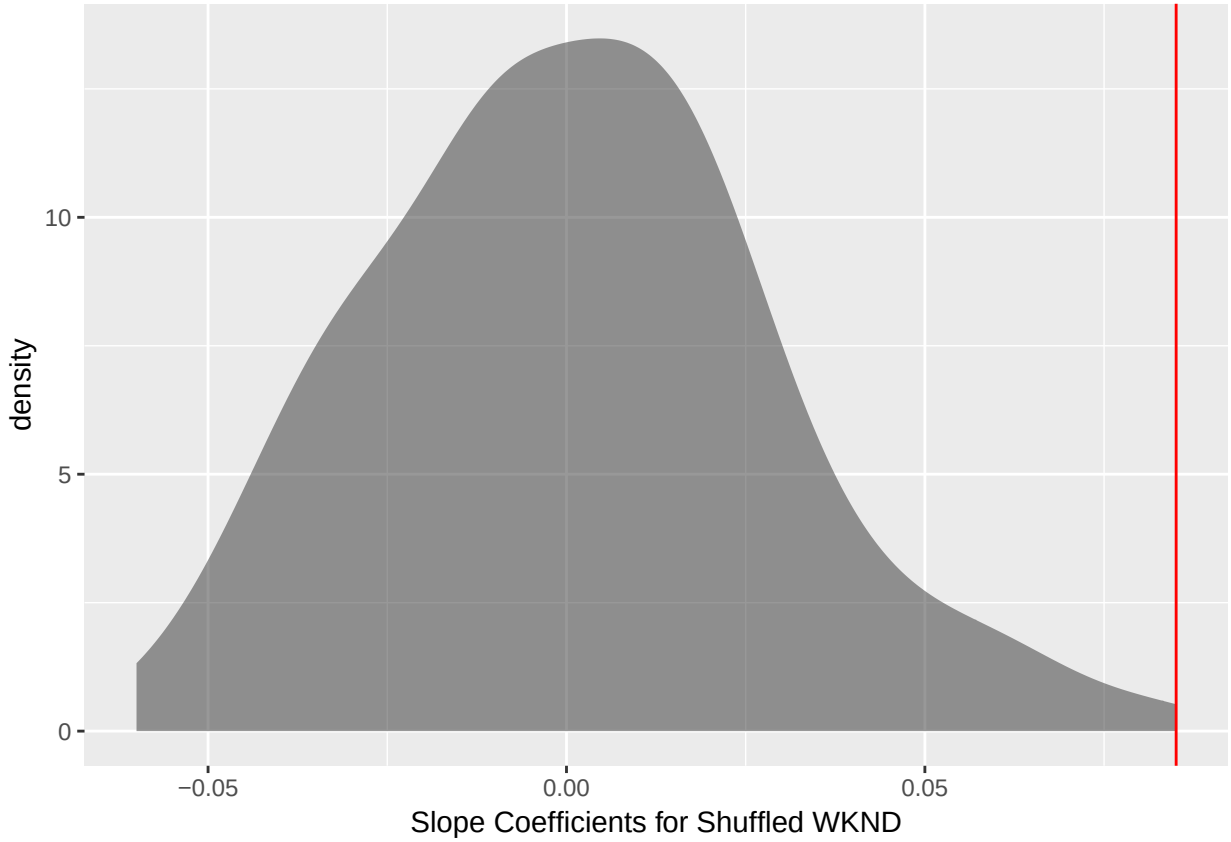
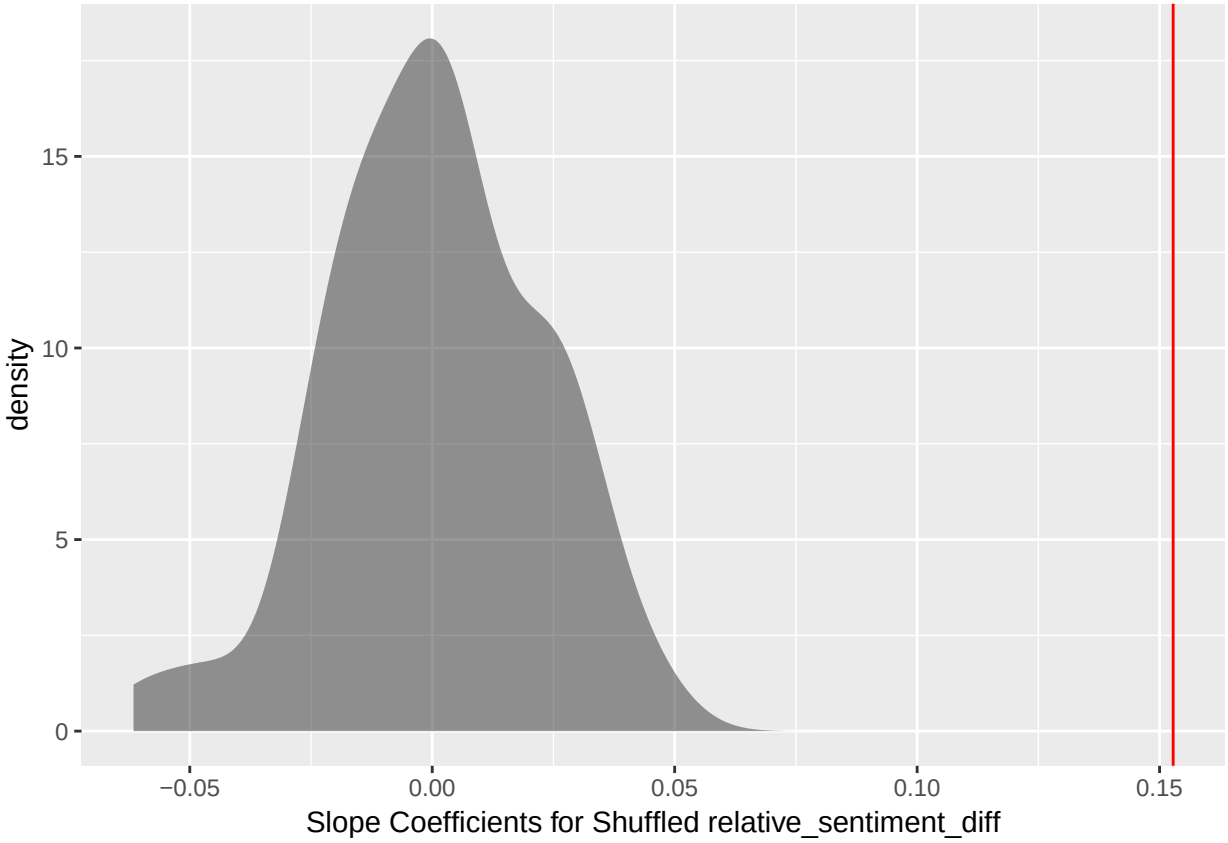

density
10
5
0
−0.05
0.00
0.05
Slope Coefficients for Shuffled WKND

Figure 2.3



Testing Accuracy of Model

With a final model in hand with statistically significant predictors, the researchers then created a function to predict the accuracy of that model using the testing set that was created at the beginning. The researcher's found that the accuracy of the final model was 96.3%, but that the accuracy of the intercept only model was 97.7%.

```
## [1] 0.9633465
```

```
## [1] 0.977226
```

Next Steps: Balanced Dataset

Researchers were concerned with the efficacy of a model that predicted a variable with so few observations. The researchers then decided to resample the initial dataframe in order to have more observations where a comment was marked as controversial (as compared with the 2.28% of comments). This new, balanced dataset, contained exactly 50% of comments that were controversial and 50% of comments that were not controversial. The researchers again split the data into training and testing subsets.

```
## Controversiality
##    0    1
## 9064 9064
```

Modeling with Balanced Dataset

The researchers used the same process with the balanced dataset as with the original dataset to determine the best model to predict controversiality. The final model that was determined by the researchers contained 5 predictor variables and the can be seen below in *Table 3.1.* It's important to note that the *WKND* and *comments_in_subreddit* variables were found to be non-statistically significant using this balanced dataset (while they were found to be statistically significant in the researcher's final model using the original dataset).

Table 3.1

```
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -0.2888174  0.0324828  -8.891  < 2e-16 ***
## WordCount                    0.0022532  0.0008622   2.613  0.00897 **
## subreddit_scaled_total_mean -1.8597738  0.1453709 -12.793  < 2e-16 ***
## subreddit_diff_QDAP         -0.8391867  0.1828052  -4.591 4.42e-06 ***
## subreddit_diff_LM           -0.8567301  0.3130343  -2.737  0.00620 **
## relative_sentiment_diff      0.1700559  0.0789073   2.155  0.03115 *
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 12444  on 8976  degrees of freedom
## Residual deviance: 12189  on 8971  degrees of freedom
##   (86 observations deleted due to missingness)
## AIC: 12201
##
## Number of Fisher Scoring iterations: 4
```

The researchers then found the accuracy of this model to be 56.4%, as compared with the 50% accuracy of the intercept only model.

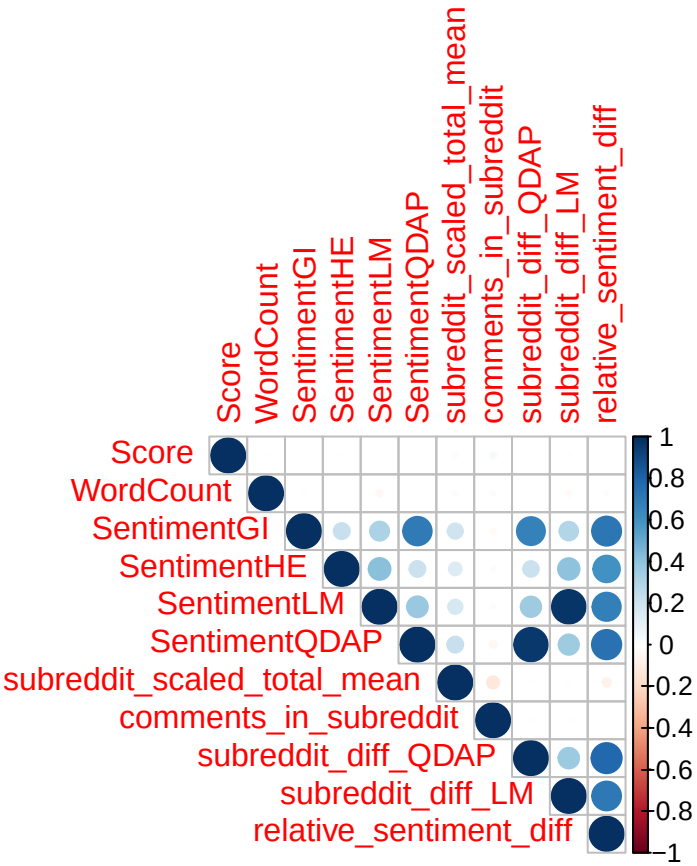
```
## [1] 0.5640375
```

```
## [1] 0.5
```

Additional Insights

The researchers also looked at a correlation plot (*Figure 4.1*) between the quantitative variables in their dataset. *SentimentLM* is highly correlated with *subreddit_diff_LM*, as well as *SentimentQDAP* with *subreddit_diff_QDAP*. This means subreddits are relatively homogeneous and have similar average sentiments. This makes sense intuitively as most subreddits will have a positive and negative comments that will balance out. Another takeaway from *Figure 4.1* is that all nominal sentiment variables are highly correlated, which means that each library is relatively similar and contains much of the same information.

Figure 4.1



Discussion

The objective for this study was to determine whether the sentiment of a comment was a factor in determining how much attention a comment drew. In addition, the researchers looked at other non-sentiment variables to see whether they also proved to be good at predicting how much attention a comment received. The binary variable of controversiality was used as the parameter for "attention a comment receives."

Researchers found that the sentiment of a comment and other variables associated with each comment included in the final model proved to be only marginally better than the intercept only model in the balanced dataset, and very slightly worse than the intercept only model in the original dataset.

The exceedingly low p-values of the coefficients in the final model indicate that they are highly significant predictor of controversy, yet the best models produced in this study, only outperformed the intercept only model by 6-7% in the balanced data set, and the models produced in the unbalanced data set failed to noticeably improve on the intercept only model. These two conclusions may seem paradoxical, but in fact they reveal a deeper truth about the data set, namely that although the rather enormous number of observations allowed the researchers to detect significant but minor associations between the sentiment variables and controversy, ultimately those minor associations are not very helpful in accurately predicting controversy. In other words, sentiment predictors are associated with controversy, but that association is largely overcome by the randomness in controversy not captured by sentiment predictors.

In terms of answering the original research objectives, researchers did find an association with sentiment and controversiality, in Reddit comments, but found that sentiment was not an very useful predictor. Given the scope of this study, researchers can conclude that sentiment analysis does provide some insight for predicting

controversy, but that information is not comparatively that useful absent further analysis of other variables. Ancillary conclusions include the relative homogeneity of sentiment across subreddits, indicated by the high correlation between relative and absolute sentiment. Additionally, researchers found information provided through sentiment from the four dictionaries to be relatively redundant, as indicated by the high correlation between the sentiment values produced from each of the four dictionaries.

Limitations & Areas for Concern

The scope of this study is limited by its design in a number of ways. First of all, this is an observation study, not a randomized experiment, thus association between the variables considered can be analyzed, but causal claims would not be appropriate. These limitations can be grouped into three broad categories; computation time, dataset characteristics, and complications from data cleaning.

One of the primary problems encountered in this research was the lack of computation time available to the researchers. The initial data set included close to 8 million observations of some two dozen columns and running operations on a set of this magnitude rendered unfeasible run times on the hardware used. A more full analysis would have ideally incorporated a greater number of days of comments, but this was simply not feasible given the computational constraints. As such, caution must be taken in broadly applying any of the conclusions of this study outside of the date range analyzed.

Second, there are fundamental limit in the substance of the data set that limit the conclusions that can be drawn. The analysis is limited to Reddit.com and is not necessarily representative of internet comments more generally as Reddit.com has a user base that tends to lean younger and more male than other internet forums. In addition, the nature of Reddit is such that a handful of very popular communities (subreddits) tend to produce the vast majority of the comments, meaning that comments from these popular subreddits tended to dominate the data set, limiting the conclusions of this study to the larger subreddits. In addition, of the four days selected, at least one of them was in close temporal proximity to a major world event, vis. the shooting in Parkland Florida. As such it is possible that comments collected from days that were in close proximity to high level world events may not be fully representative of Reddit comments in general. Another potential limiting factor is the presence of 'bot' comments in the data sets. 'Bots' are Reddit accounts that follow automated protocols producing generally formulaic content for moderation or similar purposes. The researchers were not able to remove these bot comments from the data set and their presence means that the claims made in this study cannot be unconditionally applied to human comment only contexts. This likely does not confound the data at all, but does make associations more challenging to detect. Lastly there are limitations for this study derived from the data cleaning process. Duplicate comments were only able to be removed after the sample was pared down to ~400,000 comments due to computational time limits, meaning comments that were appeared more than once in the original ~8 million observation data set were disproportionately likely to make it into the final data set. That being said, the relative infrequency of duplicate comments and the size of the data set render this very unlikely to meaningfully influence the conclusions.

Lastly, a variety of malformations were present in the original uncleaned data arising from, among other things, unusual or improperly escaped characters in the body of the comment. As such, the incidence of these malformations was greater among longer comments, so the removal of malformed rows increased the proportion of short comments to long comments. Only about 400,000 comments were removed in this fashion, or about 5% of the original data set, so the proportion is not likely to have meaningfully shifted.

Areas for Further Research

`As mentioned above, one of the primary limitations of this study was the limited computing power the res`

References

Data collected by Reddit User u/glitch_in_the_matrix, hosted on pushshift.io by Jason Baumgartner (http://files.pushshift.io/Reddit/comments/daily/)