NYU Center for Data Science

Is Twitter Outrageous?

Alex Simonoff and Rob Hammond
NYU Center for Data Science
Advisor: Billy Brady, NYU Jay Van Bavel Lab

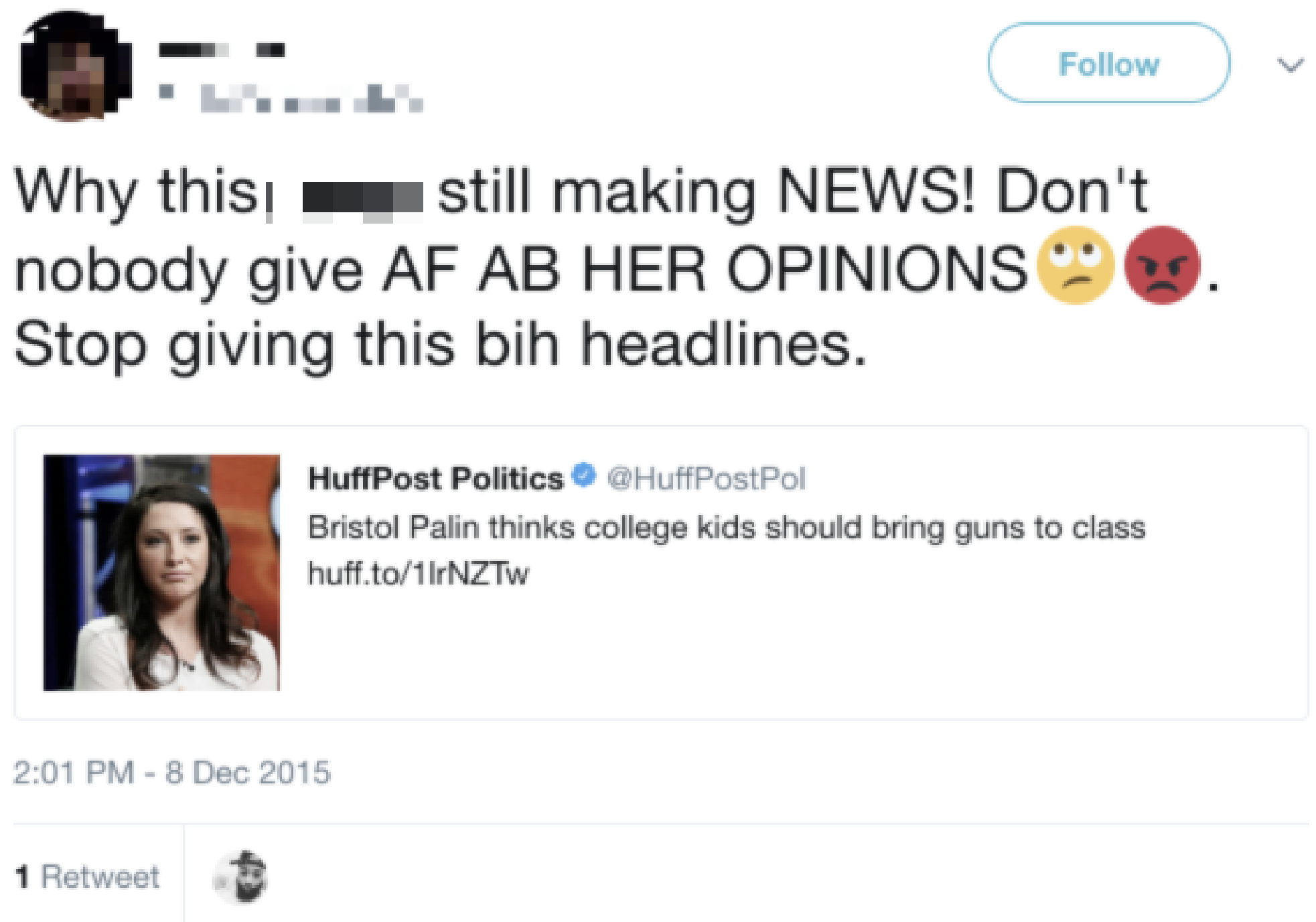
JAY VAN BAVEL
SOCIAL PERCEPTION AND EVALUATION LAB
NEW YORK UNIVERSITY DEPARTMENT OF PSYCHOLOGY

Introduction

Over the past two decades, social media has grown in popularity and taken on a role that extends well beyond simply connecting people on the web. A considerable amount of people are not only consuming news online, but also responding to it and engaging with people who share or disagree with their opinions. For our project we looked at social engagement with morally and emotionally charged messages on Twitter, focusing specifically on tweets related to the divisive issues of same-sex marriage, gun control, and climate change. We use various techniques to measure the level of outrage present in the language of a tweet and gauge the level at which this outrage might have prompted engagement.

Background

Following up on work done done by Billy Brady and Julian Wills of the Jay Van Bavel Lab at NYU, we sought out to understand how outrage manifests in language on Twitter. Work previously had largely focused on word counts, however we wanted to try more sophisticated methods in an attempt to develop a more robust measure. Using over 500,000 unique tweets from late October 2015 to mid December 2015, we used word counts, emoji mining, word mapping (see: Valence and Arousal), text classification and latent Dirichlet allocation (LDA) to build an outrage index. This index will look at various measures of outrage and average across them to develop a more robust and less noisy prediction.



[redacted]
[redacted]

Why this [redacted] still making NEWS! Don't nobody give AF AB HER OPINIONS 🙄😡. Stop giving this bih headlines.

HuffPost Politics @HuffPostPol
Bristol Palin thinks college kids should bring guns to class huff.to/1IrNZTw

2:01 PM - 8 Dec 2015

1 Retweet

Figure 1: An example of a tweet in our dataset

Emoji Mining and Outrage Counts

In our data there were several tweets that included 'emojis' (emoticons) represented as strings following a pseudo-unicode format. We established a happy emoji and three outraged emoji to build out a net emoji sentiment score (1 indicates the presence of an outraged emoji, -1 a happy emoji and 0 indicates no presence or a net neutral). In the example tweet shown in Figure 1, the red outrage emoji gives this tweet a net emoji score of 1. We also looked into a basic count of words in an outrage dictionary provided by our advisor. We then used Google News pre-trained word2vec vectors to find words close to our outrage words and established an expanded outrage dictionary.

Valence and Arousal

Valence is a measure of how positive or negative of an affect a tweet has, whereas arousal is a measure of how pacifying or stimulating a tweet is. In our case we are particularly interested in tweets with a low valence and a high arousal rating, which is where outrage inducing words and tweets will cluster.

Table 1: Examples of different levels of valence and arousal

Tweet Body	Arousal	Valence
If the USA didnâĂŹt give guns to all and sundry they wouldnâĂŹt have to worry about mass shootings #TRUMP #IDIOTOFTHEDAY	0.81105	0.76921
Obama Spends 27 Percent Of Speech Scolding Americans On Guns, Racism; 8 Percent On ISIS Threat https://t.co/osKS1KJ5QX via @dailycaller	0.89147	0.19705
Climate change getting so bad the State Dept. has issued a worldwide, travel warning https://t.co/CiGCVfdMbW	0.48094	0.61565
Enrique Marquez, Buyer of Guns Used in San Bernardino Attack, Is Studied by RICK ROJAS and IAN LOVETT via NYT https://t.co/NJwpfgzzfp	0.32946	0.68201

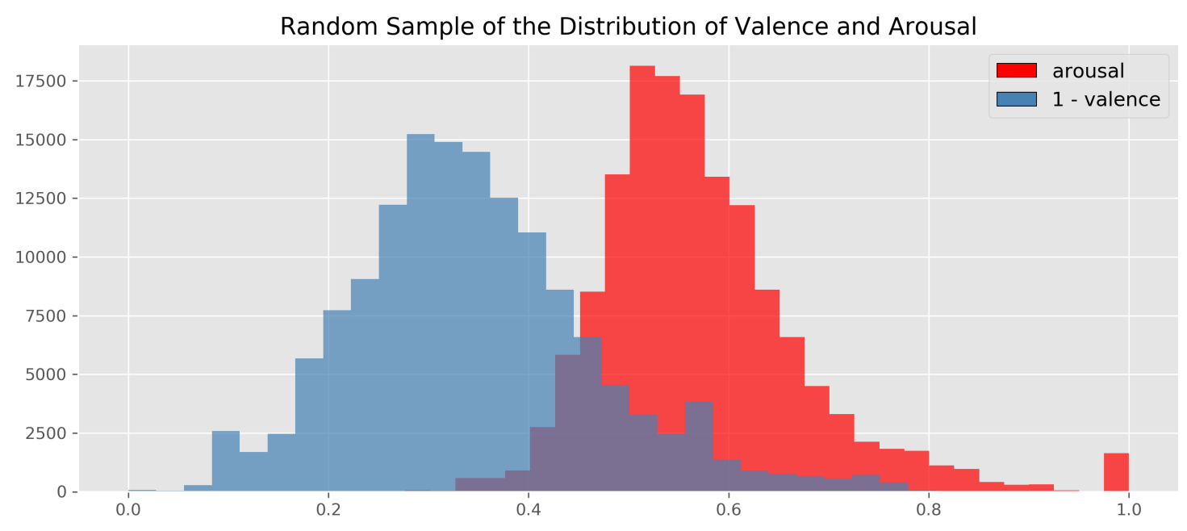


Figure 2: Distribution of 1-valence and arousal scores across a random sample of tweets.

Naive Bayes Sentiment Model

Given our data was not labeled, we sought out an alternative Twitter datasets with labels and ultimately used Sentiment140's sentiment labeled dataset of 1.6 million tweets. Using a TFIDF vectorizer with a bigram approach, we trained the model on Sentiment140's data and then fit it on our dataset. We unsurprisingly saw a bell curve with a peak at 0.5 as the model predicts this when there are bigrams in our dataset that the model is unfamiliar with.

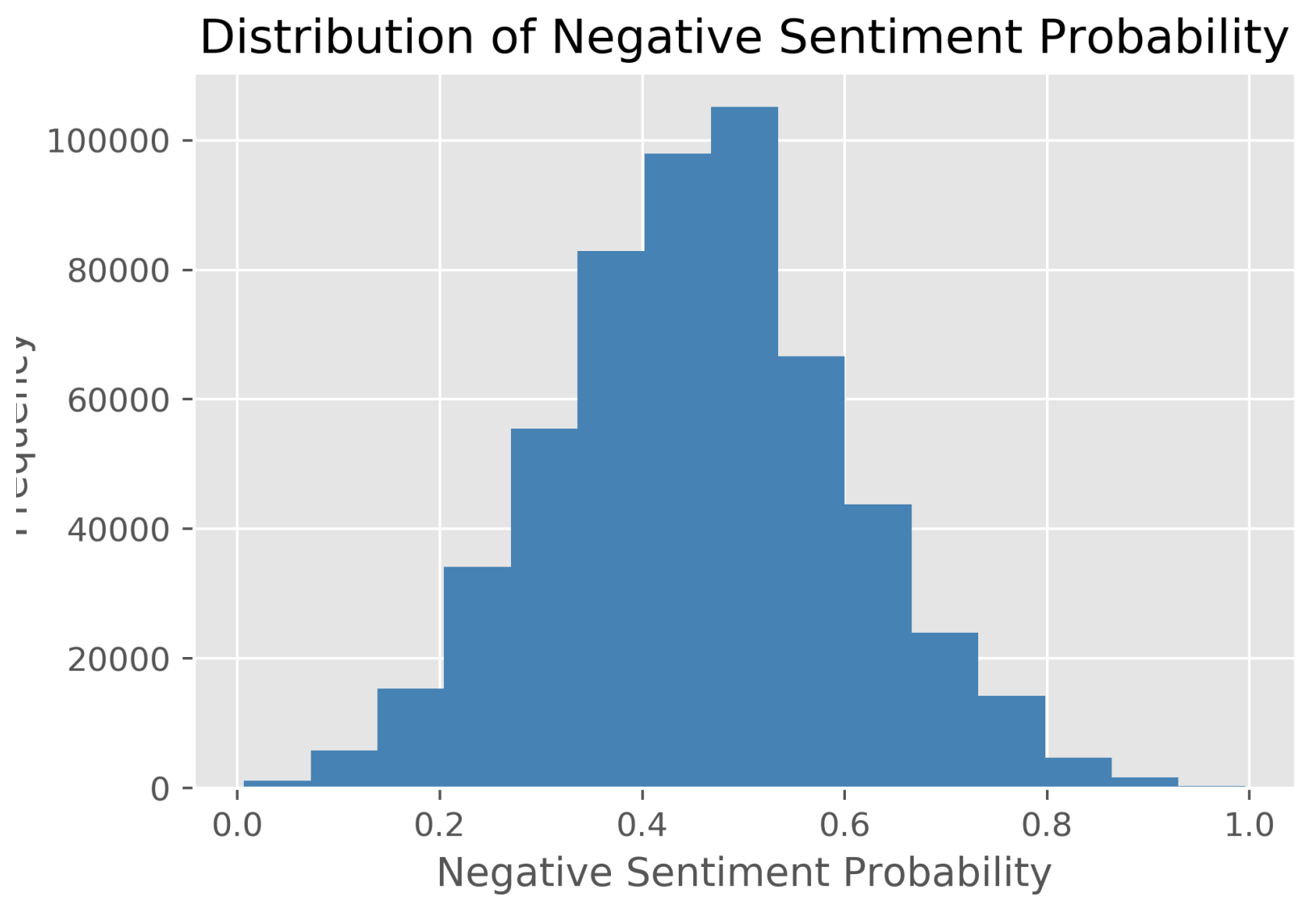


Figure 3: Distribution of sentiment scores (probability of negative sentiment) across tweet corpus

LDA for Topic Modeling

Latent Dirichlet allocation(LDA) is a model that transforms a term frequency-inverse document frequency matrix for a corpus into a word and topic distribution using a Bayesian framework with Dirichlet priors. LDA was favored over more recent innovations such as clustering with word2vec, because it has more easily interpretable outputs, and seems more accepted within the social sciences. However, we did use a word2vec coherency metric as one of our measures for choosing an optimal model. Ultimately, we decided on 60 topics as the optimal topic number through our different coherence metrics, and through looking at the top words over a number of the topics.

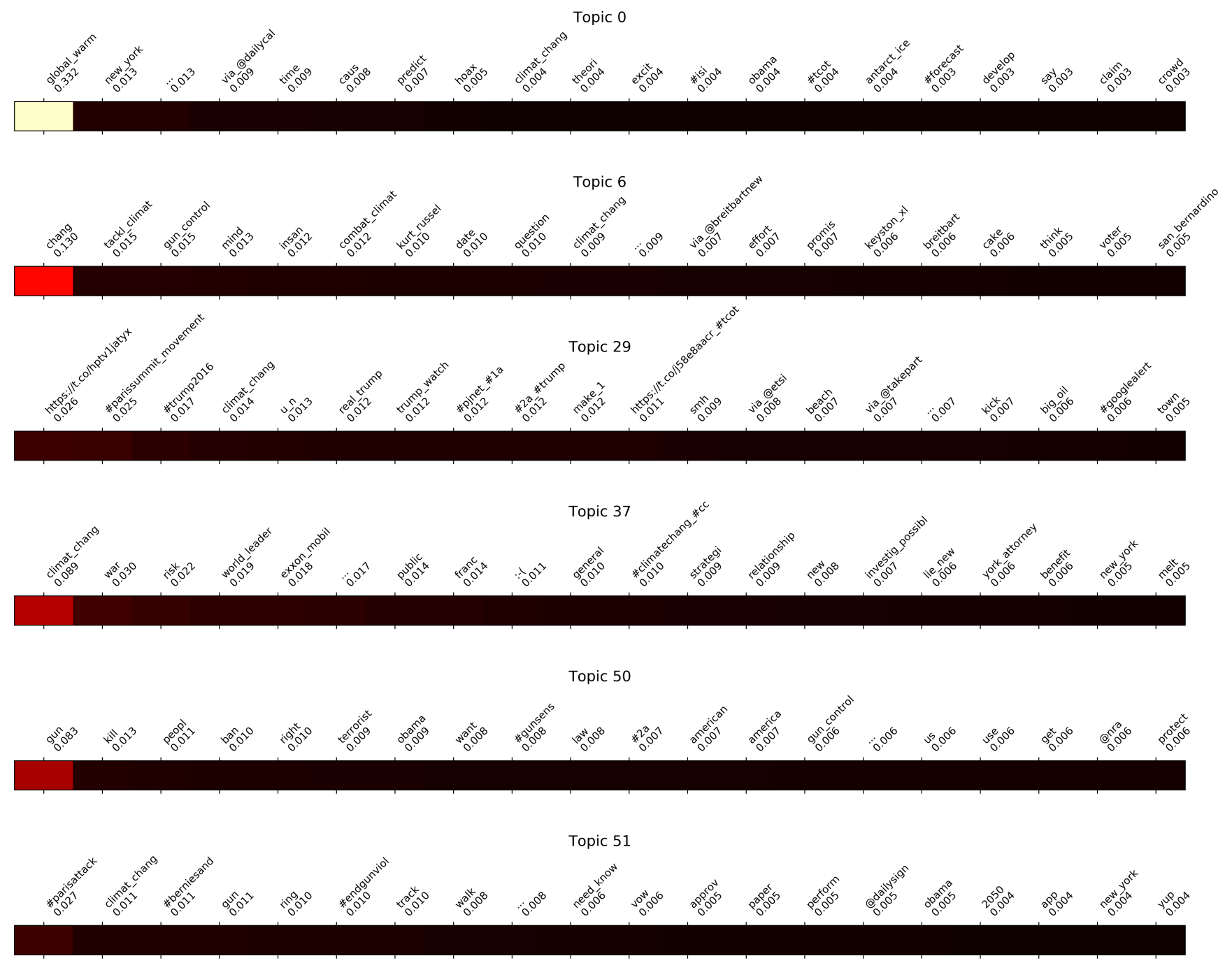


Figure 4: Examples of extracted topics demonstrating outrage and/or cohesion

In addition to capturing the probability that a tweet belongs to one of the topics, we want to know how much valence and arousal each tweet elicits. By weighting each word's valence and arousal by the probability that it belongs to a specific topic and rescaling to a 0-1 range we are able to see how much "outrage" a topic is inducting. As previously mentioned, we are concerned with the cluster of high arousal and low valence topics.

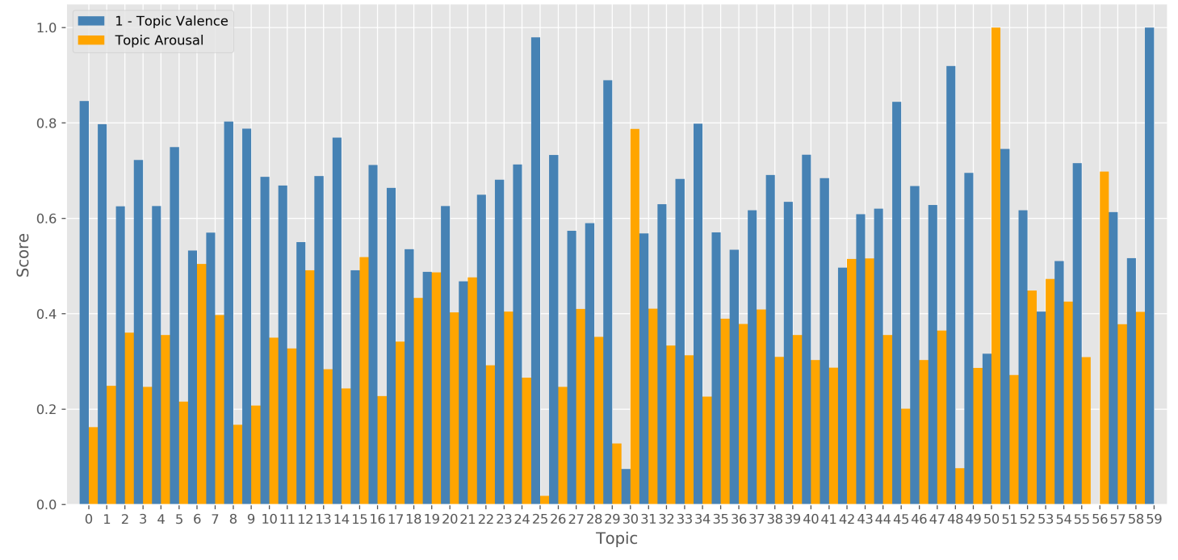


Figure 5: Distribution of scaled valence and arousal scores across all 60 topics

Outrage Index

In order to create an effective outrage metric we had to scale each of our metric components to a 0-1 range, and create a weight for each component, so we can arrive at a final measure in the 0-1 range. The 0-1 range is quite useful for creating an easy-to-understand metric. For example in our earlier tweet we have a score of **0.541** and we have gone through the features contributing to this score in the table to the right.

Outrage Index Continued

Table 2: Outrage Index Elements and Weights

Feature	Topic Valence	Topic Arousal	Valence	Arousal	Outrage	Sentiment	Expanded Outrage	Emoji
Outrage Index Weight	0.15	0.14	0.14	0.14	0.12	0.11	0.10	0.10
Example Tweet Values	0.504	0.366	0.348	0.548	0.5	0.628	0.6	1.0

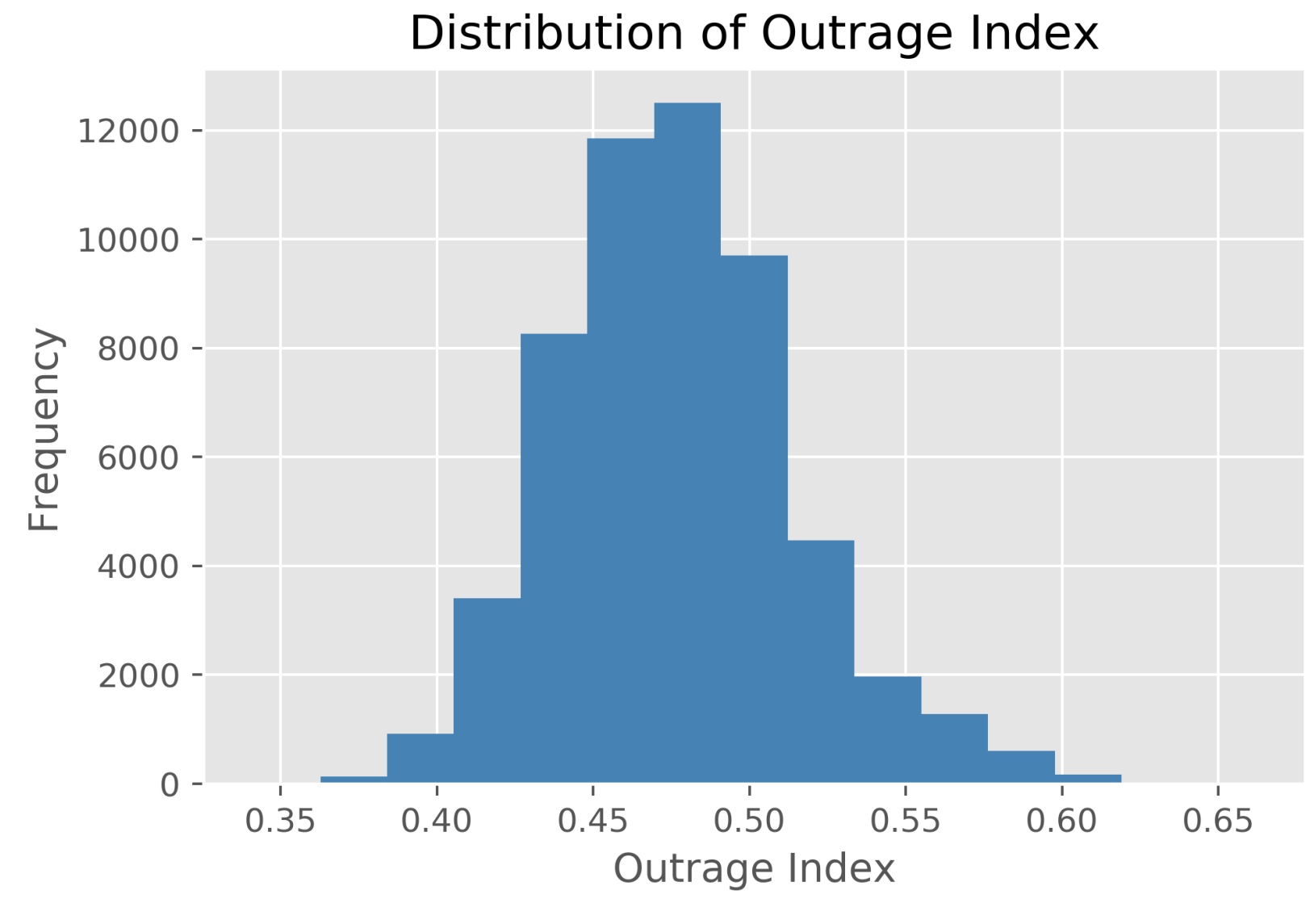


Figure 6: Distribution of randomly sampled outrage scores

Engagement

After running a number of baseline tweets we found that we are unable to accurately measure engagement (the log of the number of retweets) with our outrage indices or final score. Using LASSO and Ridge regression we got poor results with $R^2 < 0.2$, however we did find that topic valence and topic arousal had the highest coefficients with our LASSO model with all features included, and the outrage score had a higher weight than the number of followers when all other features are removed! For the Random Forest approach we found similarly low $R^2 = 0.09$ but the number of followers was far more predictive than the outrage index. When we simply look at the Pearson's correlation between outrage and engagement, we see roughly a 0.04 correlation (significant at $\alpha < 0.05$; This value increases to 0.24 when we interact the outrage with log followers. With that said, when we remove the tweets that have 0 to 4 retweets and repeat, we don't get a significant correlation between outrage and engagement (though the interaction is still significant at $\rho = .17$).

Results and Discussion

Ultimately, we accomplished what we set out to do. We established a more rigorous approach to determining outrage in tweets and produced code that will be made open source and can be applied to new tweets fed to the model. While our focus was not as heavy on engagement, we did find some indications that engagement might be linked to outrage, however we could not determine it to be statistically significant.