Classification of Twitter Outrage Using Supervised and Unsupervised Techniques

Rob Hammond and Alex Simonoff
New York University
Center for Data Science

Abstract

This project sets out to model outrage in Twitter messages surrounding the divisive topics of gun control, climate change, and same-sex marriage. Taking a comprehensive approach to this task we measured eight key features that relate to the emotion and sentiment of each message. We built a topic model using latent Dirichlet allocation (LDA) to uncover the latent topics present in the messages. We then modeled the valence (the measure of positive and negative emotion) and arousal (the amount of emotional stimulation) of each message and each topic to find the most negatively stimulating messages. In addition, we counted the number of outrage related words and their closest neighbors in word2vec to capture the presence of outrage sentiment. Finally, we measured the sentiment using both emoji indicators and a Naive Bayes sentiment model. Using these eight metrics were able to create an outrage score to be used in studying the effects of outrage on the transmission of messages over Twitter.

1 Introduction and Background

In this paper we consider the problem of using unsupervised and supervised methods to model outrage, which we understand as a measure of anger, frustration, etc., in terms of the way users phrase messages on Twitter surrounding divisive topics. The goal of this work is to create a comprehensive metric to allow researchers to better understand how the framing of messages with more or less outrage can affect how humans interact with each other over social media through retweets and likes. Do higher levels of outrage lead to higher rates of transmission through those means?

Our research stems from the work of Brady et al. (2017) where the authors found that the presence of moral emotional (as opposed to the non-moral emotional) words affected the transmission of messages on Twitter within and outside of one's own political ideology. This was done in their work by studying messages about climate change, same-sex marriage, and gun control. Given the success of their work in finding that the presence of moral emotional words increased the transmission of messages by about 20% per word, we are extending this study by specifically investigating whether or not the existence of emotional outrage in Twitter messages similarly affects its rate of transmission. If the trends can be found on a broader level, can those elements be pinpointed to specific emotions?

In order to consider this task, we will use the Twitter data collected by Brady et al. (2017) that concern three divisive topics in American politics around major events pertaining to gun control, climate change, and same-sex marriage (see 2 for more information about the data).

Because of the results from Brady et al. (2017) we know that the language in this data set leads to higher levels of transmission, which then makes it ideal for the task of measuring the presence of outrage. Given the inherently unstructured and unlabeled nature of Twitter data, it was necessary for us to construct methods to capture the underlying ideas surrounding outrage. Because of the lack of labels or other guides we have mainly utilized unsupervised learning techniques; however, we have also been able to use supervised techniques for some tasks.

To create a comprehensive metric, our outrage score, we consider a number of different features that we have identified as good indicators of the existence of outrage. Specifically, we measured the valence of each message: the amount of positive or negative language present, the arousal: the degree to which a message is emotionally pacifying or stimulating, the overall sentiment of a message and how likely it is to be negative, whether or not emoticons (emojis) were present and if they contribute to a negative sentiment, the amount of outrage-specific words present, and a topic model to uncover the latent topics of messages and how they relate to valence and arousal. Overall, we used eight metrics defined in Section 3 below that allowed us to create one single outrage score to indicate level of outrage present in any given Twitter message.

It is important to consider a wide variety of unique and overlapping elements in this score so that we are able to validate how certain messages can be high or low sentiment. For instance, if a message contains several words that are in our outrage dictionary, then we can presume that it is negative with a fair amount of confidence. However, when we pair this with an overall sentiment, low valence (negative), and high arousal (stimulating) we can say with very strong confidence that a message contains outrage.

Overall, as will be explained in this paper, we will demonstrate that outrage can be measured through eight key metrics: valence, arousal, topic valence, topic arousal, outrage word counts, expanded outrage word counts, emojis, and sentiment modeling. These metrics can then be rescaled to compute one end outrage score, so that researchers can explore the effects of the outrage emotion on message transmission on Twitter.

2 Data

Our data consists of over one million messages collected over a period of twenty-two to forty-two days surrounding a major event for three divisive topics in contemporary politics: same-sex marriage, climate change, and gun control. For same-sex marriage, data was mined following the Supreme Court of the United States's ruling in favor of same-sex marriage in *Obergefell v. Hodges, 135 S.Ct. 2071 (2015)*, June 2015. The major event we used for climate change was the 2015 United Nations Climate Change Conference, COP 21 or CMP 11 held in Paris, France, and often referred to as the "Paris Summit." And, the major event triggering our gun control data was the San Bernardino shooting that occurred in early December 2015. For more information on the search criteria please see Supporting Section 1 of [5]. To see a sample message, see Figure 4 in the Appendix. In order to use the data we employed both Python's NLTK [3] and gensim [10] packages for message tokenizing, stemming, and creating phrases from commonly co-occurring words.

In preprocessing our data we collapsed all forms of a word together, in order to create a more useful understanding of the meaning of a word, as opposed to the nuanced nature that lemmas can provide. For example, we have found that words such as "absolute" and "absolution" are actually not much different in sentiment for the purpose of this study even though they hold different meanings and are used in different contexts, as such, we will take the stemmed form of "absolut." This is an important distinction given our focus on sentiment to understand outrage. We created these stemmed tokens using NLTK's TweetTokenizer and SnowballStemmer to both tokenize and stem the words: two widely accepted functions in natural language processing (NLP). The use of TweetTokenizer was essential to capture the wide array of language and form found on Twitter, e.g., hashtags, links, excessive punctuation, etc., so that we can maintain the meaning without stripping the characters that are essential to Twitter itself, information that would be imperative to this study.

Creating the phrases proved to be a bit more challenging, as there are not many named entity recognition (NER) functions for Python without using Java, which would have been out of the scope of this project. Instead, we used gensim's phrases module that allows for pseudo NER by grouping commonly co-occurring words to allow for new phrases such as "New York" or "New York Times." This is particularly important because the words "new," "york," and "times" all hold vastly different meanings separately than when placed together. Pseudo is important here as it also allows us to consider more than just pronouns such as "climate change" or "paris summit," etc., which on their own take on a completely different meaning; but, when grouped together, we know that both terms have to do with the perilous nature of our future, and not (what the individual terms might suggest): long term weather or a mountain near Paris.

3 Methods

In this section we will describe, in detail, each of the methods that were used to create the outrage score. We will talk about topic modeling, sentiment classification, building outrage dictionaries, emoji mining, measures of valence and arousal, and how they relate to one another.

3.1 Topic Modeling

Latent Dirichlet allocation (LDA) is an unsupervised generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document [blei].

We chose to use LDA for topic modeling as an outrage measure since it is capable of capturing the latent topics of a message. This is useful because it can be linked to nuanced understandings of a topic and how they relate to outrage. In 3.3 we will link these topics back to valence and arousal for validation. While there are more sophisticated unsupervised methods for latent discovery such as word2vec [8], tweet2vec [7], or lda2vec [9] that create topics by using various clustering techniques such as KMeans clustering, LDA satisfied more of our requirements in various ways.

The first is that the model is probabilistically robust. LDA uses a Bayesian method that relies on a Dirichlet prior that creates a continuous probability distribution over its topic-word and document-topic distributions, which allows for easy interpretability. Specifically, LDA captures document level associations by creating sparse representations. On the other hand, word2vec is dense vector of distributed representations that capture very local word associations. LDA is a topic model, whereas word2vec is mainly used as an embedding for NLP tasks utilizing neural network models. In addition to the interpretability of LDA topics from these probability distributions, it is already an accepted topic model, even if it may be contentious within its own field, as opposed to more recent models such tweet2vec or lda2vec, which are still in preprint. Both of these models seem quite promising, but have not yet proven to be effective in practice.

The fact that this method is widely adopted and probabilistically robust makes it more likely to be accepted within the psychology field compared to newer, less interpretable, hybrid methods. Additionally, LDA and word2vec represent two different worlds of modeling. With LDA, a user is looking to explicitly uncover latent topics for document modeling, whereas word2vec is often the first step of neural network modeling, and is most commonly a form of supervised learning where topics or other information is known ahead of time. This would not be as useful for our project as we do not have labeled data and we want to specifically uncover message level associations present in our data as opposed to local associations.

To train our LDA model we relied on the gensim ldamodel function. The advantage of gensim's implementation is that it utilizes Cython for large performance boosts; it also allows for parallelization of multicore machines for more speedup. But most importantly is that it allows for automatic tuning of the α and η priors for LDA. This is beneficial because it frees up effort from having to research optimal hyperparameters ahead of time, and allows us to focus more on convergence and optimal topic numbers.

The next step for this project was to model the optimal number of topics to represent the corpus. We knew ahead of time that there were three overarching topics, thus anyone could presumably create a model of association rules to uncover the three main topics. We therefore sought to understand the nuanced meanings of each overarching topic and how they may interact by creating a series of models ranging from ten through one hundred topics. We then tested the `c_v`, `u_mass`, and `c_w2v` coherence measures for each topic to arrive at an optimal number of topics. The three measures are validated through human readers [11] or offer a more contemporary view of latent discovery within data science. In addition, the three measures all use a different framework whether it is pointwise mutual information (PMI), conditional probabilities, or normalized PMI (NPMI) vector representations for word similarity, and therefore finding one that can optimize all three is ideal.

After arriving at sixty topics for our ideal model we then performed a fine-grained search by training new models in a range of fifty-eight to sixty-two topics in an effort to ensure sixty was in fact our optimal topic number. Because sixty topics still proved to be the optimal number of topics we then

ran our model once more for more iterations to allow the model to further converge utilizing ten passes over the randomized corpus with chunks of ten thousand documents, as opposed to just five passes in our initial tests.

In Table 1 (see below) we have provided a sample of some our most coherent (human-assessed) topics from the LDA model. In the table we can see that Topic 29 is relating to activism in relation to both climate change, the Second Amendment, and Donald Trump. In Topic 37 we can see that the topic entails anger about climate change considering the "exxon_mobil" phrase and sad face. Topic 50 is about guns and terrorism, and seems to have a high propensity for outrage given the words "kill," "ban," and "terrorist". In addition we are also likely seeing some mix of the "right" complaining about then President Barack Obama taking away guns and people likely also talking about "#gunsens" from Obama's attempted policy initiatives. In either case we can safely assume that this topic is highly stimulating and quite negative. Finally, in Topic 51, we can see that there is an array of emotions but generally talking about the attacks in Paris on November 13, 2015. One may think it is peculiar to see "#berniesand" and "climat_chang" here, but this was also at the Democratic Debate when Bernie Sanders stated that "climate change is directly related to the growth of terrorism."

Topic 29	**Topic 37**	**Topic 50**	**Topic 51**
https://t.co/hptv1jatyx (0.026)	climat_chang (0.089)	gun (0.083)	#parisattack (0.027)
#parissummit_movement (0.025)	war (0.030)	kill (0.013)	climat_chang (0.011)
#trump2016 (0.0170	risk (0.019)	peopl (0.011)	#berniesand (0.011)
climat_chang (0.014)	world_leader (0.019)	ban (0.010)	gun (0.011)
u_n (0.013)	exxon_mobil (0.018)	right (0.010)	ring (0.010)
real_trump (0.012)	... (0.017)	terrorist (0.009)	#endgunviol (0.010)
trump_watch (0.012)	public (0.014)	obama (0.009)	track (0.010)
#pinet_#1a (0.012)	franc (0.014)	want (0.008)	walk (0.008)
#2a_#trump (0.012)	:-((0.011)	#gunsens (0.008)	... (0.008)
make_1 (0.012)	general (0.010)	law (0.008)	need_know (0.006)

Table 1: Top Ten Words By Topic: In this table we show four of the most interesting topics from the sixty produced using LDA with their top ten words and topic-word probabilities ranked by probability.

3.2 Valence and Arousal

An important concept in understanding the overall sentiment of each message is to analyze the valence and arousal of each word in aggregate. These two concepts, which stem from the field of behavioral psychology that we utilize frequently in this project, are important because they represent how positive or negative a word is (valence) and how pacifying or stimulating a word is (arousal).

Given these two metrics we wish to consider messages with high arousal and low valence to capture negatively stimulating messages, which would be highly indicative of outrage. To do this we used [12] dictionary containing the mean and standard deviation valence and arousal rating for 13,915 words, and used the standard deviation weighted mean in Equation 1 below for the entire message

$$Valence = \sum_{w} \mu_w * \frac{\frac{\sum_i \sigma_i}{\sigma_w}}{\sum_i \frac{\sum_j \sigma_j}{\sigma_i}} \tag{1}$$

where w is each word in a message. Using the same method we can also extract the weighted arousal rating for each message.

The advantage of using the valence and arousal dataset from Warriner et al. (2013) is that it was specially created with the idea of being able to understand large scale sentiment through the characterization of attitudes and opinions by utilizing word co-occurrence measures. Specifically, the purpose of their work was to allow for research on the interplay of language and emotion, which fits the purpose of this project. We will be studying the sentiment and outrage of messages to see if it affects how users interact with one another on Twitter.

In Figure 1 below we have a plot of each message's scaled arousal and scaled valence scores. As was previously mentioned we are particularly interested in capturing the lower right hand corner of the plot, which has far fewer messages than the high valence, low arousal messages. This is interesting because it gives some indication that people are not always conversing in a fit of rage.

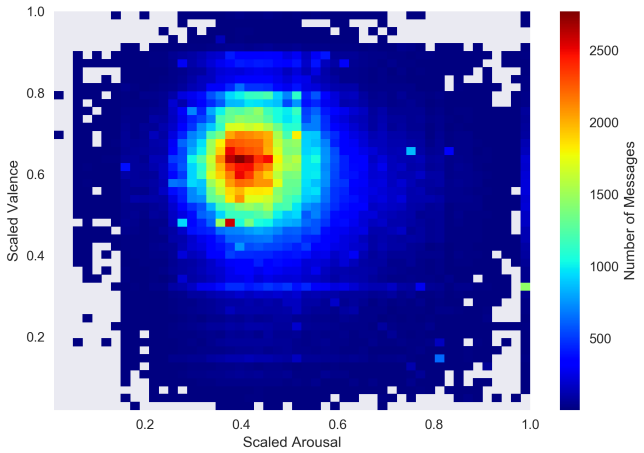

Figure 1: Valence vs. Arousal: In this plot we are observing the pairwise distribution of the scaled arousal and scaled valence measures for each message.

3.3 Topic Valence and Arousal

As was seen in the previous section, we know that the combination of valence and arousal can be a good indicator of outrage for our context when we minimize valence and maximize arousal. But we also have our topics; and, although they are coherent, we still want to understand how they relate to outrage. To solve this we have computed the weighted mean valence and arousal just as we did for messages, but in this case we have additionally weighted with the topic's word probability. If we did not do this, then all of our topics' valence and arousal scores would have been the same as each topic is a continuous probability distribution over all the words in the model's dictionary.

One downside to using every word in the distribution is that you end up with most topic scores, for both valence and arousal, hovering around the mean. Another factor in this is due to the low probabilities assigned to the vast majority of words in the topic distribution; hence, the sparse representation, so most of the score is relying on the words with highest probabilities. An additional complicating factor here is that the LDA model utilizes phrases in addition to unigrams, which can cause more skew as phrases would not have a valence or arousal score on their own.

Below in Figure 2 we can see the distribution of topic valence and topic arousal. As was previously mentioned, most topics fall into the middle of the range of valence and arousal scores with a few key exceptions. Topic 50 as was called out in Table 1 and in its preceding description was seemingly filled with outrage, and in this table we can see that it has the most arousal of any topic in the model with a far lower valence score (though not the lowest). In combining the findings in our topic table and our topic valence and topic arousal distributions it is easy to see that outrage is difficult to measure alone in terms of emotional sentiment as there are highly nuanced ideas going into what constitutes outrage. In the next sections we will explore more metrics that are explicitly related to outrage.

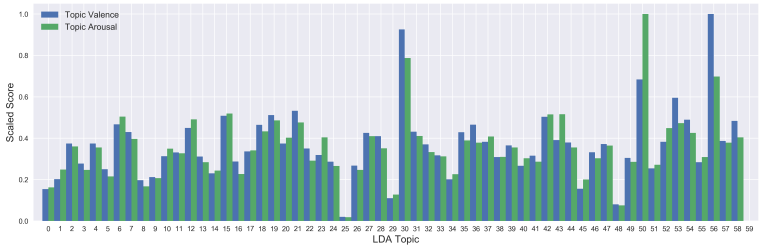

Figure 2: Topic Valence and Arousal: In this plot we are showing the scaled valence and scaled arousal score for each of the sixty topics from the LDA model.

3.4 Outrage Dictionary

In order to understand what language could be most pointedly associated with outrage our project advisor, Billy Brady, provided us with a hard-coded outrage dictionary. This dictionary contained 72 words which were considered to best encompass the language that would broadly signify outrage. For instance, words like corrupt, vicious, and offensive are words that are typically associated with an emotional response of outrage. Much like our messages we used only the stems of the words and removed any duplicates in this new dictionary leaving us with 60 outrage specific words. Then for each message we counted the number of occurrences of these words in the message to give us an overall count of the number of outrage words in a message.

3.4.1 Enhanced Outrage Dictionary

Whereas a human-coded outrage dictionary was an undoubtedly useful resource for our project, we wanted to leverage additional techniques to improve the dictionary and expand our set of words, so that we had a base outrage count and an enhanced outrage count. Again, we leveraged gensim to accomplish this task. This time we used their Python implementation of word2vec that uses pre-trained word and phrase vectors provided by Google. The pre-trained word vectors contain a 300-dimension vector space to represent three million words and phrases.

After downloading the pre-trained Google News corpus word vector model from the Google code word2vec archive [8], we looked for the words in the dataset that were most similar to our base outrage dictionary words. We found the five words in the pre-trained vectors that were closest to each of our outrage words using the cosine similarity distance metric between the word vectors. It is important to note that this is the most useful aspect of word2vec for this project: the ability to find similar words in an embedding space because of its local word associations. After compiling all of the words from our outrage dictionary and their five nearest word neighbors, we then stemmed all of the words and removed the duplicates to establish our enhanced outrage dictionary. This dictionary contained 255 words, over four times the size of our original set, making it far more likely to reliably detect outrage without requiring additional human coding, which would still likely not be exhaustive. For us, however, this is a good proxy for our task given our goal of finding as many indicators as possible in a narrow time frame.

The disadvantage of using the Google News pre-trained word vectors for our task is that these vectors were trained on news articles, not Twitter messages, which are only 140 characters in length (all data was extracted before Twitter increased the character limit in 2017) and frequently contain abbreviations of words or phrases that are unlikely to be present in the formal long-form text style of news articles. Similarly, Twitter users are often not as refined as published journalists, who write professionally and in a process with editorial feedback, and as such tend to use different discourse. In particular, there is a dominant presence of slang words, emojis, colloquialism, and other jargon. These would certainly be very useful to understand emotion; however, news articles preference concision and facts (or perceived facts) more than would be seen in user generated social media content. That said, given that our base outrage dictionary was created from words that could be seen in a more formal context, as opposed to Twitter's shorthand, we felt that the most similar words from a new article corpus would help maintain this level of formality within our study as opposed to the context-specific, nuanced nature of short form jargon. On top of maintaining formality, we are also ensuring reliability as jargon or slang terms can change meaning when placed in different contexts far more frequently than common words. It is therefore entirely more practical and meaningful, from our perspective, to focus on common words with more known or generally understood dictionary definitions, terms with more stable etymologies, over a sole focus on slang, which further validates our use of the Google News word vectors. Therefore, even with this level of expansion, we expected to be able to get a much more holistic view of outrage in messages.

3.5 Emoji Indicator

In addition to using words to express emotions on social media, users also use symbols such as emojis. Emojis are frequently used as a way to take one character space in Twitter and express some specific emotion that could otherwise only be represented through numerous entire words. Emojis are present in our data using a unicode-like encoding, for example, `<f0><U+009F><U+0099><U+0082>` is the text representation of a smiley face in a message. Given that there is no universal mapping

from unicode string representations to their visual representations as every company uses a different symbol (a common problem for messaging services), we manually searched our data and found three outrage related emojis and one happy emoji, corresponding to angry faces and happy faces respectively.

We used these tags to search through the text of every message to identify whether or not the text contained the text code for an outrage emoji or a happy emoji. This gave us a binary flag for both emoji outrage and happiness. In order to take action on these flags we created a net emotion emoji score that would take the outrage emoji flag less the happy emoji flag to give a -1, 0, or 1 classification of the emoji emotion in a message; 1 corresponds to a message with one or more outrage emojis and no happy emojis, 0 is a message with either both happy and outrage emojis or neither, and -1 is a message containing one or more happy emojis and no outrage emojis.

The three codings are important as a happy emoji should penalize the outrage score by containing positive indicators and an outrage emoji should boost the outrage score because we are affirming in another manner its level of negative sentiment.

3.6 Sentiment Modeling

One of the biggest challenges we faced in this project was that our data was unlabeled. While we wanted to measure the outrage of a message, we did not have any messages that had been labeled with their outrage score. Trying to classify a value without having the ability to leverage already known classifications is typically not encouraged, as it is often considered incredibly hard to validate that the model produces accurate results. One way to mitigate this issue is to use another dataset that might provide insight that we currently do not have with ours. Sentiment140 [2] is a resource for natural language processing that includes a dataset of 1.6 million messages labeled with a sentiment score of 0 (negative), 2 (neutral) or 4 (positive). Given that we have Twitter data, and that we believe a strongly negative sentiment score is a relatively good proxy for outrage, we built a model using this data and ran it on our dataset to get a sentiment score.

We took the 1.6 million message corpus and split it into a 25% test set and a 75% training set with an additional 500 messages provided by Sentiment140 as a validation set. We chose to use a Multinomial Naive Bayes approach with unigrams and bigrams, as we had initially believed the dataset contained three classes. We soon realized the training data only contained two classes (0 and 4, excluding neutral messages) while the provided validation set contained all three. We then tested with a Gaussian Naive Bayes approach, but did not get significantly different results, so we continued using our multinomial model. After fitting across various values of alpha we had the greatest accuracy with alpha = 1.0, namely the model performed best with a high smoothing parameter. The model has a 75.7% correct classification rate on our test set of 400,000 messages that are either labeled 0 or 4 while the model has a 59% correct classification rate on our validation set of 500 messages labeled 0, 2, or 4.

For our outrage score (discussed in Section 3.7) we focus on the probability of a message having negative sentiment, as opposed to a binary flag. By considering the probability of negative sentiment we do not have to set a threshold that would remove a lot of the information built into our model, and instead we were able to get a more reliable and informative measure of negative sentiment. Although outrage and negative sentiment are not exactly the same thing, we would argue that it is very unlikely that a message would have positive sentiment and express outrage, and we therefore believe this prediction of negative sentiment will give us strong signal in understanding the messages in our data that express outrage.

When we look at the probability of negative sentiment in our message dataset in Figure 3 in the Appendix we can observe an approximately normal distribution. We have a frequency spike at 0.5 as many of our messages are likely not captured by our training dataset and the model has no information as to the sentiment of the message. In this instance, the model naturally predicts a probability of 0.5, as it cannot attest to the positivity (0) or negativity (1) of the message so it assumes the middle point.

3.7 Outrage Score

After building several models and measuring outrage in different ways, we now want to build a comprehensive metric out of these scores. In data science, ensemble methods are generally more

robust to outliers and noise than any individual model and it is our belief that, although our scores and weights are not as rigorously constructed, our weight and combine method can also achieve this task. We have built eight different measures of outrage that aim to capture outrage itself, valence, arousal, negative sentiment, and the valence and arousal of the topic that a message most likely belongs in to capture different facets and nuances of outrage that each one individually cannot explain.

Whereas the measurements were discussed above, we will go into more detail into how they relate to each other and into our justification of weights. In order to use straightforward and interpretable weights, and get a sense of how each of the scores contribute to the overall outrage score, we scale all values to be between 0 and 1. With the exception of the sentiment model output (this is already a probability falling between 0 and 1) and the outrage and extended outrage word counts, we scale each variable using a min max scaler. As valence and arousal can be NA values if a message contains no words in our valence and arousal dictionary, after scaling, we coerce NA values to 0.5 as this would represent a neutral arousal and valence as it would be our best estimate for an unknown word. For topic arousal and variance we avoid this issue as we do not have a situation where a topic does not contain any words in our valence and arousal dataset. The topic valence and individual valence scaled between 0 and 1 are then subtracted from 1 as we want to look at low valence (low sentiment) as an indicator of outrage. Emoji after transformation appears as 0 for positive emojis present only, 0.5 for neither positive or negative emojis or both or 1 for outrage emojis only.

For outrage and expanded outrage word counts we scale from 0.5 to 1. If a message contains no words from our outrage dictionary, we code it as a 0.5 outrage word count, as we believe it should neither contribute to nor detract from a neutral score. Most messages have no outrage words and frequency decreases as word count increases (having a single outrage word is the second highest word count group). If a message has six words from our basic outrage dictionary is will scale to a 1, however if a message has three words we would prefer it scale to 0.75, as opposed to 0.5, as three outrage words is strong signal and should be represented accordingly.

We weighted topic valence and topic arousal highest, as we believe our topic model uncovers true groups of language that are cohesive and provide additional information about a message than message valence could. We weight message valence and message arousal marginally lower, as we believe each give strong insight into the overall language of a message, as it relates to the messages emotional response. We consider these four to be our strongest predictors as they are based on validated, rigorous work that was published by Warriner et al. (2013).

We weighted the basic outrage word count the same as message valence and arousal, as this measure is based on a user-created word dictionary that is likely as strongly vetted, but not exhaustive. We weighted the expanded outrage word count lower as it is possible that the words nearest to ours do not quite capture outrage and they were not human-validated. Sentiment fell between these two features as our sentiment model performed well on the test set and we believe that it captures negative sentiment in a message well. While the outrage dictionaries will capture words out of context, the sentiment model does a better job of considering the entire context of a message. We originally weighted emoji lowest as our list of emojis was in no way comprehensive, but we found low incidence of emojis (less than 1% of tweets in our corpus had any nonzero emoji score) which ultimately manifested by contributing nothing to our overall outrage score and rather narrowed our distribution artificially. By looking at the messages containing outrage emojis we did not get the sense they were truly more outraged than messages that did not include them.

Although we did weight features differently, we also made sure to not let any individual feature strongly dictate the score. Whereas we believe topic valence is more informative than our sentiment score, we do not think it should outweigh the effect of this feature if it is high.

Below we have outlined our weight method as it relates to each of the features. We also work through a sample message (see Figure 4 in the Appendix) that reads: "NRA is DEAD WRONG Good Guys Can't Kill Bad Guys if 50% of US GUN OWNERS Have NEVER FIRED Their GUN https://t.co/BtgJKZVze4." This is a message from our corpus with one of the highest outrage scores of 0.665. We see very high message arousal and topic arousal scores that are definitely in line with the message itself, while topic valence is significantly lower than message valence, which seems to represent the message well. There are few words that individually are strongly associated with outrage, and therefore the outrage and extended outrage word scores are somewhat low, though still above average. Lastly, the sentiment model predicted a reasonably high probability of negative sentiment in the message.

Feature	1 - Topic Valence	Topic Arousal	1 - Valence	Arousal	Outrage	Sentiment	Expanded Outrage	Emoji
Outrage Score Weight	0.16	0.16	0.15	0.14	0.14	0.13	0.12	0.00
Example Tweet Values	0.392	0.841	0.749	0.838	0.583	0.667	0.583	0.0

Table 2: Outrage Score Model: In this table we have our outrage score's elements and their respective weights as well as an example message's weights producing an outrage score of 0.55289.

When looking at the below Figure 3 of outrage score histograms grouped by topics some clear trends are visible. Same-sex marriage posts tend to be lowest on the outrage scale. When scanning the messages it seems that the data extraction for on-topic messages does not distinguish between same-sex marriage as a political issue and marriage as a topic in general. As such, people posting about weddings they are attending do appear in our corpus. This is most certainly biasing the results about how people talk about same-sex marriage, which tends to be more politically divisive. Climate change and gun control appear to cause a bit more outrage with gun control having some of the highest scoring outrage messages. As gun violence is a serious issue that seems to occur on a regular basis, gun control tends to prompt more anger between different ideological camps of those in favor of gun control and those opposed to an infringement on the Second Amendment. Climate change, on the other hand, tends to fall in the middle of the scale; we have noticed that many people are still climate change deniers and tend to share sarcastic messages and while non-deniers are posting in outrage over the way people ignore a scientifically proven issue that is causing problems around the world.

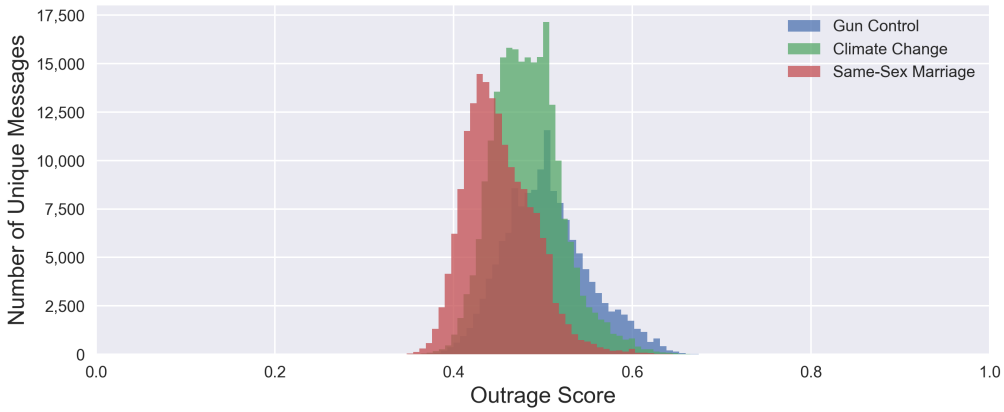

Figure 3: Outrage by Topic: This plot is a grouped histogram of each message's outrage score grouped by overarching topic.

4 Engagement Modeling

After developing our outrage score, we wanted to understand if higher engagement is associated with a post having a higher outrage score. It has been generally suggested that social media has moved towards a personalized news feed that pushes users into ideological echo chambers to the point that a user will follow other users that are like themselves based off of their activity. We believe this might create large networks of people who share beliefs and, thus, are more likely to react to and engage with an emotionally charges post relating to political topics that spark strong opinions.

As a baseline we tested out a LASSO and Ridge regression model with the purpose of understanding how our features and outrage score compared with the features that were already present in the data.

In particular LASSO is very interpretable as feature weights are shrunk to zero if they have little or no contribution to the model, and important features are given higher weights. One downside with LASSO is that it arbitrarily shrinks highly correlated feature weights, and this is where Ridge regression becomes our other baseline as it equally shrinks both highly correlated feature weights. This is important because we expect that our features are highly correlated as they help to validate each other. However, in our model, we were left with poor results when predicting engagement using number of followers, and all of our outrage metrics. In each model, even with hyperparameter tuning, we arrived at $R^2 < 0.2$ in each model. While it is expected when predicting human behavior that R^2 tends to be below 0.5, 0.2 did not seem to be a "good" fit for a model.

One advantage of using these two models is that we were able to see which features were the most important. When running with all features (and not just our end score) we found that topic valence and topic arousal consistently ranked in the top five features.

We also tried to model engagement with an extra trees regressor. An extra trees regressor is an ensemble method that fits several decision trees on random samples of the dataset and averages the results to create a model more robust to overfitting and more predictive than a single tree. Given we had a low number of predictors (two) we probably could have simply used a decision tree regressor; however, an extra trees regressor produces similar results. We chose to use both the outrage score and the log of a user's follower count (replacing $-\infty$ with 0) in our model as we believe that engagement will scale depending on the follower count. A user is much more likely to get a retweet if they are followed by more people. The model returns an R^2 of 0.09, implying that this model also does not accurately capture engagement and we are missing several variables that would contribute to a higher score. Additionally, when we look at feature importance, a user's follower count is nearly eight-times more predictive than the outrage score of a message (outrage score has importance .117 and follower count .883) signifying that even with our low predictive power, most of it is coming from follower count as opposed to outrage.

This is not particularly surprising as we do expect engagement to be strongly related to follower count. If we wanted to understand the true effect of outrage on a messages engagement in a ideal world, we would run a test where we have many users with similar follower count that tend to get the same number of retweets post both outrage-filled and outrage-less messages and monitor retweet rate (essentially an A/B test). Even in that world, however, we would not be able to capture the inherent entropy of social media; engagement with a post can depend both on the language of the post itself as well as the time of day or where the post ends up on a user's homepage.

With that in mind, we wanted to see just how much, if at all, engagement correlates with outrage when we are looking at our large sample of messages. When we again look at a message's outrage score and the poster's following, we get a high degree of correlation. The interaction of outrage and follower count has a Pearson's correlation coefficient of 0.235 with engagement. This points to a weak but nontrivial link between the interaction of these features and engagement, however we want to understand how significant a contribution the outrage score makes in this interaction. When we look at the correlation between followers and engagement we unsurprisingly see high correlation with a statistically significant Pearson's correlation coefficient of 0.227 implying a substantial amount of the correlation in the interaction is likely attributable to follower count.

The correlation between our outrage score and engagement is only 0.048. While the relationship is positive, it is still very weak. Ultimately it looks like engagement and outrage might not have a significant relationship, though given engagement has such a long tail (even when log transformed) it is challenging to capture its distribution in a meaningful way. For instance, if we were to build a model to predict engagement based on any number of predictors, the model could predict zero engagement and be correct a majority of the time. When we look at a plot of outrage score against (see Figure 6 in the Appendix) we see very clear bands of messages that fall into buckets of 0, 1, 2, 3, 4, or 5 retweets.

Given that we log transformed the data and chose to map zero retweets to log(retweets)=0, we are collapsing one retweet and zero retweets both to take on a value of 0. As zero and one retweet are different, we also looked into the correlation when we exclude messages with no engagement. In other words, we drop messages without retweets and look only at one retweet or more. The coefficient of correlation of the interaction between followers and outrage score increases marginally to 0.268 and the correlation of outrage score alone with engagement is still very low at 0.060. While this is

slightly higher, it still does not convince us that there is a nontrivial relationship between engagement and the level of outrage present in a message.

5 Results and Conclusions

We embarked on this project hoping to accomplish three tasks: accurately model the level of engagement in a message, understand the relationship between social engagement and outrage, and make our code and models open source and reproducible. We have successfully accomplished all three: we developed an outrage index score which combines insights from several different types of models and features; we studied the correlations between outrage and engagement and looked at predictive models considering both outrage as well as a user's following; and we created a GitHub repository that houses all of our code and has a class that will digest new messages and produce their outrage scores using pickled models, vectorizers, and scalers.

In building an outrage score, we wanted to look at several features to produce an index that would consider a number of aspects of text to provide a holistic approach to "outrage language." Our LDA model produces cohesive topics that seem to capture and reflect distinct groups of language and emotion. We considered the outrage word count as had been done in Brady et al. (2017) with moral and emotional language, but also added an expanded dictionary through word2vec that takes into account words that are similar to our outrage words that we fail to capture in our dictionary. We were able to leverage freely available and sentiment labeled Twitter datasets online to build a Naive Bayes classifier to predict the probability a message in our dataset has negative sentiment. We also considered valence and arousal, at both the individual message level and at the LDA topic level, as we know that messages in the same topic are likely to share similar language and emotion. This allowed us to measure the amount of strongly negative sentiment that a message contains. Combined, each of our methods offered its own validation of one or more of the other metrics. Emojis, when present, validated sentiment, valence and arousal circuitously validated sentiment, topic valence and topic arousal were validated by valence and arousal, outrage counts were similarly validated through valence and arousal, and more.

Whereas each of the metrics could offer validation for another metric, validating our outrage score was not an easy task as we don't actually know the true values of outrage in a message because there was previously no existing metric. We were able to look at the messages associated with the top and bottom 50 outrage scores to gauge how well our model predicts extreme levels of outrage on both ends. Out of the 50 lowest outrage scores, only one was misclassified and this was due to the language being sarcastic. Sarcasm is sometimes challenging for humans to detect, and is still an open challenge in NLP, so a message that reads "Thank you #climatechange" with a URL does not imply low sentiment. In reviewing the top 50 outrage scores we found that all had negative sentiment, though some were expressing unhappiness as opposed to outrage, which is expected given the array of emotions valence and arousal represent. Most of the top 50 are expressing outrage, though roughly a quarter of the messages are simply expressing sadness or solemness. In addition to validating the extreme values we also reviewed a random sample in the middle of our distribution to see if we were missing any obvious outrage messages. From looking at 50 messages in the middle of the pack with regards to outrage score, none exhibited outrage language; and from our initial validation we feel quite confident that our outrage meter does successfully measure outrage.

Even though small batch validation shows strong signs that we have an accurate index to model outrage, this is no substitute for rigorous fact and label based validation. Outrage is not defined rigorously as emotions exist on a scale and are not easy to quantify, so a next step in the validation process is to crowd-source our validation to ensure accuracy. Using human coders to gauge the level of outrage in a message or to confirm or deny our outrage scores for a set of messages would be time consuming and costly, but worthwhile, if we want to guarantee a high degree of confidence in our metric. As long as our data is not labeled accurately and our score is not validated, we are likely to miss instances of outrage in our data as we do not have a model of outrage to base our score on.

After developing our outrage score, we looked to understand how outrage relates to engagement. Using several techniques ranging from regression models to basic correlation computations, we saw minimal relationship between outrage and engagement. When we include a user's following we get a much higher correlation, but it appears that the correlation is mostly dictated by the follower count and not as much the level of outrage. When we look at users over 22,000 followers (just over 6%

of our corpus) we see higher correlation with a coefficient just over 0.16. It seems if we are able to better control for user popularity we might see a stronger correlation. This suggests that a relationship exists but is not particularly straightforward. It would be interesting to look into individual users engagement by outrage score but this would require significantly more data We could look to do propensity score matching on ideology, average engagement and follower count as well, however we did not get the opportunity to focus on these techniques.

In addition to creating an outrage score and quantifying the relationship between outrage and the transmission of messages on Twitter we have made our work open source, and reproducible. All code being available at `https://github.com/NYU-CDS-Capstone-Project/NotablyLoftyPotential` with the exception of the data and LDA model. Those are too large to be uploaded to GitHub, but can be available upon request. The goal of reproducibility and the open source nature of the work is in the same light as our advisor's own work. The theme of open source and reproducibility is one that is finally gaining traction within the sciences to ensure the validity of studies being done. In addition, this work is going to be continued after this course with the goal of being published will be needing to fit within the framework of our advisor's own work, and so we sought to do this throughout the entire process. In addition to making our models, data, and processes open source, we have also built a class in Python that allows a user to access the model and classify new messages, or even add more corpora to the model to fit the model to new topics.

In the process of making our work openly available it has ensured that our own methods are valid because of the potential for increased scrutiny, and has forced our hand in keeping our processes concise and (mostly) easy to follow.

At the end of our Capstone we were satisfied that we improved upon the work our advisor had completed by providing a more robust measurement of outrage. While the techniques we used in developing this score range from rather simple to relatively complex, the outrage score is easy to understand and very flexible. Weights and models can be changed over time to capture important trends and by making our code open source we ensure that this is an approachable task for even a non-technical user.

6 Discussion

Our work to date, while accomplishing quite a lot, still has areas for further exploration. As the work progressed we found limitations to each of the metrics that we were creating, but had little time to create a solution for each of these issues.

A major limitation to topic modeling is that LDA is a very slow model to run, even when parallelized, as the speed of reading of the data cannot be improved. This makes it hard to ensure that a model has fully converged, and that it is in fact better than any other model that you are testing against. Further, when parallelizing the data it is not possible to optimize α as it must be symmetric. Additionally with our topic modeling, it would have been preferable to also model lemmas, bigrams, and trigrams to compare our model performance, but again training an LDA model takes far too long to feasibly test every scenario.

When measuring the valence and arousal of messages we found that the abbreviated nature of discourse on social media caused us to neglect a large number of words that would have been in the dictionary. Similarly, because we used phrases we were not able to assess the meanings of the words that would have been combined in a straightforward manner. In an ideal world we would have used the mean scoring method from 3.2 to get the mean and standard deviation of the phrase. Fortunately, this scenario is only relevant to our topic valence and topic arousal scores.

As was previously mentioned mining for emojis is a nontrivial task as it requires manual validation and ultimately offered trivial value from our own methods. In the future it would be useful to create a mapping for each of the emojis in the dataset and create a positive/negative weight for each.

Another difficulty in our project was the engagement modeling. Given the high skew in some of our features and the small feature space it was quite difficult to create a well performing and/or generalizable model. In the future we would consider more than just machine learning techniques for prediction, and look to study main effects and other statistical measures as was done in Brady et al. (2017). As we look to the future to refine our project and continue this work it is highly likely that we will be using more classical statistics and spend more time on feature engineering than we had

originally done. It is entirely possible that we will see a stronger connection between outrage and engagement when we control for other factors as we briefly alluded to earlier.

Further, our outrage score could certainly use some refinement. We would like to explore considering interactions between features to obtain a better understanding outrage. If a message has a very high probability of negative sentiment and several outrage words, these predict a strong likelihood that a message has a significant level of outrage. Looking at the intersection of these features might give us more information that looking at either feature separately. The same can be said about valence and arousal; sitting at the intersection of high arousal and low valence are messages that are very probably filled with outrage and looking at this region might give us a strong signal. While this might complicate scaling our features it provides us with an avenue of interesting discussion.

In addition to our difficulties and limitations, we had additional ideas on how to further explore the data that we did not get an opportunity to implement during the course of our Capstone. For example, we believe people tend to use uppercase letters when they are having a particularly strong emotional reaction; therefore, if we look at the percent of characters that are uppercase (excluding URLs and emojis) from this count we could capture this emotional response. We also would like to do more in the ways of data cleaning when it comes to tokenization by removing URLs and references to another user (e.g. tagging or '@'ing users) could help to clean up our text while looking closer at 'hashtags' might give us stronger context. If we are able to use hashtags to further our understanding of what a message is referring to we might be able to better understand the message contextually. If a message is discussing gun control and adds a "#sanbernardino" in the text, our sentiment model and valence and arousal have very low chance of giving any sort of signal that this is likely to be an emotional response to a horrific gun-related incident. While our LDA model does a good job of grouping together hashtags, we can do a bit more mining into these topics ourselves to enhance our model.

With regards to engagement, we would like to look into ideology. Our advisor provided ideology scores (generated by SMaPP [1]) of the users associated with messages in our dataset, and we would like to explore engagement as it relates to user ideologies. The type of engagement that users are getting might be more digestible if we consider the engaging users' ideologies. Given that we are dealing with politically motivated topics, if a strong conservative is getting a lot of engagement from strong liberals, it is probably a much different type of engagement than if strong conservatives are engaging. For example, it may be suggested that while two users with similar ideologies will likely engage to show support or agreement, it is more likely to be a mocking or disagreeing response when two ideologies clash. While these limitations did exist due to what was feasible in the scope of this project, they are more ideas for future iterations, as the work in its current state successfully accomplished the tasks that we set out to achieve.

Acknowledgments

We would like to thank our Capstone course advisor Prof. Michael Gill and our project advisor Billy Brady for their support and contributions to our work.

References

[1] Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). messaging from left to right: Is online political communication more than an echo chamber?. *Psychological science*, **26** (10), 1531-1542.

[2] Bhayani, Richa, Go, Alec, & Huang, Lei (2013) *Sentiment140- A Twitter Sentiment Analysis Tool* http://www.sentiment140.com/

[3] Bird, S, Loper E. & Klein E. (2009) *Natural Language Processing with Python: analyzing text with the natural language toolkit* "O'Reilly Media Inc.".

[4] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003) Latent dirichlet allocation *Journal of machine Learning research* **3** (Jan):993-1022.

[5] Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017) Emotion shapes the diffusion of moralized content in social networks *Proceedings of the National Academy of Sciences* **114** (28):7313-7318.

[6] *OSF | Brady Et Al. (2017) PNAS Moral Contagion Wiki* osf.io/59uyz/wiki/home/

[7] Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., & Cohen, W. W. (2016) message2vec: Character-based distributed representations for social media *arXiv preprint arXiv:1605.03481*.

[8] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013) Distributed representations of words and phrases and their compositionality In *Advances in neural information processing systems* (pp. 3111-3119).

[9] Moody, C. E. (2016). Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec *arXiv preprint arXiv:1605.02019*.

[10] Rehurek, R., & Sojka, P. (2010) Software framework for topic modeling with large corpora In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

[11] Röder, M., Both, A., & Hinneburg, A. (2015, February) *Exploring the space of topic coherence measures* In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408) ACM.

[12] Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013) Norms of valence, arousal, and dominance for 13,915 English lemmas *Behavior research methods*, **45**(4), 1191-1207.

Appendix

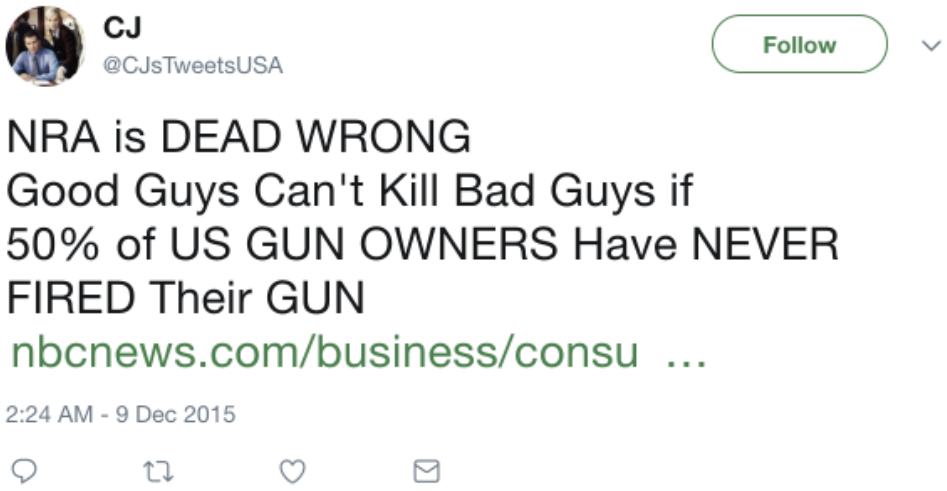

Figure 4: Sample message in our dataset.

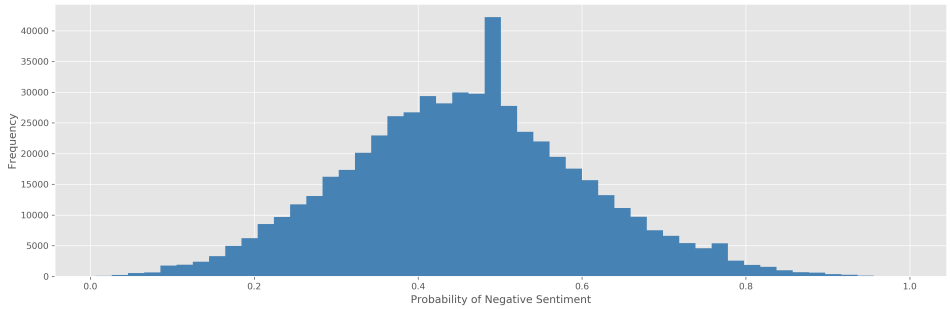

Figure 5: Histogram of sentiment scores for our corpus.

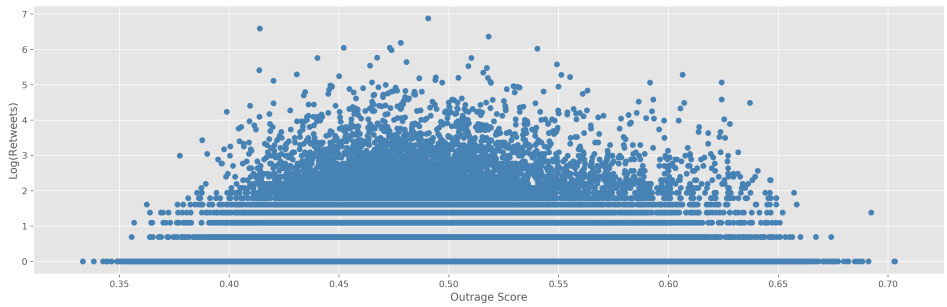

Figure 6: Scatterplot of outrage scores against log(retweets) for our corpus.