Social Engagement with Morally and Emotionally Framed Messages

Original Proposal

What is the project?

We are measuring how the presence of moral and emotional framing of political messages affects their retweet and "like" (engagement) rates in Twitter and Facebook social networks. We also test how these variables affect liberals and conservative users differently (we use ideology estimation techniques from the NYU SMAPP lab). We have a large amount of Twitter data on specific moral/political topics (e.g. gun control) and have Facebook data from pages with posts and corresponding comments, where we are attempting to replicate our findings from Twitter.

We would like help with developing new methods for measuring morality and emotion in Twitter messages (language analysis techniques). Right now, we are using a simple top-down dictionary method which is a bit limited (a "bag of words" technique where we simply count if specific words from a dictionary appear in the message). We think natural language processing techniques could help us improve the precision of our current measures of morality and emotion. For example, we've seen one group measure morality in Twitter using latent semantic analysis techniques where they use an existing dictionary but populated it with words that covary (in the Twitter messages) with the dictionary words. This is one example of something we're interested in, but are open to other techniques that could enrich our current measures. One broader goal is to create a method for measuring discrete moral emotions (e.g. anger, disgust, outrage) that could be helpful to our own project as well as other psychologists and political scientists in the field.

Although not our main priority, we could also use help with time-series analytic models. For example, we are interested in how the effect of moral and emotional language on retweet rates changes over time depending on the specific political topic (e.g. gun control). We have some experience with these models, but are always looking to improve.

What does the data look like?

We have already collected a lot of data which consists of Twitter data stacked such that each row represents one message. We have the message text, relevant meta-data for each message (e.g. author, messageID, retweetID, followers, retweet count), and some of our own dictionary count variables we have formed.

We also have Facebook data from Clinton and Trump pages leading up to the 2016 presidential election. These data are nested such that there is a original post with full test (with an id) as 1 row, and then there are corresponding comments with full text (with ids) in the following rows.

For Twitter data, we have collected 500k messages, and the facebook data we have nearly the same amount of post + comments. Thus, we will provide the majority of the data, and the student would be forming models with that existing data. We have the ability to collect new data if the student may need something more specific.

What is the proposed scope of the projects?

We assume the project could be completed with 1-2 students. They would work with a small team of 1-2 psychologists who have been working with social media data and this particular project for 2 years. We consider the problem a small, well-delimited problem, but its solution could also have broader implications for the field if we can make it open-source and available to other researchers who are becoming interested in testing psychological hypotheses related to morality and emotion in online social networks. We anticipate that the project could be done with a weekly or biweekly meeting with the small team, and anticipate that it would be between 6-10 hours per week depending on the current week and current difficulties faced.

What are the rubrics of success?

Our main goal is to produce new count variables of broad categories of emotion (moral emotion vs. non-moral emotion), and also new count variables for discrete emotion concepts (e.g. disgust, anger, outrage). The mark of success would be the development of a program that could read in social media data messages, and output new count variables. These new count variables would go into analytic models that could test how these new variables compare to our old variables (in other words, do our conclusions change, remain the same?). We would also perform face-validity checks to see if the new language variables classify messages with better precision than our old methods, as accessed by human coders viewing a random sample of tweets/classifications. As a final goal, we would like the code to be accessible enough for other researchers to be able to use if needed for their own models with social media data. We would prefer the code be in Python or R, because they are these programming languages are the most widely used in our field (R is most used).

What is the relevant background needed with respect to the project?

To see the type of project and methods we are aiming to improve, the best place would be to skim our recent publication:
Brady, Wills, Tucker, Jost & Van Bavel (2017) Emotion shapes the diffusion of moralized content in social networks. PNAS
It provides a background of the theory and also showcases our current methods and data.

What is the relevant organizational, project or institutional history?

The Van Bavel lab is a psychology lab that studies how morality and group identity (e.g. political ideology) shape cognition and behavior. For more information, please see our lab website:
http://www.psych.nyu.edu/vanbavel/lab/

Our research is of broad interest scientists and the public. It has been disseminated in scientific journals with a broad audience such as PNAS, nature neuroscience, and cognition. We also disseminate our work to the public via NY Times op-eds and other news outlets.

How will the organization support and mentor the students?

The students will primarily work with Dr. Jay Van Bavel, and his two senior graduate students William Brady and Julian Wills who are in the last year of their PhD. Dr. Van Bavel will be present at bi-weekly meetings, and William Brady and Julian Wills will be available for weekly meetings. William and Julian have been working with social media data for 2 years, are proficient in R and data analysis involving multi-level models and count data, and have produced multiple successful projects with previous data. However, we do not have training in machine learning or natural language processing. William is going on to do post-doctoral research in Fall 2018, and Julian is transitioning to work with Facebook's news team in Fall 2018 (and has already completed a summer internship). Thus, together we hope we can help provide a template for students of best practices in terms of what academic data output should look like, as well as a comparison to what industry data jobs like Facebook look for. As it turns out, we find these two goals have a lot of overlap. This project will also turn into a tangible research product, for which the student will receive authorship. This could be useful for resume purposes, as it showcases a complete and written-up project. Julian also found it useful for industry jobs interested in analyzing trends in social media.

Research Question

For our project we are working with the Jay Van Bavel Lab in the NYU Psychology department to look at engagement with morally and emotionally charged messages on Twitter. The main goal for this project is to use the data science toolbox to enhance the current work modeling moral and emotional messaging on Twitter. Once we are able to derive additional meaning from various NLP methods we will use that information to understand how the emotional messaging can impact engagement on Twitter within and outside of one's own ideology. For instance, does a highly emotional tweet prompt more retweets and likes, or does it get cast to the side? And, who is engaging with the tweet in the first place: do disconnected ideological networks exist within this same framework?

Data

The raw data is just over one million tweets with about 330K gun control tweets, 500K tweets on climate change and 290K tweets on same-sex marriage. In order to only consider unique tweets we have aggregated all retweets by counting retweets for each original tweet to get a measure of engagement. This reduces the data just over 500K unique tweets: 145K gun control tweets, 177K same-sex marriage tweets, and 231K climate change tweets. The data spans from late October 2015 to mid December 2015 and have the SMAPP ideology rating at the time of extraction for the twitter user appended to each entry.

General engagement is very useful, but it does not provide context for the types of engagement that exist. We surmise that there are a few types of engagement: supporting, divergent (both same ideology and opposite ideology are engaged), and disagreement. In order to separate the potential types of engagement we have counted the number of retweets by ideology 'bins'. The bins were generated by consulting a histogram of the ideology of users and establishing cutoffs along the scale. Ideologies range from -1.5 to 2.7 where 0 is non-partisan, negative numbers lean liberal and positive numbers lean conservative. The bins are defined across strong liberal (<-0.7), moderate liberal (-0.7 to -0.1), nonpartisan (-0.1 to 0.1), moderate conservative (0.1 to 0.7) and strong conservative (>0.7). Once we begin looking at engagement we hope to develop a better understanding of users are engaging with each other, i.e., does a person with opposing ideology retweet your tweet out of contempt or agreement with a particular point.

Because Twitter data is unstructured text with only a maximum length of 140 characters we have to use a tokenizer that is capable of understanding the type of abbreviated and often syntactically void text that is present. For example a tweet full of hashtags is not a sentence, but is likely contributing to a larger conversation across Twitter. After much research we arrived at the NLTK TweetTokenizer as it has incorporated up to date research on tweet structure including emojis/emoticons, smiley faces, hyperlinks, etc. To ensure consistency we then force all text to the correct utf-8 format to allow us to use NLTK's PorterStemmer to identify stemmed words and not be overly concerned with word morphologies. Lastly we applied a term frequency -- inverse document frequency (tf-idf) transform on the tokenized and stemmed field to capture the term-document ratios.

In addition to stemming and transforming the text field, we also mined emotion out of emoticons present in tweets. Emoticons are represented in a format that was not easily discerned so manual checking was done to determine two 'positive' emoticons (smiling face and thumbs up) as well as four 'outrage' emoticons (angry, crying and frustrated faces). The number of each type of emoticon in each tweet was then added up and netted out to generate a relative emotion/outrage score based on emoticon alone.

After testing our latent semantic indexing and latent dirichlet allocation we found that punctuation had a massive influence in most of our underlying topics. In order to control for the undeserved importance of commas, periods, and other extraneous characters we decided to remove all standalone characters except for the following from the data set: !, ?, @, #, &, and '.

Methodology

Our goal is to first build a model to estimate the level of outrage present in a tweet. As we do not have labeled data to work with we will be using unsupervised learning techniques. We will be building a series of outrage scores which will be weighted to build an index that should give a high level understanding of the level of emotional outrage in a tweet. The index will be comprised of the several scores of varying levels of statistical robustness.

First we are planning to build a model on Stanford's twitter sentiment data (labeled positive versus negative sentiment tweets). We are hoping that by developing a general model on labeled data and running our tweets through the model, we can understand directionally the overall sentiment of a tweet. This will not be specific to the political and moral context in our data, but it can be used to estimate the presence of positive or negative emotion in a tweet.

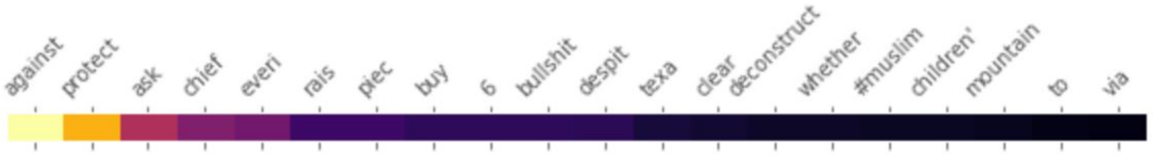
Our second goal is to uncover any latent topics present within tweets. The purpose of the topic modeling is twofold: we want to be able to understand how tweets group together to understand if emotional responses exist within tweets and and we want to use the predicted topic as a feature for a twitter engagement model. To uncover the latent topics we are exploring two models in particular: latent semantic indexing (LSI) and latent dirichlet allocation (LDA). LSI is a probabilistic variant of latent semantic analysis (LSA), a traditional model for understanding the underlying meaning of documents by mapping the words and documents into a concept space and comparing them. LSA, in essence, is a tf-idf model, which can have tens of thousands of features. In order to create a more informative and smaller final model truncated SVD is used to reduce the dimensionality to 100 to 300 topics. Applying the truncated SVD is what separates LSI from LSA. LDA then takes the model one step further and assumes a dirichlet prior to create a Bayesian probabilistic model. In general LDA tends to be more accurate, but comes with major losses in efficiency.

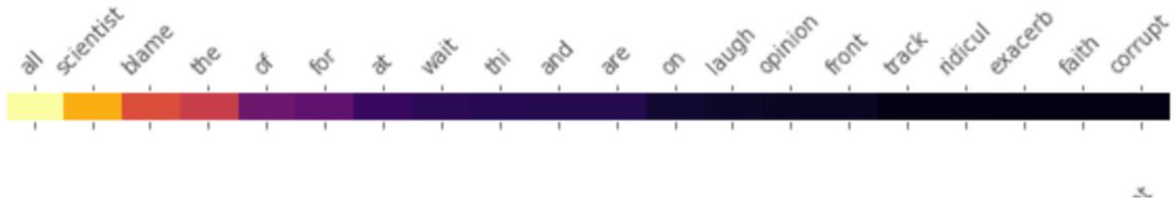
Lastly, we will be using a basic outrage dictionary to detect the presence and counts of words associated with outrage. As this will only detect presence of words and not necessarily the context they are used in, considering the counts in conjunction with our overall sentiment model should get us a reasonable estimate of outrage level.

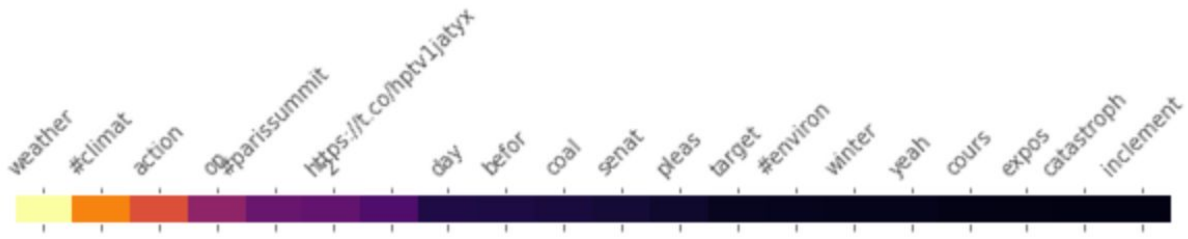
Results

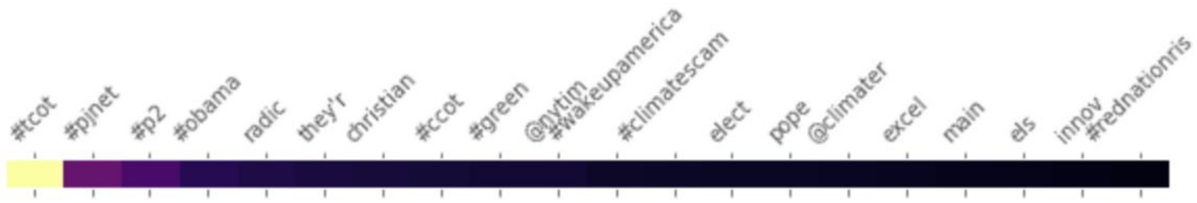
In our initial test runs our LSI and LDA models tended to place a large amount of weight on periods and commas, so we had to go back to the data to remove those from the data and rerun as was mentioned in the Data section. We then found that there were plenty of other standalone characters that would skew the data in place of commas and periods, so we decided to remove all the characters except for the ones mentioned in the Data section.

Our latest iteration has provided some promising topics that seem to account for outrage as can be seen below. The topic words are ordered from most important to least important with their associated topic probability mapped to matplotlibs inferno colormap, so black represents near zero probability, and yellow is the maximum probability for that topic.









Discussion

We are currently rerunning our LSI and LDA models to learn more about the topics that show up in the overall model and to further refine the types of words that are being deemed important to the model. For instance we will remove stop words from the model so that we do not have such heavy weighting being applied to objectively meaningless words.

We have also found that a lot of our topics also correspond to one of the three categories that are being considered for our data set. While this is useful to know that the models can differentiate these categories, it is not useful when our focus is to uncover moral and emotional messaging within tweets. Therefore, we will be further refining our models to run on each category separately to understand the emotional framing within our three categories. We likely additionally refine our filtered words as there seem to be a number of codes and other unimportant features such “https” or similar items. After we get a more refined model we will work to improve our visualizations to better understand the results and to find which words are the most important for understanding emotional responses as well for later engagement modeling.

Once we have a refined topic model we will need to determine an overall measure of outrage, which is going to be a challenge considering that our data is unlabeled. To build an outrage index off of our various scores and model predictions we will have to determine the relative importance of each, scale the scores and use them in a clever way. This will require us to look at the data and make decisions off of our interpretation of the more reliable scores.