Project #1:
Twitter Sentiment Analysis
Final Report

Problem Statement/Challenge

The questions to be answered/addressed in the project are, "Do Americans support U.S. President Donald Trump's position on immigration and, specifically, what many cite as the separation of families as they try to enter/cross the U.S. border? And do those opinions vary by geographic region?" Answers to these questions would be very useful and relevant to political allies and spokespeople—as well as opponents and critics—of the President. In addition to increasing our understanding of how Twitter data provides reflections of these sentiments across the country, this information could be used to shape advertising and marketing campaigns, political platforms, policy formulation, etc.

Further, this model and associated analytic techniques could be used to predict how states will react to news, announcements, and potentially controversial topics. To extend the reach and scope of the current model, one could simply replace the current query string with any topic(s) of interest and immediately gain a sense for the public sentiment (support or opposition) surrounding the topic.

Project Details, Description of the Data, and Data Wrangling

To address these questions, I used Twitter data obtained directly from the standard Twitter API. My initial goal was to gather all of the tweets on this topic, perform various analyses to find patterns and identify key variables and statistics, demonstrate how Americans

feel about the President's position on immigration and to determine if those sentiments varied by region or state.

To acquire the data, I first needed to authorize a Twitter API client. As I explored the Twitter API, I realized that standard application access would only allow me to see (and query) tweets for the last seven days. Therefore, I changed topics from my initial proposal. As opposed to focusing on an event that happened a few weeks ago (i.e., the summit with North Korean leader Kim Jong-Un as initially proposed), I focused on a current topic such as the aforementioned presidential position on immigration and how to deal with families as they try to enter the United States. This felt like a minor change and the code (and subsequent analyses) allows for the simple insertion of any topic. If I were to pay for the Premium Twitter API, I would have been able to access tweets for at least 30 days.

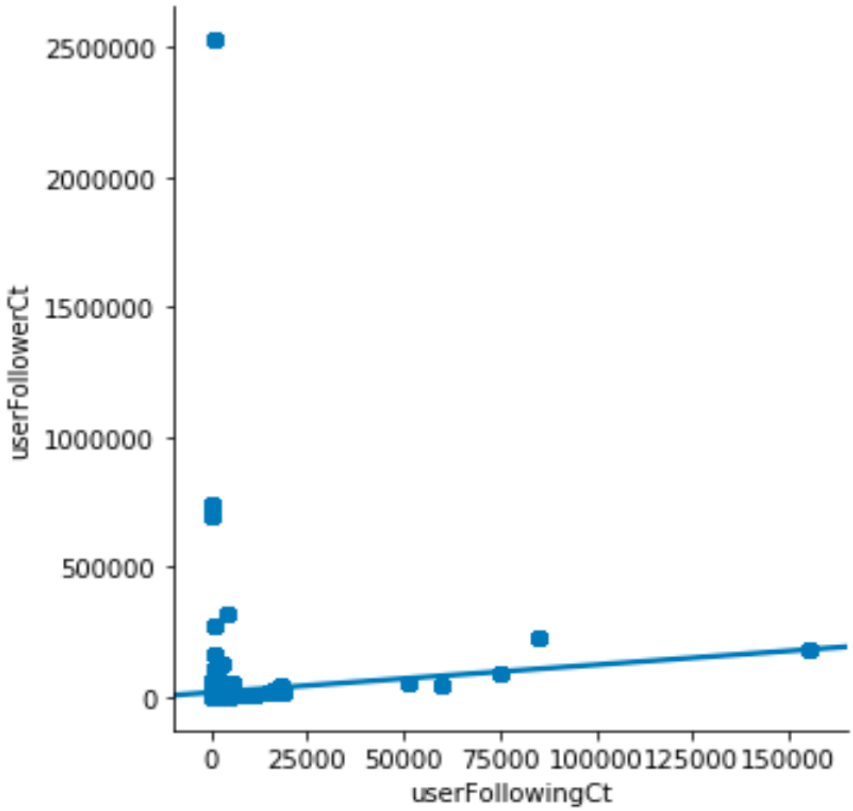
Next, I learned that specific location information is not typically provided/enabled by the average Twitter user. For the desired degree of fidelity—specifically geographic coordinates—to be available, a Twitter user would have to have to actively and purposefully enabled those services. Therefore, I was forced to rely on one object in the geolocation metadata called 'place_id'. This antiquated process of relying on the 'place-id' has since been updated and upgraded in the Premium API and would have allowed for a much more geographically precise query strategy. After considering the costs associated with the Premium API and consulting with my Springboard mentor, I decided not to subscribe for the Premium API, but it is clear that if I were to pursue future Twitter analyses or projects, I would definitely make the investment.

The next step was to establish a list of the top 100 cities in the United States (based on population) and building a dictionary to establish the state and region for each city. I then queried Twitter API for the place_id codes for each city and used those place_id codes to build the query for my terms of interest. Again, for this project, I focused on terms such as "immigration" or "family" or "separation" combine with a variety of 'President Trump' terms.

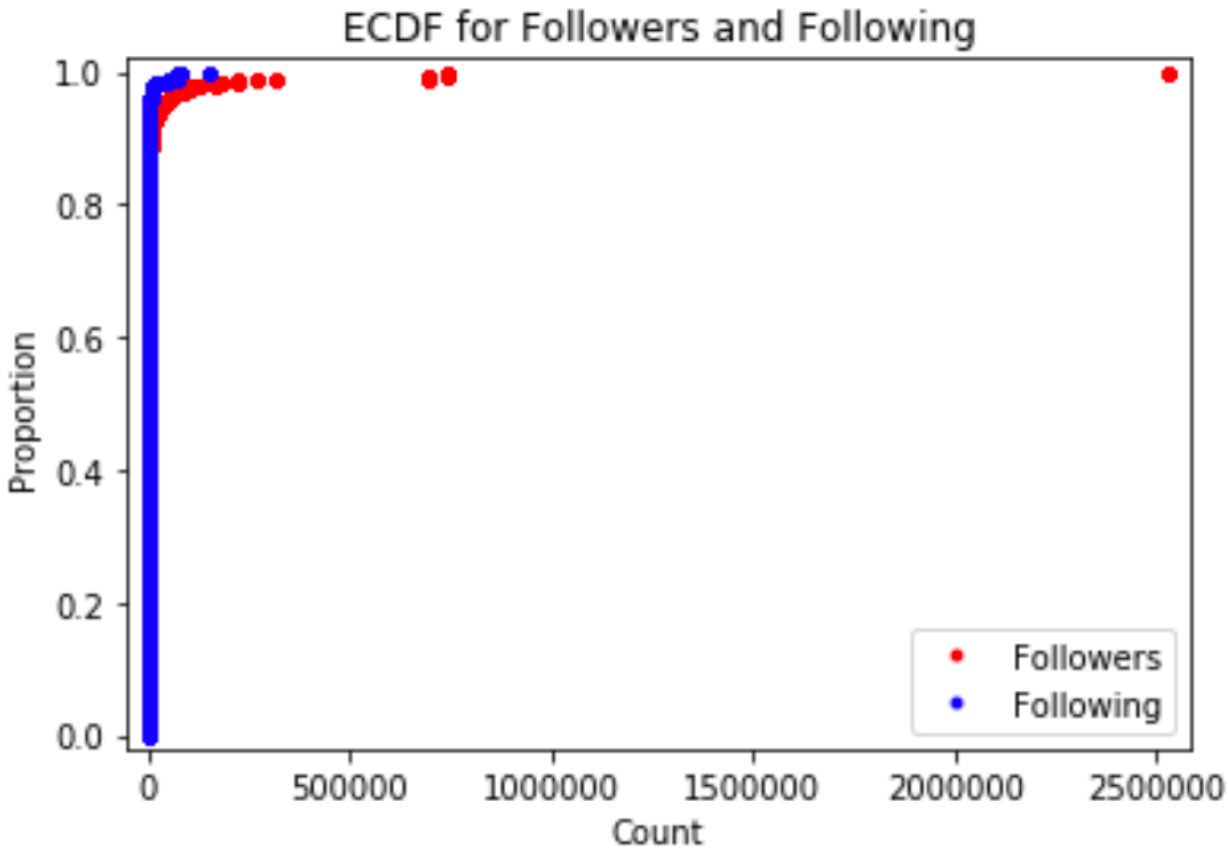
Fortunately, after overcoming this hurdle, the data was relatively clean and required little to no additional wrangling. Finally, as discussed later in this paper, I created a 'sentiment' column with the results of my sentiment analysis that labeled each tweet as "positive", "negative", or "neutral."

Exploratory Data Analysis

The final data set had 49,000 rows and 21 columns. The columns consisted of features such as tweet ID, tweet text, number of times the tweet was retweeted, source of the tweet (e.g., iPhone, Android, etc.), the date the tweet was created, User ID, user description, number of followers, and the number of other Twitter users the sender is currently following. As you can see from the below chart, the number of Twitter followers is positively correlated with the number being followed (and outliers are very clearly identified):



I created an empirical cumulative distribution function graph to see that the distribution of the number of user's followers is very similar to the distribution of the number of people they are following. The following scatterplot reveals this general relationship, subdivided by sentiment:



Additionally, I found that very few of the variables were correlated with each other. The retweet count and number of times a tweet was favorited/like, for example, were significantly positively correlated with a correlation coefficient of 0.98.

Machine Learning and Further Analyses

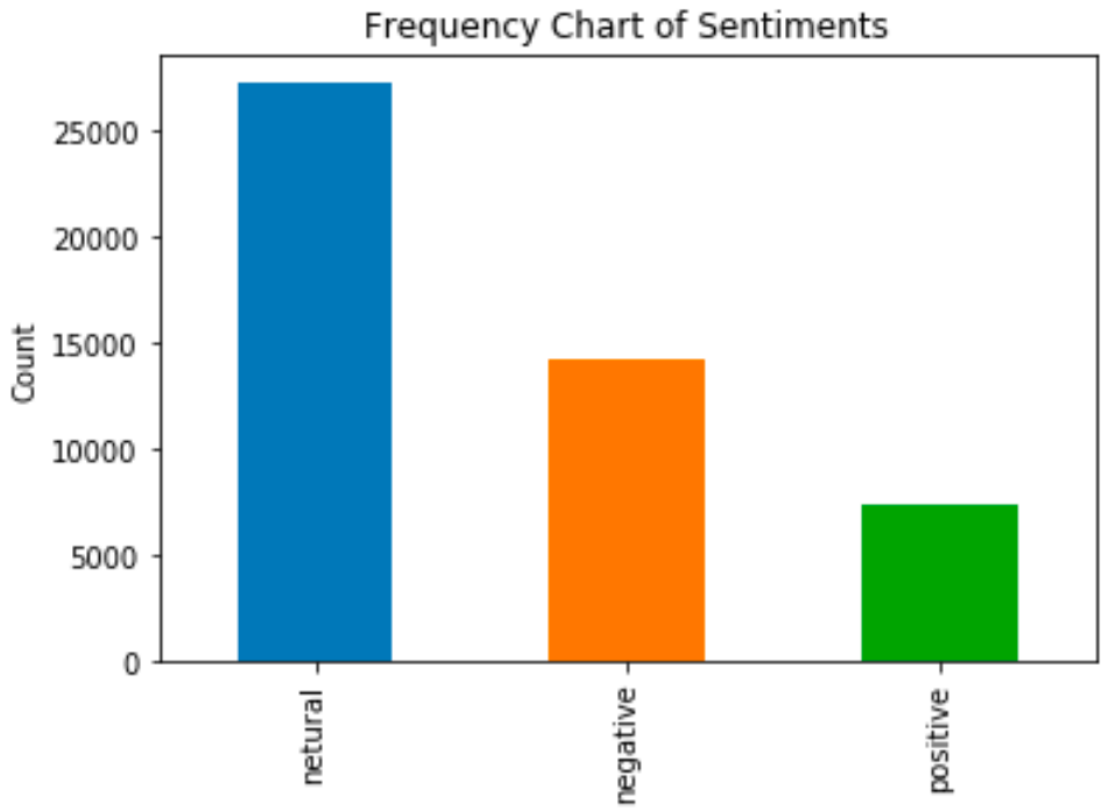
In close consultation with my Springboard mentor, I decided that most of the machine learning techniques presented in the curriculum would not be appropriate for my project. While I look forward to using linear and logistic regression models, Support Vector Machines

and Decision Trees in the future (and hopefully in my second capstone project), I only needed Bayesian techniques for text analysis; specifically, sentiment analysis/classification. Before doing this, however, I needed to utilize regular expression to preprocess the text portion of each collected tweet.

The documentation for *textblob* describes it as providing "a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more." Further the sentiment property "returns a named tuple of the form sentiment (polarity, subjectivity). The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective."

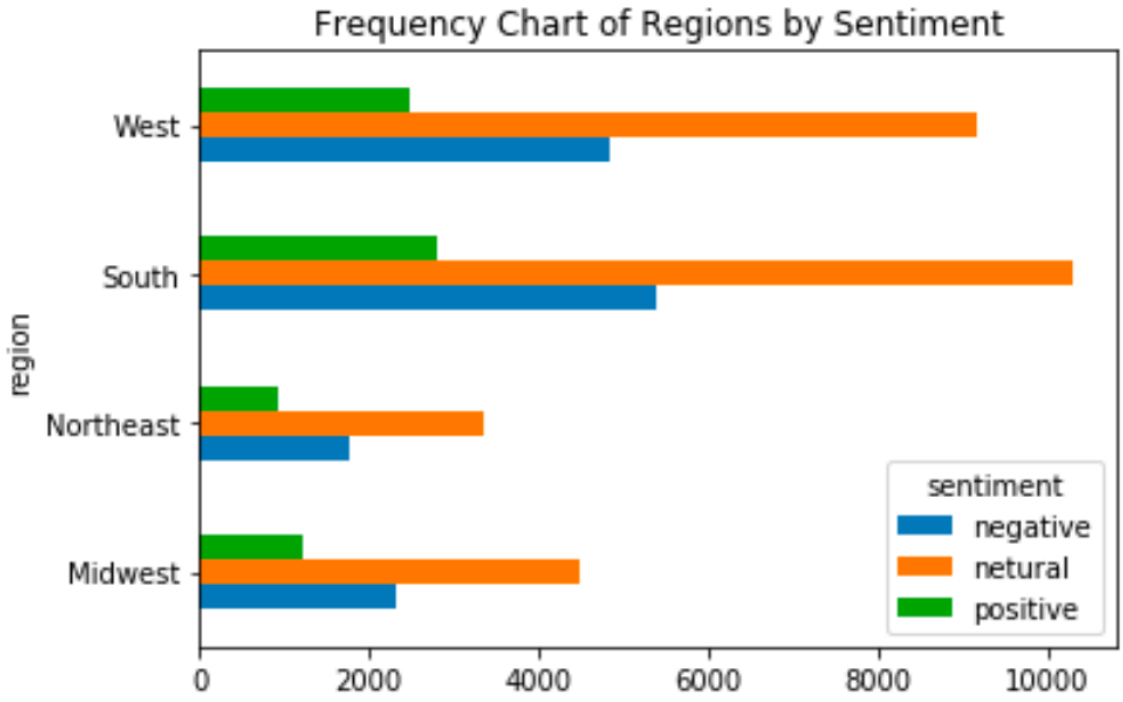
I found the method very clear and easy to use. So, I leverage the *textblob* package to label each tweet as "positive", "negative", or "neutral" and created a new column in the DataFrame to store the label for each row/tweet.

In analyzing this data, I first discovered that the number and percentage of "positive", "negative", and "neutral" tweets were pretty interesting. Specifically, there were far more "neutral" tweets (55.7%) than "negative" (29.1%) and "positive" (15.1%) tweets, as one can see in the following bar chart:

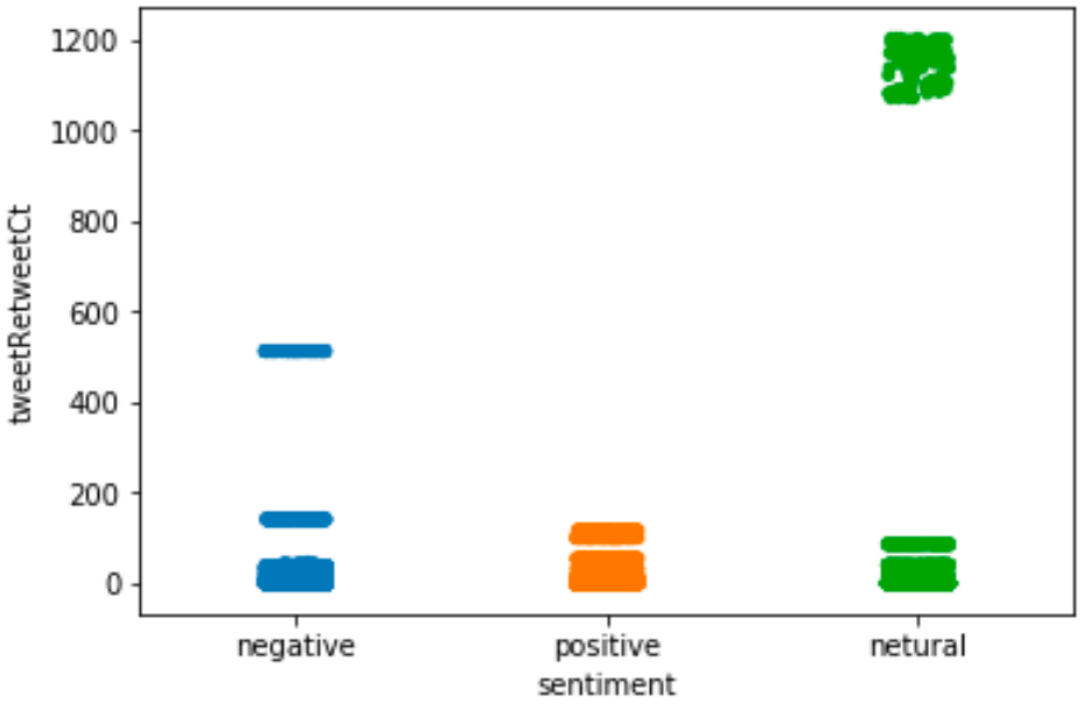


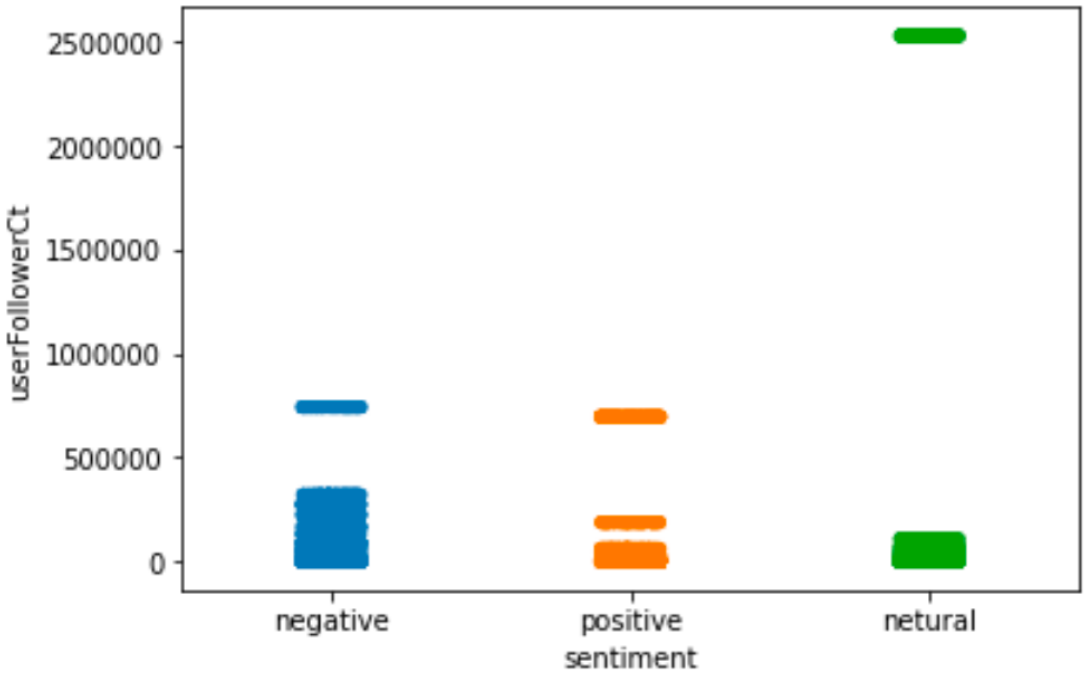
I utilized bar charts to highlight this as well as dividing the tweets by region and then by sentiment. As you can see from the below, there were more overall tweets from the South and West. The percentages of “neutral”, “negative”, and “positive” by region were very consistent with the overall nationwide percentage mentioned above. For example, 15.06% of the tweets in the West region were classified as “positive” while 29.2% were classified as “negative.”

By comparing the regions against each other, one can see that the West region had the highest percentages of “negative” tweets and the Midwest region had the lowest percentage of “positive” tweets. These are highlighted in the following chart:

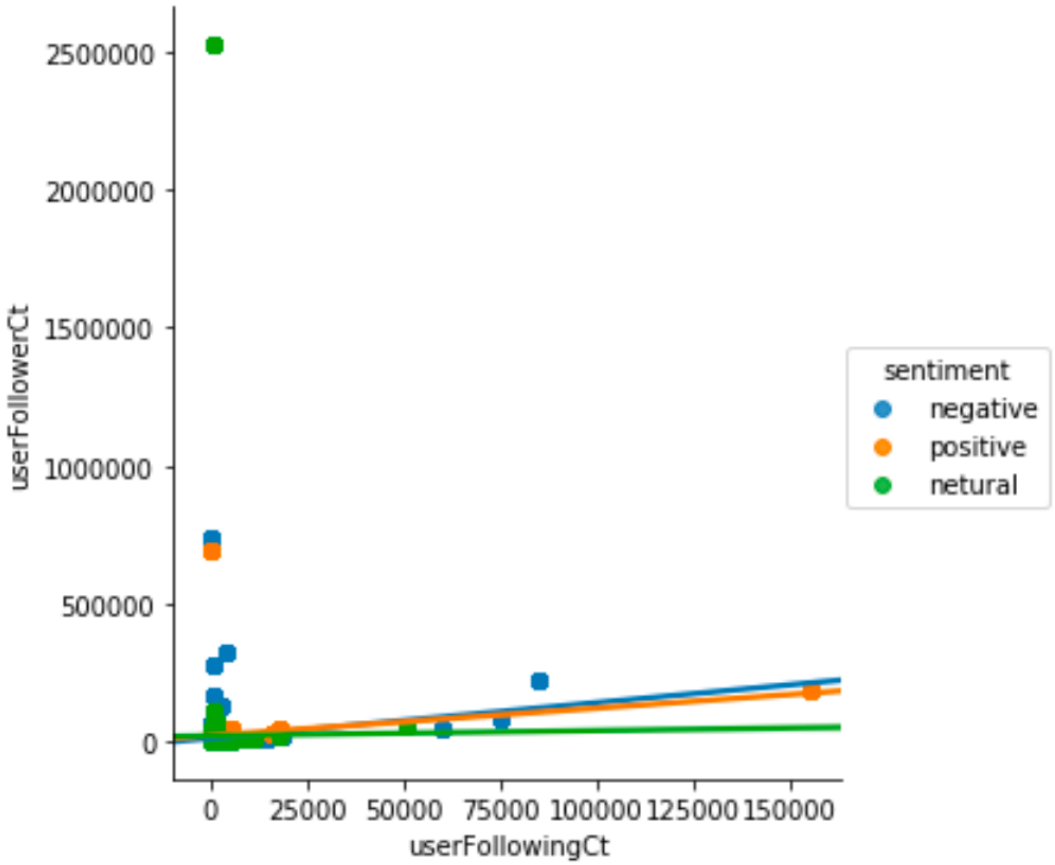


I utilized a number of Seaborn strip plots to learn how the sentiments vary across, for example, number of retweets and a user's number of followers.


1200
1000
800
600
400
200
0
tweetRetweetCt
negative
positive
netural
sentiment


2500000
2000000
1500000
1000000
500000
0
userFollowerCt
negative
positive
netural
sentiment

As mentioned above in the Exploratory Data Analysis section, the number of Twitter followers is positively correlated with the number being followed and were evidenced by linear regression models and plots. The relationship is even stronger among those with "negative" sentiments about the President's immigration position.



Conclusions and Recommendations for Further Analyses:

This model and subsequent analyses can, in fact, be used to answer questions surrounding the decisions and policies of President Donald Trump. In this specific case, the questions were "Do Americans support U.S. President Donald Trump's position on immigration and, specifically, what many cite as the separation of families as they try to enter/cross the U.S.

border? And do those opinions vary by geographic region?" By using Twitter data from the standard public Twitter API, I demonstrated that Twitter data can provide reflections of these sentiments across the country, which can be used to make policy and platform decisions. Of the Twitter users measured in this study, almost twice as many had a negative view than a positive view of the President's position on immigration. Further, those opinions are equally distributed across the primary geographic regions of the United States; there is little to no difference between the regions.

Additionally, by simply replacing and altering the current query string with any topic(s) of interest, this model can be customized and modified to examine public sentiment on any topic. Finally, by upgrading Twitter API access to the Premium API, one could gather historical tweets and collect more precise geolocation information that could be used to conduct trend analyses and more specific (i.e., county- and city-bound) analysis.