Abstract

Over 400 million Tweets are processed per day by Twitter, creating a trove of information. This rich data set can help gauge overall public sentiment and reactions to current events, campaigns, products, ideas, etc. This paper aims to answer the question if we can Twitter sentiment to predict the winner of the New York City mayoral election.

Introduction

New York City is a particularly interesting use case for this analysis, as Twitter age and gender demographics mirror the population fairly closely. Additionally, as show in Figure 1 below, the New York City population using Twitter is spread across the five boroughs. In fact, Jack Dorsey – the founder of Twitter, called New York a "tweeting town" and said "New York City has more Twitter users than any other city in the world and the second most Twitter developers." In 2009, New York City's tweeters comprised 1.44% of all accounts, but produced created 2.37% of all tweets.

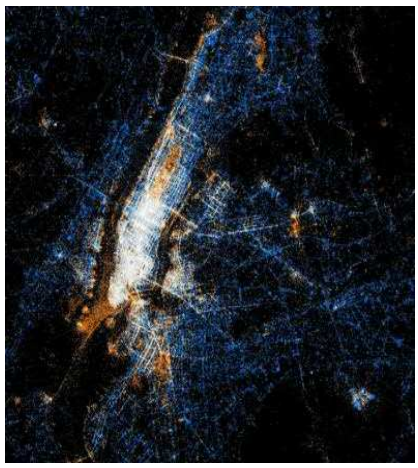

Figure 1

Historically, New York City sways democratic with over 65% of New Yorkers registering themselves as democrats. However, the last two mayors of New York City have been republican. Therefore, it is difficult to predict if a particular political party will take office, as New Yorkers tend to vote on the candidate, rather than political affiliation. The candidates in the 2013 mayoral election were Bill de Blasio (democrat) and Ray Lhota (republican). As shown in Figure 2, Bill de Blasio, won the election on November 5, 2013. We will now examine if we can use Twitter sentiment on de Blasio and Lhota to determine the New York City election.

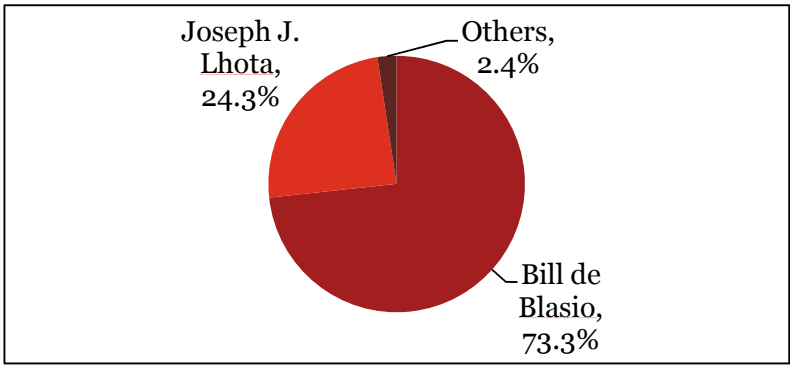

Figure 2

Methods

I had one specific question I wanted to answer using tweets where 'Lhota' or 'Deblasio' was mentioned:

- Can we use Twitter sentiment to predict the New York City mayoral election

To answer this question, I extracted Twitter data from the final five days leading up to the election, October 31, 2013 to November 5, 2013—which led to a total of 6,092 tweets.

To begin, I needed to classify the tweets as positive, negative, or neutral. To do this, there were two methods – either manually classify the tweets or use Naïve Bayes to classify the tweets. Since the sentiment was the ultimate feature that I was using for this analysis, the accuracy of the sentiments was imperative. I therefore elected to manually classify the tweets.

After classifying the tweets, I noticed that there were distinct topics that people discussed via Twitter. I was very intrigued by this and so I then performed a topic model and word frequency on the corpus.

Setup and Analysis

Retrieving the data via Twitter's API was fairly seamless, however what proved to be more difficult was how far I could go back. Twitter sets limits (which changes frequently) on how many tweets one can obtain at a given time. In my first call, I was able to obtain the final five days of tweets. When I attempted to modify the max_id to call older tweets, Twitter prohibited me from going that far back. In the future, it would be better to begin extracting the tweets earlier, so that the population is larger. When calling tweets, Twitter provides the fields illustrated in Figure 3. For this analysis, I selected text, max_id, and created_at as my features. Initially, I was only interested in the text of the tweet and when it was tweeted in relation to the campaign. For future analyses, it would be helpful to obtain the coordinates, to understand the borough of the tweet.

contributors	id_str	retweet_count
coordinates	in_reply_to_screen_name	retweeted
created_at	in_reply_to_status_id	retweeted_status
current_user_retweet	in_reply_to_status_id_str	source
entities	in_reply_to_user_id	text
favorite_count	in_reply_to_user_id_str	truncated
favorited	lang	user
filter_level	place	withheld_copyright
geo	possibly_sensitive	withheld_in_countries
id	scopes	withheld_scope

Figure 3

While reviewing the sentiment, I removed the tweet if it was not in English. Additionally, if the tweet seemed ambiguous and posted a link, I tagged the tweet as neutral – I did not visit each individual link mentioned in the tweet to discern if it was positive, neutral, or negative. Classifying the tweets was the most time consuming portion of this analysis. After classifying these tweets, the breakdown of each tweet is listed in Figure 4.

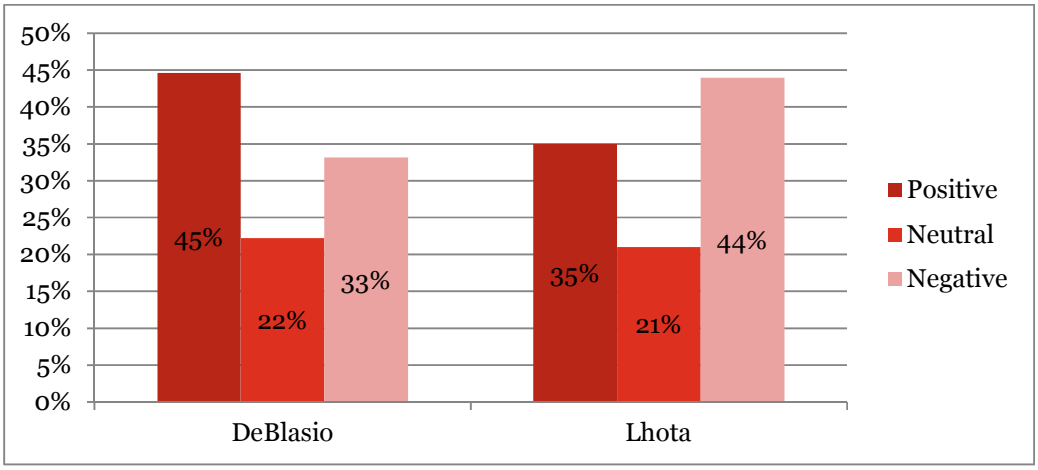

Figure 4

After manually classifying the data, I tried to use Naïve Bayes to classify the tweets. I split the corpus into a 70/30 training/testing set. I then tested the Naïve Bayes classification against manual classification. With the Naive Bayes classifier, the tweets were only correctly identified about 42% of the time. When I looked closer to understand the discrepancy, the Naïve Bayes classifier didn't do very well with sarcastic or ambiguous tweets. For example, the tweets listed below were falsely labeled as neutral tweets, when they are in fact negative tweets. Potentially using more data and a larger training set would help classify sarcastic tweets better.

- RT @AlexisinVT: Deblasio says he'll bring fundamental progressive change to #NYC Sound familiar? Wake up NYers! #nycmayor #tcot #teaparty #\u2026
- #deblasio is not his real name. It's his mother's maiden name. Wilhelm Hitler is #deblasio 's real name. He is a descendant. @BilldeBlasio

As I classified the tweets, I noticed the tweets were often centered on specific topics. As such, I ran a topic model on the tweets. To prepare the data for a topic model, I performed several transformations on the data. These transformations include converting the tweets to lower case, converting URLs to 'URL', replacing @username with 'AT_USER', removing # from hashtags (but still maintaining the hashtag word), removing punctuation and white spaces, and removing stop words.

Before	After
Joe Lhota for New York City mayor http://t.co/nzusbXV8Us #tcot	joe lhota new york city mayor URL tcot

Figure 5

Next, I used a topic model on the candidates. The topic model uses Latent Dirichlet Allocation (LDA), which provides a scoring system to a series of words. The figure below lists the most popular topics by candidate.

Lhota	Deblasio
Giuliani	Stop
Frisk	Frisk
Stop	Socialism

Figure 6

According to various analysts' reports, stop and frisk was the most important issue for New Yorkers in the mayoral election. The other topics of importance to New Yorkers included labor unions, education, and the income gap. However, using Twitter as a proxy for the most important issues, these additional topics

seem less important. (Note: this could be swayed by the fact that we have only the last five days of election data).
Of particular interest, On October 31, 2013 when the court blocked the order requiring changes to the NYPD stop and frisk program, 83% of all tweets that mentioned 'Debelasio' were about stop and frisk. On October 31, 2013 when Rudy Giuliani stated he would campaign with Lhota, 43% of tweets were about Giuliani. As the top topics for both candidates, I wanted to see what the overall sentiment was for each candidate. For Lhota, 86% of the tweets that discussed Giuliani were positive – presumably signifying the public valued Giuliani's support. For Deblasio, the stop and frisk sentiment was more lukewarm. For Deblasio, 39% of tweets swayed negative, 26% neutral, and 35% positive.

Next, I performed a word frequency on the data. The top words coincided with the topics from the topic model and therefore did not add much additional value.

Conclusion and Future Research

While I originally wanted to answer the question if we could use Twitter sentiment to predict the election outcomes, in performing the analysis, I found more value in understanding sentiment towards specific topics. Overall, this analysis does show that if a candidate has more positive tweets he or she may be more likely to win the election.

With additional data, I could understand the public's sentiment and reaction to more topics. Specifically, the Twitter response to mayoral debates might serve as a significant indicator in determining mayoral elections. It would also be interesting to collect the candidates tweets and how they respond to certain tweets. This could help gauge if a candidate can mitigate negative publicity through Twitter or if there is a correlation between a candidate's Twitter activity and election outcomes.

By obtaining all the user information for the tweets, Twitter has a user-call that we can use to obtain user information and specifically their followers. We could then use the demographic information, along with follower information to 1) create a proxy for political affiliation or 2) create social networks to see if the people tweeting positively about the candidates are in a similar network. Lastly, I would like to analyze more elections in New York City, especially elections where the outcomes are closer.