# #GrabTwitterByTheSentimentAnalysis:

## Using Twitter as a Corpus for Political Sentiment Analysis

Kat Aiello and Lisa Chen


## Abstract

In the last decade, microblogging has become a popular tool for Internet users to express their ideas and opinions regarding their daily lives and recent events. As a result, these sites provide valuable data and information for sentiment analysis. Twitter, the largest microblogging platform, has become one of the prominent microblogging sites, with over 300 million monthly active users. In this study, we will use the context of the recent 2016 presidential election to investigate Twitter can be used as a tool for accurate measurement of candidate support and prediction of polls. We built a Naive Bayes classifier, trained on almost 3000 tweets, and conducted an analysis on multiple corpora of over 500 tweets containing a reference to a politician. Our results showed that while our approach to political sentiment analysis does not strongly correlate with polling data, it may indicate strong backlash during scandals. Preliminary evaluations show that our proposed techniques may still require further processing in order to more accurately determine political sentiment.


## Introduction

Twitter is a microblogging service that launched more than a decade ago, has has since gained approximately 300 million unique monthly users, of which, 70 million comprise of users from the United States. Users publish messages containing less than 140 characters, known as "tweets", which cover opinions, discussions, and ideas regarding a variety of topics. Twitter in recent years, has become a widely utilized communication channel in the political arena. In weeks leading up to elections, political issues and major events are increasingly discussed amongst users. In addition, politicians use social media platforms as they attempt to finalize and mobilize the support of the electorate and voters. There has been some debate amongst political analysts concerning the reliability of data collected from the "Twittersphere." Some have acknowledged Twitter as a potentially valuable tool for predicting political sentiment, others consider tweets to be no more than "pointless babble" (pearanalytics 2009; Tumasjan et al. 2010) in the complex realm of politics. In our study, we will evaluate whether tweets reflect offline candidate support in a meaningful way, and whether Twitter may provide reliable poll-like data.

### *Background on the recent United States election 2016*

In our study, we used a total of almost 3000 tweets published at various timepoints corresponding to major events leading up to the US presidential election, which took place on November 8th, 2016. After what was predicted to be a clear win for former Secretary of State Hillary Clinton, Donald Trump unexpectedly took traditionally "blue wall" states such as Michigan, Pennsylvania, and Wisconsin, in addition to winning the perennial swing states of Florida, Iowa, North Carolina, and Ohio. This came as a surprise to pollsters and voters, as most forecasting models continually projected Clinton to win in the months leading up to Election

Day. We thus turn to Twitter for possible answers, expecting political sentiment analysis of Twitter to produce polling data that is unbiased, non-self-selective, and more encompassing and representative of the general voting population.

### *Related work and research questions*

The rapid growth of Twitter has recently garnered the attention of researchers from various disciplines interested in exploring the predictive powers of microblogging data. Well known studies have successfully exploited Twitter data to predict the stock market (Bollen et al 2011; Gayo-Avello 2013) and flu outbreaks (Lampos et al 2010; Gayo-Avello 2013), all of which suggest that Twitter is a powerful and versatile forecasting tool. Tumasjan et al. (2010) evaluated whether or not tweets on Twitter accurately reflected offline political sentiment in any meaningful way. The study took it an additional step further by analyzing whether Twitter could be used to predict party popularity in the real world. Their tweets honed in on the German election of 2009, analyzing over 100,000 tweets. They found that even number of tweets may reflect election results, ultimately, coming close to traditional election polls.

Wang et al. (2012) provided a system for real-time analysis of public sentiment towards presidential candidates in the 2012 U.S. election based on tweets on Twitter. They developed a sentiment model that automatically classified tweets into four categories: positive, negative, neutral, or unsure. The data obtained through their model provided accurate insight into the online political landscape, as there was a strong correlation between emerging political events and news, and the resulting public sentiment changes in response.

While many studies have focused primarily on political sentiment en masse, few have explored whether or not a model can successfully predict an individual's candidate support based on his/her/their tweets. There is also currently no existing literature on the recent election. Additionally, the rise in Twitter usage during the 2016 election will provide us with an extensive amount of data to work with and utilize. Ultimately, the goal of the present explorative study is to address the following research questions:

- How accurately can Twitter inform us about the population's support for political candidates?
- Can twitter successfully serve as an accurate poll generator?

# Methods

### *Data set and time point background*

We examined Tweets 72 hours following major events leading up to the election due to the spikes in political tweeting during these timepoints. These events included the three major presidential debates in addition to the three other notable dates. The first was September 10th and 11th, during which Hillary Clinton referred to a group of Trump supporters as a "basket of deplorables" and appeared to faint at the 9/11 memorial. Also during this period, a comment Trump made in the aftermath of 9/11 came to light when he bragged that, now that the World Trade Center had fallen, the Trump building was the new tallest building in Manhattan. (It actually was not. Maybe he used his tiny hands to measure?) The second date we looked at was

October 7th. On this day, Wikileaks released thousands of Hillary Clinton's campaign emails. A tape also surfaced where Donald Trump can be heard making lewd comments about women, bragging about being able to "grab them by the pussy", often referred to as "Pussygate" by the media.The third was October 28th, when the FBI announced that they were reopening an investigation into Hillary Clinton's home email server.
We collected tweets using import.io[1], a web scraping tool, and Twitter Advanced Search[2]. Our training corpora contains almost 3000 tweets from the six time periods we selected containing either "#ImWithHer" or "#MAGA". Our testing corpora consists of around 500 tweets from each time period mentioning either Hillary Clinton or Donald Trump. For debates, we collected both tweets just containing "#debate" and tweets containing either candidate's name.

### *Analysis and Processing*

In processing, we removed all URLs, hashtags, usernames, emojis, and punctuation from tweets and replace all uppercase letters with lowercase ones.

We then built a Naïve Bayes classifier using Bayes' theorem:

$$P(s \mid T) = \frac{P(s) * P(T \mid s)}{P(T)}$$

Where s is the sentiment and T is the text of the tweet. $P(T \mid s)$ would be calculated by the following:

$$P(G \mid s) = \prod_{g \in G} P(g \mid s)$$

Where G is the set of n-grams in the tweet. We ran our analysis using bigrams.

We calculated the probability for each tweet that it supported Hillary Clinton or Donald Trump. We then found the percentage of support for the entire corpus of tweets was calculated using the following:

$$\%\ supporting\ candidate = \frac{\sum_{t \,\varepsilon\, T} P(s_{ci} \mid t)}{\sum_{c \,\varepsilon\, C} \sum_{t \,\varepsilon\, T} P(s_c \mid t)}$$

Where $c_i$ is the specific candidate, T is the set of testing tweets, and C is the set of all testing candidates. In our case, this only included Hillary Clinton and Donald Trump.

[1] https://dash.import.io/
[2] https://twitter.com/search-advanced?lang=en

# Evaluation

We collected tweets from the 72 hours following each of the three debates. First, we simply searched for tweets containing the hashtag "#debate", but our data did not present any semblance to poll data (Table 1).

| Dates | % Supporting Clinton (#debate) | % Supporting Trump (#debate) | % Supporting Clinton (Names) | % Supporting Trump (Names) | % Supporting Clinton (Polls) | % Supporting Trump (Polls) |
|---|---|---|---|---|---|---|
| Sept. 26, 2016 | 49.6 | 50.4 | 86.8 | 13.3 | 46.4 | 44.9 |
| Oct. 9, 2016 | 25.5 | 74.5 | 15.4 | 84.6 | 48.8 | 43.2 |
| Oct. 19, 2016 | 37.5 | 62.5 | 53.6 | 46.4 | 49.7 | 42.8 |

Table 1. Percentage of candidate support for presidential debates.

Looking at the probabilities for each tweet containing "#debate", we found that neutral, short tweets containing mostly common types had much higher probabilities of supporting either candidate than did most tweets that actually strongly supported a candidate (Table 1). For example:

1) "Her fight is our fight! RT to stand with @HillaryClinton, who has put forward the most pro-equality agenda in history #Debate night" had a $1.0139 * 10^{-35}$ chance of supporting Hillary Clinton and $8.3199 * 10^{-45}$ chance of supporting Donald Trump.

2) "What are you guys watching tonight?! #debate" had a $7.0302 * 10^{-12}$ chance of supporting Hillary Clinton and $4.9705 * 10^{-12}$ chance of supporting Donald Trump.

Although the difference between the probabilities that the second tweet (2) supported either candidate was relatively small, it was massive compared to the probabilities that the first tweet (1) supported either candidate. We suspect that these large differences significantly skewed the results, thus reducing the impact of tweets that clearly showed support. This may be due to the fact that tweets supporting either candidate often mention a specific issue, which is usually a less frequent type and thus decreases the probability of the tweet supporting that candidate more than a more frequent type would. We then decided to collect tweets from the same time periods that either contained the phrase "Hillary Clinton" or "Donald Trump". We figured that we would then have a combination of positive and negative tweets mentioning either candidate, and would be able to filter out the more general, non-partisan tweets. Although we found much greater variation in support of either candidate, very few tweets were general and non-partisan. Thus, despite the fact that there were many fewer neutral tweets with high probabilities of either candidate, our results remained skewed, suggesting that additional processing may still be required for more accurate results.

| Dates | % Supporting Clinton (Tweets) | % Supporting Trump (Tweets) | % Supporting Clinton (Polls) | % Supporting Trump (Polls) |
|---|---|---|---|---|
| Sept. 11, 2016 | 56.9 | 43.1 | 46.7 | 43.0 |
| Oct. 7, 2016 | 93.7 | 6.3 | 48.9 | 43.2 |
| Oct. 28, 2016 | 1.6 | 98.4 | 49.6 | 43.9 |

Table 2. Percentage of candidate support for other major time timepoints.

We saw an even greater difference in support when we analyzed tweets from major scandals.

# Discussion

We analyzed each testing corpus containing approximately 500 tweets that mentioned the names of the presidential candidates in the 2016 election. Overall, we found that while our model may occasionally reflect voter preferences similar to those of traditional polls, it does not accurately and reliably indicate candidate support. Thus, we can not draw conclusions regarding Twitter's ability to generate unbiased and encompassing poll data.

While our results greatly differed from the polls at the time, we noticed that favor was most skewed toward one candidate or another immediately following a scandal. This was clearly observed in the data from on October 7th, 9th, and 28th. The severe decrease in Trump support on the 7th was likely attributed to the massive criticism Trump received for Pussygate. Soon after, WikiLeaks leaked thousands of emails from the Democratic National Convention. Clinton had climbed back up by our next time period on October 19th, but then on October 28th, suffered an extreme dip in support likely attributed to the FBI reopening an investigation into her use of a home email server.

This study had several limitations. First, as evidenced from our results, there was an abundance of neutral tweets that may have skewed our results. In order to account for these neutral tweets, we should have analyzed the salience values of specific words and excluded words that were not strongly associated with either candidate.

$$salience(g) = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} 1 - \frac{min(P(g \mid s_i), P(g \mid s_j))}{max(P(g \mid s_i), P(g \mid s_j))}$$

Where N is the number of sentiments, in our case 2. Lower values of salience indicate the n-gram does not strongly indicate either sentiment and should be excluded.

This would eliminate any tweets that did not exhibit strong sentiment towards either candidate, and in return, could possibly output a more accurate forecast.

Even with the best tools, we learned that Twitter may not be the best indicator of political sentiment. We found that in the aftermath of scandals, favor for the affected candidate dropped significantly. However, this does not indicate as large a change in the polls. Presidential Elections are usually won by a relatively small margin. Certain events in an election are unlikely to sway polls more than a few points. Our hypothesis is that in the aftermath of a scandal, one can expect significantly more negative tweets about the affected candidate, changing the proportion of tweets that support that candidate, making it appear as if very few support them anymore.

# References


Bollen, Johan, et al. "Twitter mood predicts the stock market." *Journal of computational science* 2.1 (2011): 1-8.

Gayo-Avello, Daniel. "A meta-analysis of state-of-the-art electoral prediction from Twitter data." *Social Science Computer Review* 31.6 (2013): 649-679.

Lampos, Vasileios, and Nello Cristianini. "Tracking the flu pandemic by monitoring the social web." *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*. IEEE, 2010.

Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. No. 2010. 2010.

pearanalytics. 2009. Twitter Study. Retrieved December 15, 2009 from http://www.pearanalytics.com/wpcontent/uploads/2009/08/Twitter-Study-August-2009.pdf

Tumasjan, Andranik, et al. "Predicting elections with twitter: What 140 characters reveal about political sentiment." *ICWSM* 10.1 (2010): 178-185.

Wang, Hao, et al. "A system for real-time Twitter sentiment analysis of 2012 US Presidential Election cycle." *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012.