Khagay Nagdimov

Text as Data: DS-GA 3001-04

Professor Arthur Spirling

TA Kevin Munger

05/11/2016

*Twitter Sentiment Can Partially Predict Winners of the Primaries*

***Introduction***

Every day twitter users share their opinions with the world. Persons from New York may have fundamentally different views of the world than Texas citizens and Twitter serves as an outlet but, more importantly, an archive of those opinions. Based on these tweets, predictions can be made about who will win the presidential race but, there is not much proof as to why this can be done. In this paper, a study that was performed will be summarized to compare whether tweet sentiment and primary wins have a relationship, meaning does have an average amount of positive tweets for a candidate correlate to that candidate winning the primary.

***Literature***

There has been much research in this area. Mostly, classification methods using dictionaries have been utilized in previous studies. One such study using LIWC (Linguistic Inquiry and Word Count) dictionary created by James W. Pennebaker. (Tumasjan, Andranik, Sprenger O., Timm, Sandner G., Philipp, Welpe M., Isabell, 2010). Results confidently showed that Twitter is used for “political deliberation.” Other studies have used Naïve Bayes classifiers and support vector machines (SMVs) (Pak, A., & Paroubek, P, 2010). Other researchers have shown that computer algorithms are better at annotating text than human coders, providing the fact that results from sentiment analysis algorithms are trustworthy. (Makazhanov, A. & Rafiei, D. & Waqar, M., 2014). These papers show that text can be analyzed based on the words and the frequency of those words. Such a methodology will be applied in this paper, specifically the use of a dictionary (Hu and Liu) and a data set of tweets.

***Theory and Hypotheses***

There has not been research that has compared the results of the sentiment analysis with the actual results of the primary. This has mainly been an issue because only recently have the primaries been taking place (started February 1, 2016). In this paper, the question of whether the opinions on twitter directly correlate to who will win the primary will be answered. I predict that the two do not have a direct relationship because of the many other factors such as whether people on Twitter are actual voters, whether people change their opinion last minute before voting, whether the voters from the state geocode their tweet, and other factors. However, what can be answered from such a study will be whether twitter can be used to predict the primary winners.

***Data and Methods***

To begin with, the Twitter API was used to collect tweets from April 8, 2016 to May 5, 2016. In total, 27,486 tweets were collected. This data set is new when compared to other papers because it contains tweets written during the primaries.

The result from the Twitter API was a JSON file containing several key-value pairs for each collected tweet: an identification number, the state the tweet was sent from (will be discussed further), the text of the tweet, the candidate the tweet is mentioning, and several others. Not all tweets are geocoded and thus, this project focuses on geocoded tweets which will be important to remember when results are analyzed.

The candidate and state key-value pairs were extracted from the JSON file using the Python programming language. A folder for each candidate (Donald J. Trump, Hillary R. Clinton, Ted Cruz, and Bernie Sanders) was made and filled with files containing tweets for each state. The files for each candidate was then loaded into the R statistical programming language where the analyses were performed using the Quanteda package developed by Ken Benoit and Paul Nulty.

The Hu and Lui sentiment lexicon dictionary was used to define which words represent a negative or positive meaning. This method of analyzing text has been used before but, on a data set that is not as recent.

For each state, a data term matrix a.k.a. data frequency matrix (dfm) was prepared for the tweets from each state using the Hu and Lui dictionary and library package quanteda. The data frequency matrix was then filtered to the number of positive and negative words for each tweet. The difference in the two resulted in the sentiment score of the tweet.

*Results*

The results of the sentiment analysis for each state for each candidate are shown in Tables 1 and 2. The tables show the winner of the primary based on who has a higher sentiment score. Likewise, the winner of the actual primary is also shown, provided by Bloomberg Politics. The first chart shows the results for the two republican candidates: Ted Cruz and Donald Trump. The second chart shows the results for the two democratic candidates: Bernie Sanders and Hillary Clinton. In the tables, there are green highlighted states. Those states had tweets collected before and after the primaries. The other states had tweets collected only after the primary took place.

**Table 1: Republican Sentiment Analysis Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| State | Ted Cruz | Donald Trump | Winner of Primary (Based on Sentiment) | Winner of Primary (Based on Bloomberg Politics) |
| Alabama | -0.2012 | -0.6851852 | Cruz | Trump |
| Alaska | 0.0667 | -1.058824 | Cruz | Cruz |
| Arizona | -0.3901345 | -0.5702128 | Cruz | Trump |
| Arkansas | -0.3333333 | -0.4625 | Cruz | Trump |
| California | -0.2194534 | -0.6581843 | Cruz | N/A |
| Colorado | -0.1729323 | -0.3545455 | Cruz | Trump |
| Connecticut | -0.2531646 | -0.6952381 | Cruz | Trump |
| Delaware | -0.2307692 | -0.3809524 | Cruz | Trump |
| Florida | -0.1773472 | -0.5679348 | Cruz | Trump |
| Georgia | -0.2553648 | -0.564482 | Cruz | Trump |
| Hawaii | -0.05714286 | -1.142857 | Cruz | Trump |
| Idaho | 0.0625 | -1.173228 | Cruz | Cruz |
| Illinois | -0.1574803 | -0.6848485 | Cruz | Trump |
| Indiana | -0.3136646 | -0.847352 | Cruz | Trump |
| Iowa | -0.1206897 | -0.4 | Cruz | Cruz |
| Kansas | -0.1186441 | -0.7647059 | Cruz | Cruz |
| Kentucky | -0.0859375 | -0.6141732 | Cruz | Trump |
| Louisiana | -0.1785714 | -0.5103448 | Cruz | Trump |
| Maine | 0.15 | -0.1538462 | Cruz | Trump |
| Maryland | -0.1624365 | -0.7932692 | Cruz | Trump |
| Massachusetts | -0.1034483 | -0.660066 | Cruz | Trump |
| Michigan | -0.1377049 | -0.4203822 | Cruz | Trump |
| Minnesota | -0.3245614 | -0.5882353 | Cruz | Rubio |
| Mississippi | -0.07142857 | -0.9782609 | Cruz | Trump |
| Missouri | -0.02816901 | -0.8222222 | Cruz | Trump |
| Montana | -0.2142857 | -1 | Cruz | N/A |
| Nebraska | -0.1724138 | -0.8571429 | Cruz | N/A |
| Nevada | -0.3085106 | -0.3785714 | Cruz | Trump |
| New Hampshire | -0.2619048 | -0.5405405 | Cruz | Trump |
| New Jersey | -0.04601227 | -0.664311 | Cruz | N/A |
| New Mexico | -0.1904762 | -0.8717949 | Cruz | N/A |
| New York | -0.1457883 | -0.5720601 | Cruz | Cruz |
| North Carolina | -0.1809524 | -0.7386364 | Cruz | Trump |
| North Dakota | -1.125 | -0.8333333 | Trump | Cruz |
| Ohio | -0.007109005 | -0.6336336 | Cruz | Kasich |
| Oklahoma | -0.4086022 | -0.5909091 | Cruz | Cruz |
| Oregon | -0.3846154 | -0.552795 | Cruz | N/A |
| Pennsylvania | -0.2107527 | -0.6527197 | Cruz | Trump |
| Rhode Island | 0.07142857 | -0.9428571 | Cruz | Trump |
| South Carolina | -0.06666667 | -0.7979798 | Cruz | Trump |
| South Dakota | 0.25 | -0.3 | Cruz | N/A |
| Tennessee | -0.1856287 | -0.7664671 | Cruz | Trump |
| Texas | -0.2057677 | -0.5192108 | Cruz | Cruz |
| Utah | 0.2638889 | -0.7580645 | Cruz | Cruz |
| Vermont | 0.1333333 | -0.4117647 | Cruz | Trump |
| Virginia | -0.1442623 | -0.3880126 | Cruz | Trump |
| Washington | -0.1528384 | -0.7419355 | Cruz | N/A |
| West Virginia | -0.2045455 | -0.5647059 | Cruz | N/A |
| Wisconsin | -0.129771 | -0.5181818 | Cruz | Cruz |
| Wyoming | -0.7142857 | -0.7222222 | Cruz | Cruz |

**Table 2: Democratic Sentiment Analysis Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| State | Hillary Clinton | Bernie Sanders | Winner of Primary (Based on Sentiment) | Winner of Primary (Based on Bloomberg News) |
| Alabama | -0.0222222 | -0.38 | Clinton | Clinton |
| Alaska | -0.125 | 0 | Sanders | Sanders |
| Arizona | -0.1538462 | -0.2177419 | Clinton | Clinton |
| Arkansas | -0.1081081 | -0.4545455 | Clinton | Clinton |
| California | -0.1552795 | -0.3006645 | Clinton | N/A |
| Colorado | -0.07407407 | -0.2134831 | Clinton | Sanders |
| Connecticut | -0.1320755 | -0.2363636 | Clinton | Clinton |
| Delaware | 0.06451613 | -0.2 | Clinton | Clinton |
| Florida | -0.164 | -0.350211 | Clinton | Clinton |
| Georgia | -0.2040816 | -0.2910448 | Clinton | Clinton |
| Hawaii | -0.2068966 | -0.4615385 | Clinton | Sanders |
| Idaho | 0 | 0.4 | Sanders | Sanders |
| Illinois | -0.2461538 | -0.2877698 | Clinton | Clinton |
| Indiana | -0.09493671 | -0.410596 | Clinton | Sanders |
| Iowa | -0.04597701 | -0.410596 | Clinton | Clinton |
| Kansas | -0.2608696 | 0.09090909 | Sanders | Sanders |
| Kentucky | -0.1964286 | -0.3928571 | Clinton | N/A |
| Louisiana | -0.05405405 | -0.2666667 | Clinton | Clinton |
| Maine | -0.2142857 | -0.8571429 | Clinton | Sanders |
| Maryland | -0.09756098 | -0.3894737 | Clinton | Clinton |
| Massachusetts | -0.2671756 | -0.4433962 | Clinton | Clinton |
| Michigan | -0.25 | -0.2096774 | Sanders | Sanders |
| Minnesota | -0.09090909 | -0.09615385 | Clinton | Sanders |
| Mississippi | -0.3478261 | -0.25 | Sanders | Clinton |
| Missouri | -0.3333333 | -0.125 | Sanders | Clinton |
| Montana | -0.1785714 | -0.3513514 | Clinton | N/A |
| Nebraska | -0.38888889 | 0 | Sanders | Sanders |
| Nevada | -0.1904762 | -0.4814815 | Clinton | Clinton |
| New Hampshire | -0.25 | -0.1818182 | Sanders | Sanders |
| New Jersey | -0.05194805 | -0.3656716 | Clinton | N/A |
| New Mexico | -0.8 | -0.125 | Sanders | N/A |
| New York | -0.2961165 | -0.3430322 | Clinton | Clinton |
| North Carolina | -0.2111111 | -0.1343284 | Sanders | Clinton |
| North Dakota | 1 | No Tweets | N/A | N/A |
| Ohio | -0.2797203 | -0.3469388 | Clinton | Clinton |
| Oklahoma | 0.28 | -0.7777778 | Clinton | Sanders |
| Oregon | -0.1028037 | -0.2421053 | Clinton | N/A |
| Pennsylvania | -0.21 | -0.2377049 | Clinton | Clinton |
| Rhode Island | 0.03448276 | -0.2173913 | Clinton | Clinton |
| South Carolina | 0.1578947 | -0.07692308 | Clinton | Hillary |
| South Dakota | -0.6666667 | No Tweets | N/A | N/A |
| Tennessee | 0.1714286 | -0.4871795 | Clinton | Clinton |
| Texas | -0.1950464 | -0.1976285 | Clinton | Clinton |
| Utah | -0.02777778 | -0.4 | Clinton | Sanders |
| Vermont | -0.125 | -0.12 | Sanders | Sanders |
| Virginia | -0.08108108 | -0.2021277 | Clinton | Clinton |
| Washington | -0.09474026 | -0.06 | Sanders | Sanders |
| West Virginia | -0.5333333 | -0.3846154 | Sanders | N/A |
| Wisconsin | -0.3283582 | -0.260274 | Sanders | Sanders |
| Wyoming | 0.5 | 0.6666667 | Sanders | Sanders |

For the republican candidates, 31/41 of the states that have had a primary did not vote for the person who was tweeted most favorably. That is a 75.6% discrepancy. For the democratic candidates, the discrepancy was 12/41 or 29.2%. For the states that had tweets collected before and after the primary the results are: predictions based on twitter sentiment were correct 7/8 times or 87.5% for democrats while for republicans it was 2/8 or 25%. The hypothesis that primary winners can be predicted from twitter sentiment was a valid prediction for the democratic candidates but, not for republican candidates.

***Discussion***

The results of this paper show that Twitter can be used to predict the winner of the democratic primaries, with an 87.5% accuracy rate for primaries that have occurred while for republican primaries, predictions are 25% accurate. These results are important because conclusions can be made about how well Twitter can predict the events that will happen in reality.

As mentioned in the Literature section of this paper, there are many reasons why the results are the way they are. Perhaps, Twitter users are mainly citizens that do not vote or users use Twitter to vent but, when they vote, their decision about who they want to vote for changes, or the users who geocode their tweets are not going out and voting. However, what can be concluded confidently is that if Twitter is used to predict primary winners it cannot predict the winners of the republican primaries but, it can predict winners of the democratic primaries.

Alongside the sentiment scores, results show that most tweets are negative which shows that Twitter serves as a place where users complain about the candidates rather than praise them. This is an interesting result because it shows what Twitter is used for.

Further work needs to be done to validate these results by continuing to collect tweets while the rest of the states go through the primaries. However, the current results have helped research in this field in several ways. One of which is validating whether or not sentiment scores can be used to categorize states as democratic or republican supporters. There has been much research and projects stating such results but, the validation part has been difficult, until now.

**Citations**

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, *10*, 178-185.

Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREc* (Vol. 10, pp. 1320-1326).

Bloomberg L.P. (n.d.). 2016 Delegate Count Tracker. Retrieved May 11, 2016, from <http://www.bloomberg.com/politics/graphics/2016-delegate-tracker/>

Makazhanov, Aibek, Rafiei, Davood, Waqar, Muhammad. (2014). Predicting political preference of Twitter users. *Social Network Analysis and Mining,* 4, 193.