INCLUSION OF TWITTER DATA IN PREDICTING CONGRESSIONAL ELECTION OUTCOMES

By

Trey Feldman

Topics in Econometrics

Dr. Andrew Paizis

New York University

June 29, 2015

I. Abstract

Many claims of successfully predicting the outcomes of elections based on Twitter data have been made in recent years. These claims, however, have been highly controversial and have been rebutted by many researchers. This paper examines the efficacy of including two forms of twitter data into congressional election prediction models–data involving sentiment analysis and raw numbers data. In accordance with researchers that have denied the usefulness of Twitter data in predicting elections, this paper finds a similar result. While twitter data can successfully be used in election prediction models, it is the classic election prediction variables, such as who the incumbent is, that are most useful and are necessary in correctly predicting elections.

II. Introduction

The boom of social media in recent years has supplied massive amounts of data for researchers to use in many different fields. Using just raw volume of tweets for candidates as well as sentiment analysis, many researchers have claimed to have found reliable ways to predict elections. Sentiment analysis aims at classifying the words and opinions social media users put out in forms such as reviews on Amazon, Facebook statuses, and, for the purposes of this paper, Twitter tweets. These tweet classifications can, some claim, be used to effectively predict elections. These claims, however, are considered controversial by many. This paper thus seeks to form an independent opinion on including two separate variables in a model: one variable being the portion of all positively labeled tweets from each election mentioning the democratic candidate, and the other variable being the portion of all tweets in each election mentioning the democratic candidate.

In order to test the efficacy of using sentiment analysis on twitter data to predict the outcomes of elections, a naïve Bayes classifier was used to automatically classify every tweet mentioning each candidate in the weeks before the 2012 congressional election. These tweets were classified as either ‘positive’ or ‘negative.’ The specifics of how this classification algorithm works will be briefly explained later.

A variable for positive tweets for the democratic candidate, as well as negative tweets for the candidate, were both originally intended to be used in the model. The two variables, however, were found to be highly correlated with each other with correlation coefficient of .722, and the negative tweets variable was dropped from the model.

To both enhance the prediction models as well as to assess how useful twitter data based variables are, several other independent variables were added to the models. The variables used cover incumbency, state ideologies, and the health of the individual state economies. These variables are fairly classic variables in election prediction models.

III. Literary Review

One need not look far to find a slew of articles on the subject of predicting elections with Twitter data in various ways. For example, one such study by Karissa McKelvey, Joseph DiGrazia, and Fabio Rojas concluded that, “tweets that employ free-text, hashtags, and @mentions… [are] significantly representative of the public's voting choices,” (2014). While this study focuses on raw numbers of tweets collected about candidates and claims a strong correlation between those numbers and actual poll numbers, various other studies focus on the use of sentiment analysis to predict elections. One such study, by Vadim Kagan, Andrew Stevens, and V.S. Subrahmanian, claims, in absolute terms, that the elections in Pakistan that they were studying could be cheaply and accurately predicted using their sentiment analysis techniques.

While many academic articles can be found in support of using Twitter data to predict elections, there are many researchers that have disputed it outright. These reasons are nicely summed up in an article, plainly titled, “No, You Cannot Predict Elections With Twitter,” by Daniel Gayo-Avello of the University of Oviedo. One of Gayo-Avello’s main concerns is with the scientific method employed by nearly all of these studies. First and foremost, finding an article that has actually successfully predicted an election, rather than found a model that fits the results of a previous election, is nearly, if not entirely, impossible to find. Second, Gayo-Avello claims, in accordance with other research, that sentiment analysis on this sort of data is not accurate enough to predict elections (and sometimes not much more accurate than a random classifier). Finally, Gayo-Avello points out that this sort of methodology is a classic case of self-selection bias. The large samples of data being collected from twitter for these studies are not random, but rather are coming from a small group of politically active users. More specifically, Tumasjan, Sprenger, Sandner, and Welpe found that, of all the users contributing to their pool of over 100,000 tweets in their study, only 4% of those users had contributed over 40% of the messages (2010). Thus, the samples of data do not represent all demographics, represent a very small subset of the population in general, and give more weight to those that are extremely politically active. For these reasons, many researchers do not believe it is possible to put together a model based on Twitter data that can consistently predict elections. These researchers point to better methods of predicting elections such as noting who the incumbent is, the status of the economy, and the ideologies held by certain areas of the country.

IV. Model

Two linear probability models were examined. The dependent variable in both cases was whether the democrat was elected–receiving a value of 1 if they were elected, and 0 if they were not elected. Both models included variables for whom the incumbent was, the presumed political ideology of the state, and the job creation rate for 2011 and 2012. The first model included an additional variable for the share of positive tweets that the democratic candidate received, and the second model included a variable for the share of total tweets that the democratic candidate received. The base variables included in both models were chosen as a base to both test claims that variables such as these are the main factor in predicting congressional elections, and to test the significance of adding a variable based on Twitter data to the model.

The incumbency variable is a dummy variable that received the value of one if the democrat running was the incumbent, negative one if the republican running was the incumbent, and zero if no incumbent was running. This is an important variable in election prediction models because incumbents are notoriously difficult to beat in elections. This difficulty grows even more in congressional elections where districts often sway heavily towards one political ideology. This may be due to something such as that the congressional district is in an urban area where people typically vote more liberally, or it may be due to the practice of gerrymandering. Gerrymandered districts are those that are intentionally redrawn by the party in control to favor their chances of winning. Because of this practice, there exist many districts where elections are easily won by the candidate of one party in every election, and some districts where the candidate of one party is often uncontested in the general election.[1] Due to the notorious difficulty in beating an incumbent candidate, the coefficient of this variable

[1] For the purposes of this paper, those elections which were uncontested were not included in the dataset.

was expected to have a relatively high, positive coefficient. Thus, the probability of the democrat winning would get a significant boost if the democrat was the incumbent, and the probability of the democrat winning would go significantly down if the republican was the incumbent.

Using senators as a predictor of state ideologies is thought to be a good predictor of state ideologies due to the fact that senators serve six year terms and that they are statewide elections. It can be inferred that the longer terms of senators would make it less likely for voters to vote outside of their typical political ideology. This is because a person should naturally be expected to be more wary of taking a risk on a candidate outside of their typical party affiliation for a six year term than a two year term. Additionally, because senate elections are state wide elections, they are not subject to gerrymandering and are representative of votes from all over the state. Thus, if the entire state has elected two democrats in two separate elections, it can be inferred that the state likely has left leaning political ideologies.

The second independent variable included in both models was a dummy variable accounting for the presumed political ideology of the state based on the current senators elected to the state. The variable received a value of one if both senators in the state were democrats in 2012, zero if one senator was a democrat and one senator was a republican in the state in 2012, and negative one if both senators were republicans in the state in 2012[2]. The importance of this variable is to account for the

[2] There were two senators in 2012 that had won as independents: Joe Lieberman of Connecticut and Bernie Sanders of Vermont. Both were counted as democrats for the purposes of this variable due to their liberal, and far from conservative, ideologies. The resultant values of one for Connecticut and Vermont, labeling them as states with liberal ideologies, should be uncontroversial due to the historically liberal voting record of both states.

ideologies that intrinsically belong to many states. For example, southern states, such as Alabama, are notably conservative, and New England states, such as Vermont, are notably liberal. Though there exist congressional districts within states that do not vote the same as the rest of their state generally does, a state that generally votes more conservative or liberal overall will obviously have more conservative or liberal members of congress. Thus, the coefficient of this variable was expected to be positive–meaning that the chances of the democrat being elected would go up if the state leaned democratically, and go down if the state leaned conservatively.

The final variable included in both models was a variable for net job creation rate in the years 2011 and 2012. This variable was included to account for economic conditions in each state and the assumption that a higher job creation rate would signal a better state economy and thus a higher chance for a candidate to be elected. This variable was expected to have a positive, but low coefficient. In presidential elections, the state of the economy plays a huge role in the outcome of the election. Though this variable was expected to play some role in the outcome of the congressional elections, the impact was not expected to be nearly as high as it would be for a presidential or gubernatorial election for two reasons. First, though the actual impact that the executive branch of the United States and each individual state has on the economy is controversial, blame or credit for a terrible or amazing economy most often goes to the executive branch. Second, because of congressional districts often being rather lopsided in their political views, it would take a rather disastrous economy in a specific state for large portions of any given congressional district to switch parties. Therefore, if a state has had quite a negative net change in job creation, the chances of the candidates being reelected should go down quite a bit. If the change has not been very dramatic in either direction, this variable should not be expected to have a very large impact on the election.

For the first model, an additional variable was added accounting for the percentage of total tweets classified as positive mentioning the democratic candidate. This variable was computed as follows:

$$DemShareOfPosTweets_i = \frac{DemPosTweets_i}{DemPosTweets_i + RepPosTweets_i}$$ [3]

If actually effective, the coefficient of this variable was expected to be positive due to the obvious fact that if one candidate is receiving more positive attention on Twitter, their chances of winning their election should be higher. Due to the imperfect nature of classification, discussed later, and the doubts previously raised by other researchers, the actual substantiveness of this variable was unknown.

Finally, the extra variable included for the second model (instead of the variable discussed in the previous paragraph) was a variable for the portion of total tweets mentioning the democratic candidate. It was computed as follows:

$$DemShareOfTotalTweets_i = \frac{DemTotalTweets_i}{DemTotalTweets_i + RepTotalTweets_i}$$

This variable was expected to have a positive sign due to the obvious fact that if a candidate is receiving more attention on Twitter, their chances of winning their election are likely higher. As mentioned previously, some research has claimed a high

[3] As mentioned previously, due to the high collinearity of the share of negative tweets received by the democratic party and the share of positive tweets received by the democratic party, no variable referencing the share of negative tweets could be included in the model.

correlation between the share of Twitter attention one candidate is getting to their poll numbers. What is disputed, however, is the actual impact these sort of variable has on an election prediction model. The inclusion of this variable into this model was in order to examine, precisely, what the impact this variable may have on predicting an election.

In summary, the two models being considered are as follows:

$Y_i = \beta_1 + \beta_2 INC_i + \beta_3 IDE_i + \beta_4 NJC_i + \beta_5 DPTS_i + u_i$ and

$Y_i = \beta_1 + \beta_2 INC_i + \beta_3 IDE_i + \beta_4 NJC_i + \beta_5 DTTS_i + u_i$

TABLE 1

Summary of Variables

Variable	Definition	Value
Y	Probability Democrat Will Be Elected	0-1
INC	Incumbent	Democrat = 1; Republican = -1; No Incumbent = 0
IDE	Ideology	Two Democratic Senators = 1; Two Republican Senators = -1; Else = 0
NJC	Net Job Creation	Percent
DPTS	Democratic Positive Tweet Share	$\frac{DemPosTweets_i}{DemPosTweets_i + RepPosTweets_i}$
DTS	Democratic Total Tweet Share	$\frac{DemTotalTweets_i}{DemTotalTweets_i + RepTotalTweets_i}$

V. Data

All data used for this research was publicly available data downloaded and scraped from the web[4]. Most importantly, information on the individual congressional elections was necessary in order to collect further information about each election and tweets about each candidate. A spreadsheet was initially downloaded from the Federal Elections Commission website containing the names of the winners of each election, their parties, whether they were the incumbent, and how many votes each party received in each election. In order to obtain the names of the losing opponents, it was then necessary to loop through the list of winning candidates and examine their candidacy pages on ballotpedia.org. Ballotpedia.org is an objective, online encyclopedia filled with election information. On each winning candidates' page on the site, their opponents names for the 2012 election were listed. These names were scraped and stored in order to grab their relevant tweets.

In order to obtain the tweets for each candidate, the topsy.com application programming interface (API) was used. Topsy.com is database of every single tweet that has been tweeted since 2006. Using their API, specific tweets involving each

[4] Web scraping is a process where an algorithm is used to automatically load large amounts of web pages and 'scraping' relevant information from them.

candidate were grabbed by iterating through the list of candidates. The specifics of the tweets obtained are as follows:

- ➢ Any tweet mentioning the candidate's name in full; e.g., “My vote is for Sonya Smith!”
- ➢ Any tweets with the candidate’s name in a hashtag; e.g., “Is there anyone worse than #JoanCrandy?”
- ➢ For candidates with middle initials, or more than two parts to their names, the algorithm would search for (for candidate Patty May Blue): Patty May Blue, Patty Blue, #PattyMayBlue, and #PattyBlue. This was to account for the fact that, though middle initials or extended names may have been listed, the candidate may have popularly been known by their first and last name online.
- ➢ For candidates with a suffix in their name, i.e., Tom Pinnochio, Jr., the algorithm would search for tweets containing: Tom Pinnochio, Jr., Tom Pinnochio, #TomPinnochioJr, and #TomPinnochio. Again, this was to account for certain candidates that may not necessarily commonly include the junior at the end of their name.
- ➢ All tweets obtained were from October 8, 2012 to November 5, 2012–the four weeks before the election which occurred on November 6, 2012.

In initially assessing the tweets obtained, it quickly became apparent that many candidates shared a name with professional athletes. To avoid having these irrelevant tweets throw off the numbers, the search algorithm rejected any tweets containing any word in a list of banned words. The list of banned words included the names of every sports team in the National Football League, the National Basketball Association, and Major League Baseball. Additionally, popular sports words such as recruit, championship, traded, player, game, and touchdown were also on the banned words list. Though this helped tremendously to keep the tweets obtained on topic and relevant to each candidate, it is absolutely impossible to avoid every single irrelevant tweet. After removing the large amount of sports related tweets, however, the issue of irrelevant tweets was believed to be a rather minor issue and mainly applicable to those few candidates with more common names. To avoid spam bots and tweets that had been retweeted a bunch, all retweets and repeat tweets were also not included.

For several reasons, 100 elections were removed from the dataset, leaving 335 elections to be examined. Two elections were removed due to candidates sharing the same name. Including these two candidates would have made it extremely difficult to differ which candidate any specific tweet was referring to without either manually looking through all of the tweets or using some sort of geospatial information about where the tweets were coming from. Several other elections were dropped from the dataset because they were uncontested elections. And finally, several elections were dropped because no tweets were found for one of the candidates. This was likely due to the candidate more commonly going by another name than the one listed on the ballot.

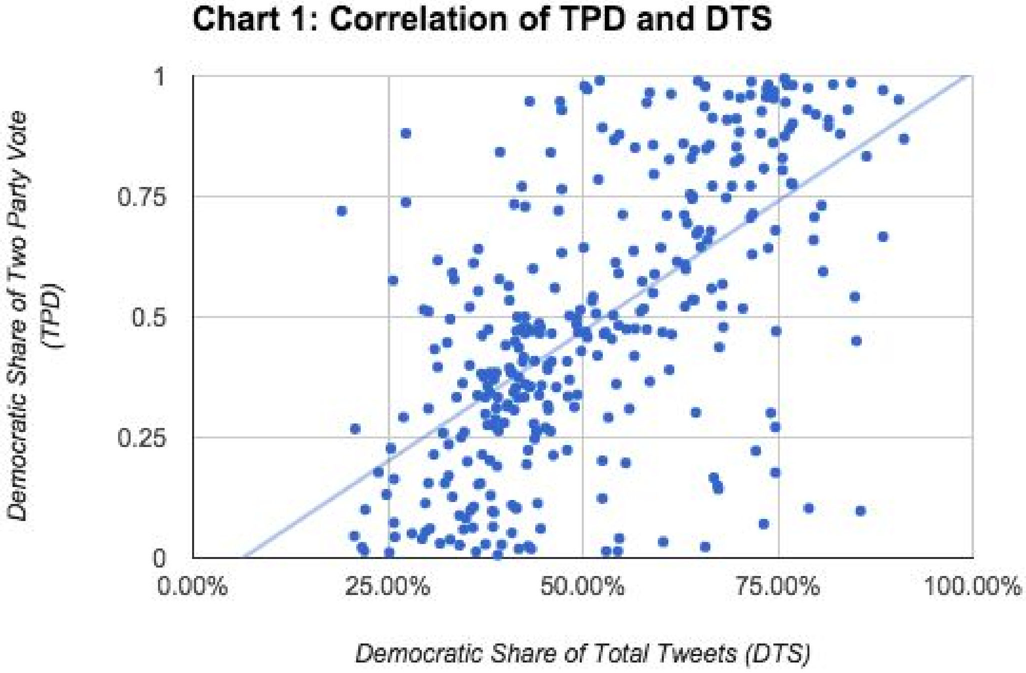
Once the tweets for all of the candidates were obtained from the Topsy database, they were classified by a naïve Bayes classifier. While the complete explanation of how a naïve Bayes classifier is outside the scope of this paper, the basic idea behind it is that it is trained to assign a piece of text a probability of being a part of some category based on the words the piece of text contains. For example, imagine a training set of tweets is fed into a naïve Bayes classifier that contains a bunch of tweets, labeled negative, that contain the word ‘terrible.’ If the positive tweets in the training set never, or rarely, contain the word ‘terrible,’ then the algorithm knows to assign a higher probability of a tweet being negative if it contains the word ‘terrible.’ While it is certainly more complicated than this simple explanation, the algorithm is essentially using statistics to decide what the probability of a phrase being positive or negative is based on what it has been trained to view as positive or negative.

In order to train the naïve Bayes classifier used for this paper, 10% of all the tweets obtained were randomly put into a list for classification. Of these tweets, 310 were arbitrarily selected and classified, by human, as positive, and 310 were classified as negative. 160 positive tweets and 160 negative tweets were then fed into the classifier and the trained algorithm was then tested on the rest of the tweets classified by humans in order to assess the accuracy of the algorithm. In ten samples of 50 tweets, the algorithm achieved a mean accuracy of 69.19% with a standard deviation of 5.2%. While such results are not spectacular by any means, the incorrect classification of about 3 in 10 tweets was not expected to throw off the results of this research if indeed sentiment analysis is an effective tool in predicting elections.

While it should be noted that the accuracy of this classifier could have been improved with more training data (with perhaps thousands of human-classified tweets), the unspectacular accuracy of this classifier on these tweets is in accordance with

Gayo-Avello and other researchers claims that sentiment analysis may not be a great tool for politically themed tweets. As Gayo-Avello importantly points out, “Political discourse is plagued with humor, double entendres, and sarcasm; this makes determining users’ political preferences hard and inferring voting intention even harder,” (2012). Beyond this, while sentiment analysis may be great for something like determining if a review is positive or negative, political tweets often mention both candidates in an election. If a tweet reads, “Vote for Otto Augspach! Marcy Pezzini is a nutjob!” a very clear problem becomes apparent. This tweet will be classified for both candidates because it mentions both of them, but it will only receive one sentiment for both candidates. In other words, if the tweet is classified as positive, Marcy Pezzini will get credit for a positive tweet even though someone just called her nutjob. Thus, the naïvety involved in these classifiers, including their inability to pick out phrases, recognize sarcasm, and assign different sentiments for different names, may pose significant drawbacks in their ability to be classified as accurate data.

To remove any doubts of classification, the second model was created to see if the raw number of tweets about a candidate were indeed a good variable to include in an election model. As can be seen in chart 1, there is a very mild correlation between the democratic share of total tweets and the democratic share of the two party vote (r^2 = .378). This is in discordance with the claims of McKelvey and colleagues that claim a high correlation between tweet share and outcome of elections (2014). Thus, before even testing the efficacy of this variable in predicting elections, the idea of this variable having significance in swaying election predictions looks to be very weak.



Finally, the last piece of data used was the data on net job creation. This data was collected from the United States Census Bureau website. The net job creation rate from 2011 and 2012 were averaged together to create the data for each election. One obvious flaw with this was that about two months of job creation that occur after the election, November and December 2012, are used to predict the election on November 6, 2012. While this was not expected to throw off the results much, it is certainly important to consider that including these holiday months likely gave the net job creation rate in each state a bump due to seasonal job creation.

TABLE 2

Descriptive Statistics

Variable	Mean	Standard Deviation
Y	0.4746	0.5001
INC	-0.1134	0.9118
IDE	0.1731	0.7929
NJC	2.0672	.7461
DPTS	0.4925	0.2860
DTS	0.4889	0.2892

VI. Regression Results

For the first model, which includes a variable for the share of positive tweets received by the democratic candidate, all variables were found to be highly significant (beyond the 99% confidence level) except for the net job creation variable. This variable was also the only variable that did not end up with the expected sign, producing a coefficient of -0.035. This variable was either a very bad choice for estimating voters views of the economy, or may show that economic results play a much larger role in executive branch elections than they do in congressional elections. Future models for predicting congressional elections should definitely avoid this variable.

R^2 for the model came out to 0.6624, signifying that about 66% of the variation in the probability that the democrat will be elected can be explained by this model. As expected, the t-value for the incumbency variable is very high, further confirming the enormous role incumbency plays in congressional elections.

An F test was also carried out to check if the addition of the positive tweets variable was significant. The F test produced a value of 15.56–a significant F value indicating that the variable is appropriate for the model.

The results for the second model were nearly identical to the first, with an R^2 value of 0.664 and the net job creation variable once again being insignificant and producing the wrong sign. Again, an F test was carried out to test the significance of adding the additional variable. The F value obtained was 17.288, again a significant value indicating that the new variable was appropriate to add to the model.

In both models, incumbency proved to be a very significant variable in predicting congressional elections. While ideology showed to be a significant variable and giving a candidate about a 6% boost in chances of being elected in each model, it proved to not play nearly as much of a role as incumbency. This is likely due to many congressional elections deviating from their overall state's ideology. This variable would likely be more suited for a senatorial or gubernatorial election.

Overall, these models do an okay job of predicting election results. The first model predicts 299 of 335 elections correctly, or about 89% of the elections, and the second model predicts 300 of 335 elections correctly, or about 90%.

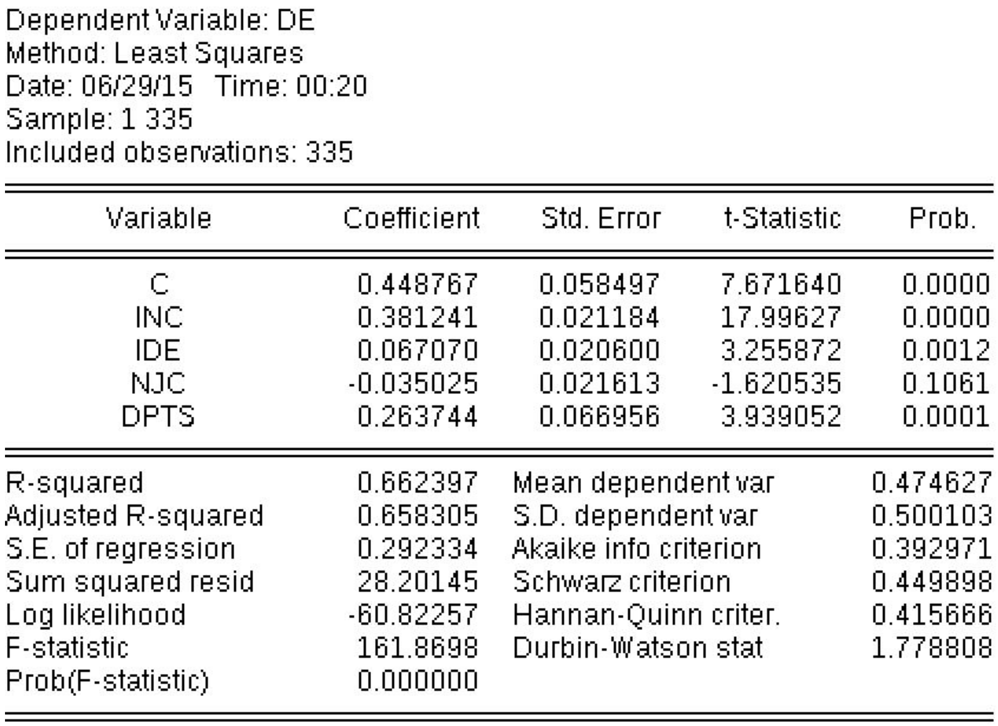
VII. Summary and Conclusions

This research sought to examine the many claims that anyone with a computer and a Twitter account can forecast any election for free. As many researchers have suggested, using Twitter data does not seem to be an effective way to predict congressional elections. Though both models were able to predict nearly 90% of the elections correctly, a model including only a variable for incumbency already predicts about 86% of elections accurately for this 2012 election. If you account for who was in office before those elections where no incumbent is running, that number would go up even higher.

In general, there are just too many issues with this form of data collection for it to be effective in accurately and consistently predicting elections. Computers are not humans, and thus their abilities to naturally process text and classify it are limited–especially in this context. While the variables including Twitter data did indeed indicate to have a significant effect on the models, the effect does not seem large enough to base an entire model around. Additionally, putting aside the efficacy of classifiers, this method of data collection, once again, does not seem scientifically sound. To truly generate accurate poll numbers, the sample needs to be more random and more representative of a population. One person, or spambot, tweeting about their favorite candidate four thousand times should not be able to throw an entire model off. Demographics must also be considered when basing a model around a form of technology. Considering there are many more younger Twitter users than older, and that younger people generally vote more liberally, models based on Twitter posts could easily be biased.

To conclude, election prediction models are probably best left to more standard predictors of elections. Though using sentiments expressed on Twitter to predict elections sounds like a really cool idea, the evidence seems to show that it is not much more than an intriguing headline.

VIII. Appendix A

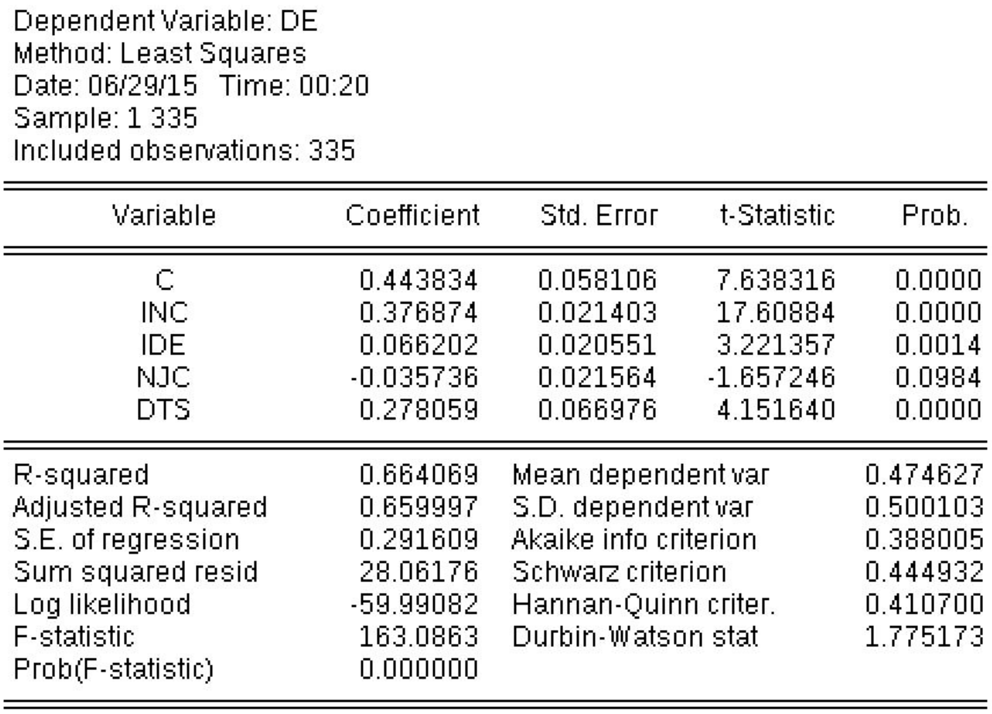


```
Dependent Variable: DE
Method: Least Squares
Date: 06/29/15   Time: 00:20
Sample: 1 335
Included observations: 335

      Variable         Coefficient    Std. Error    t-Statistic      Prob.

          C              0.448767      0.058497      7.671640       0.0000
         INC             0.381241      0.021184      17.99627       0.0000
         IDE             0.067070      0.020600      3.255872       0.0012
         NJC            -0.035025      0.021613     -1.620535       0.1061
         DPTS            0.263744      0.066956      3.939052       0.0001

R-squared                0.662397    Mean dependent var        0.474627
Adjusted R-squared       0.658305    S.D. dependent var        0.500103
S.E. of regression       0.292334    Akaike info criterion     0.392971
Sum squared resid        28.20145    Schwarz criterion         0.449898
Log likelihood          -60.82257    Hannan-Quinn criter.      0.415666
F-statistic              161.8698    Durbin-Watson stat        1.778808
Prob(F-statistic)        0.000000
```



```
Dependent Variable: DE
Method: Least Squares
Date: 06/29/15   Time: 00:20
Sample: 1 335
Included observations: 335

      Variable         Coefficient    Std. Error    t-Statistic      Prob.

          C              0.443834      0.058106      7.638316       0.0000
         INC             0.376874      0.021403      17.60884       0.0000
         IDE             0.066202      0.020551      3.221357       0.0014
         NJC            -0.035736      0.021564     -1.657246       0.0984
         DTS             0.278059      0.066976      4.151640       0.0000

R-squared                0.664069    Mean dependent var        0.474627
Adjusted R-squared       0.659997    S.D. dependent var        0.500103
S.E. of regression       0.291609    Akaike info criterion     0.388005
Sum squared resid        28.06176    Schwarz criterion         0.444932
Log likelihood          -59.99082    Hannan-Quinn criter.      0.410700
F-statistic              163.0863    Durbin-Watson stat        1.775173
Prob(F-statistic)        0.000000
```

IX. Bibliography

Gayo-Avello, Daniel. "No, You Cannot Predict Elections with Twitter." IEEE Internet Comput. IEEE Internet Computing 16.6 (2012): 91-94. Web.

Kagan, Vadim, Andrew Stevens, and V.S. Subrahmanian. "Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election." IEEE Xplore. N.p., 4 Feb. 2015. Web. 20 June 2015.

McKelvey, Karissa, Joseph Digrazia, and Fabio Rojas. "Twitter Publics: How Online Political Communities Signaled Electoral Outcomes in the 2010 U.S. House Election." Information, Communication & Society 17.4 (2014): 436-50. Web.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe. "Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape." Social Science Computer Review 29.4 (2010): 402-18. Web.