Limits of Electoral Predictions Using Twitter

Daniel Gayo-Avello*
Departamento de Informatica
Universidad de Oviedo
dani@uniovi.es

Panagiotis T. Metaxas† and **Eni Mustafaraj**
Department of Computer Science
Wellesley College
pmetaxas, emustafa@wellesley.edu

Abstract

Using social media for political discourse is becoming common practice, especially around election time. One interesting aspect of this trend is the possibility of pulsing the public's opinion about the elections, and that has attracted the interest of many researchers and the press. Allegedly, predicting electoral outcomes from social media data can be feasible and even simple. Positive results have been reported, but without an analysis on what principle enables them.

Our work puts to test the purported predictive power of social media metrics against the 2010 US congressional elections. Here, we applied techniques that had reportedly led to positive election predictions in the past, on the Twitter data collected from the 2010 US congressional elections. Unfortunately, we find no correlation between the analysis results and the electoral outcomes, contradicting previous reports. Observing that 80 years of polling research would support our findings, we argue that one should not be accepting predictions about events using social media data as a black box. Instead, scholarly research should be accompanied by a model explaining the predictive power of social media, when there is one.

Introduction

A substantial amount (22%) of adult internet users were engaged with the electoral campaigns through online social networks during the November 2010 US elections (Smith 2011). This percentage will likely increase, thus, researchers are trying to make sense of the data produced on these channels. It has been reported that the volume of Twitter chatter can be used to predict metrics such as movie success (Asur and Huberman 2010), marketability of consumer goods (Shimshoni, Efron, and Matias 2009), and even the voting results in the 2009 German elections (Tumasjan et al. 2010).

The latter should be surprising given the differences in the demographics of likely voters and social media users (Smith 2011), and it could be that such results were just a matter of coincidence, not easily repeatable.

Our work here aims to test the predictive power of Twitter metrics against several races of the recent US Congressional elections. Our main conclusion is that such predictions so far have proven to be not better than chance, thus, exposing the limits of predictability of elections by means of social media data.

*The authors are listed in alphabetical order.

†Work partially supported by a Brachman-Hoffman fellowship.

"Predicting the Present" with Social Media

The idea that what people are blogging or searching about can provide a glimpse on the collective psyche is very appealing. Since most of the online social media services provide APIs, such data can be collected allowing us to test this hypothesis. Clearly, making predictions from such data would have numerous benefits in the areas of public health (e.g. (Ginsberg et al. 2009), (Lampos, Bie, and Cristianini 2010)), business (e.g. (Asur and Huberman 2010), (Shimshoni, Efron, and Matias 2009)), economics (e.g. (Bollen, Mao, and Zeng 2010), (Choi and Varian 2009)), and politics (e.g. (O'Connor et al. 2010), (Tumasjan et al. 2010)).

Have Social Media Data predicted the Elections?

The promising results achieved by the studies above have created some hype surrounding the feasibility of predicting electoral results from social media. Most of this hype is fueled by traditional media and blogs, bursting prior and after electoral events. Shortly after the recent 2010 elections in the US, bold statements made it to the headlines. From those arguing that Twitter is not a reliable predictor (e.g. (Goldstein and Rainey 2010)) to those claiming just the opposite, that Facebook and Twitter were remarkably accurate (e.g. (Carr 2010)).

Compared to the media coverage, the number of scholarly works on the topic is relatively small, although it tends to support a positive opinion on the predictive power of social media. Thus, according to (Williams and Gulati 2008), the number of Facebook fans constitutes an indicator of candidate viability of significant importance in races of various types (though it reportedly failed to predict a substantial number of races in the US 2010 elections (Facebook 2010)).

(O'Connor et al. 2010) described mixed results: simple sentiment analysis methods on Twitter data exhibited a rather high correlation with the index of Presidential Job Approval, but the correlation was not significant when compared with pre-electoral polls for the US 2008 presidential race. At the same time, the work by (Tumasjan et al. 2010)

directly addressed the question of predicting elections from Twitter data, and concluded with a strong statement, namely that *"the mere number of tweets mentioning a political party can be considered a plausible reflection of the vote share and its predictive power even comes close to traditional election polls."*

The reader may have noted some contradiction in the papers mentioned above. Twitter volume data were supposedly able to predict electoral results in Germany in 2009 with amazing accuracy (Tumasjan et al. 2010), yet more elaborated methods did not correlate with pre-electoral polls in the US (O'Connor et al. 2010). Naive sentiment analysis and frequency counts seem to be enough in one case, but come short in another. Could it be that past promising results were just a matter of chance?

Two very recent studies provide a critical view on this topic. They do not claim that electoral predictions from social data are unfeasible, but that they are more difficult to obtain than one could expect from previous studies. (Lui, Metaxas, and Mustafaraj 2011) focused on the use of Web search volume as a predictor for the 2008 and 2010 US Congressional elections. They found that Google Trends correctly "predicted" one group of races, but that such predictions were neither consistent (i.e. group correctly predicted for 2008 obtained poor predictions for 2010), nor competent compared to predictions by incumbency[1]. In fact, they report that using incumbency as a baseline, Web search volume seems to be a poor electoral predictor.

(Gayo-Avello 2011) applied some of the recently published research methods to a set of tweets obtained during the US 2008 Presidential elections. He found that every published method would have grossly overestimated Obama's victory to the point of predicting a win even in Texas. In addition, he points out that demographic bias is a key caveat when relying on social media data.

More Experiments on Twitter and Elections

We decided to do some more work in order to put to the test the claims of predictability of election results through analysis of Twitter data. For our study, two data sets related to elections in the US during 2010 were employed. Predictions were computed according to Twitter chatter volume as in (Tumasjan et al. 2010) and sentiment analysis as in (O'Connor et al. 2010). Then, the predictions were compared against the actual election results. The mean average error (MAE) was rather high: 17.1% for the Twitter volume and 7.6% for the sentiment analysis. By comparison, MAE for professional polling services is typically about 2-3%.

To understand whether sentiment analysis was really performing better, several tests were carried out including manual labeling of tweets, and comparing the political preference of individual users as inferred by means of sentiment analysis with their presumed political orientation derived from the politicians they followed in Twitter.

[1]Incumbents (the candidates who currently hold office) do remarkably well in elections, getting re-elected about 85% of the time, as shown in (Lui, Metaxas, and Mustafaraj 2011).

Data Sets

The first data set belongs to the 2010 US Senate special election in Massachusetts ("MAsen10"), a race between democrat Martha Coakley and republican Scott Brown (see also (Metaxas and Mustafaraj 2010)). It contains 234,697 tweets by 56,165 different users. It was collected from January 13 to 20, 2010, using the Twitter streaming API, configured to retrieve tweets containing the names of any of the two candidates. The second data set[2] contains all the tweets provided by the Twitter gardenhose in the week of October 26 to November 1, before the general US Congressional elections on November 2, 2010 ("USsen10"). Filtering with the names of candidates for five contested races for the Senate, we found 13,019 tweets by 6,970 different users.

Methods of Prediction

We used the methods by (Tumasjan et al. 2010) and (O'Connor et al. 2010) with some slight changes, in order to account for differences in the nature of electoral races (e.g., German elections had 5 major parties all vying for votes, American elections are based on "winner takes all", so a tweet cannot be counted as a vote for both opposing candidates). So, while Tumasjan et al. seemed to count every candidate (party name) mentioned, we did not include tweets mentioning both competing candidates. With regards to the method by O'Connor et al., we used the same polarity lexicon to find positive, negative, and neutral words. Our modification consists in the following: while they allow a tweet to be both positive and negative, we consider it to be only one of the three options (positive, negative, or neutral) depending on the sum of labeled words. Then, the predicted vote share for a candidate c_1, was computed taking advantage of the bipartisanship nature of the races, using this formula:

$$\frac{pos(c_1) + neg(c_2)}{pos(c_1) + neg(c_1) + pos(c_2) + neg(c_2)} \quad (1)$$

In the Equation 1, c_1 is the candidate for whom support is being computed while c_2 is the opposing candidate; $pos(c)$ and $neg(c)$ are, respectively, the number of positive and negative tweets mentioning candidate c. Notice that neutral tweets were not used, since they don't express a preference for a candidate.

Results of the Prediction Methods

A more detailed analysis was possible for the MAsen10 data set since it contained tweets before the election day (6 days of data), on the election day (20 hours of data), and post-election (10 hours of data). Table 1 shows the number of tweets mentioning each candidate and the election results predicted accordingly. While the total number of tweets (containing post-result tweets) closely reflects the election outcome, the share volume for the pre-election period, incorrectly predicted a win for Coakley. On the other hand,

[2]The Twitter data for the November, 2010 election is courtesy of the Center for Complex Networks and Systems Research at the Indiana University School of Informatics and Computing.

State	Senate Race	Election Result	Normalized Result	Twitter Volume	Sentiment Analysis
MA	Coakley [D] vs. Brown[R]	47.1% - 51.9%	47.6% - 52.4%	53.9% - 46.1%	**46.5% - 53.5%**
CO	Bennet [D] vs Buck [R]	48.1% - 46.4%	50.9% - 49.1%	26.3% - 73.7%	**63.3% - 36.7%**
NV	Reid [D] vs Angle [R]	50.3% - 44.5%	53.1% - 46.9%	**51.2% - 48.8%**	48.4% - 51.6%
CA	Boxer [D] vs Fiorina [R]	52.2% - 44.2%	54.1% - 45.9%	**57.9% - 42.1%**	47.8% - 52.2%
KY	Conway [D] vs Paul [R]	44.3% - 55.7%	44.3% - 55.7%	**4.7% - 95.3%**	**43.1% - 56.9%**
DE	Coons [D] vs O'Donnell [R]	56.6% - 40.0%	58.6% - 41.4%	32.1% - 67.9%	38.8% - 61.2%

Table 2: The summary of electoral and predicted results for 6 highly contested senate races. Numbers in bold show races where the winner was predicted correctly by the technique. Both Twitter Volume and Sentiment Analysis methods were able to predict correctly 50% of the races. In this sample, incumbents won in all the races they run (NV, CA, CO), and 85% of all 2010 races.

	Coakley		Brown	
	#tweets	%	#tweets	%
Pre-elec. (6 days)	52,116	53.86	44,654	46.14
Elec. day (20 hrs)	21,076	49.94	21,123	50.06
Post-elec. (10 hrs)	14,381	29.74	33,979	70.26
Total	87,573	46.75	99,756	53.25

Table 1: The share of tweets for each candidate in the MAsen10 data set. Notice that the pre-election share didn't predict the final result (Brown won by 52%).

applying sentiment analysis and Equation 1 we get a different picture. In this case, pre-election volume seems to be a good prediction for this election.

Both prediction techniques were also applied to five more highly contested senate races from the USsen10 data set (see Table 2). In summary, for the six evaluated races, the winner was predicted correctly only half the time by each method. In one occasion the two different methods agreed correctly, and in an other agreed incorrectly.

Sentiment Analysis Accuracy

The mean average error (MAE) for the predictions was 17.1% for the Twitter volume method and 7.6% for the sentiment analysis method, outside the usual error-margin acceptable in election predictions. The MAE difference between the two methods was intriguing and we decided to study it closer. Evidence on the issues affecting simple polarity-based sentiment analysis methods was examined from three different angles: (1) when compared against human judgment; (2) regarding the detection of misleading propaganda; and (3) on relation to the presumed political leaning of the users posting the tweets.

(1) From the users that had indicated their location in the state of Massachusetts, we selected those with a single tweet. This set contained 2,259 tweets and was manually labeled according to the following criteria: opposing Brown (ob), opposing Coakley (oc), supporting Brown (sb), supporting Coakley (sc), or neutral (n). Then, these labels were compared against those assigned by the automatic method, revealing that the accuracy of the method is only marginally better than a random classifier.

(2) Given this poor result, an additional evaluation was performed on a "Twitter bomb" targeted at Coakley (Metaxas and Mustafaraj 2010). The bomb consisted of a series of tweets spreading misleading information about the democratic candidate. The automatic sentiment analysis only flagged 37% of them as negative. Thus, the subtleties of propaganda and disinformation are not only missed, but even interpreted incorrectly.

(3) An additional experiment was conducted to test the assumption underlying this application of sentiment analysis, namely, that the political preference of users can be derived from their tweets. The presumed political orientation was calculated following the approach described by (Golbeck and Hansen 2011)), in which a user receives the average ADA score[3] of the Congress members he/she is following in Twitter. About half a million Twitter users follow the Congress members with Twitter accounts and little more than 14 thousand also appear in the MAsen10 dataset.

For each of these 14 thousand users four different scores were computed: their ADA score, their opinion on Brown, their opinion on Coakley, and their "voting orientation" for this particular election. The latter three were computed from their tweets using the sentiment analysis method. While we cannot provide details here due to space limitations, we report that, although the different scores correlated as expected, the correlations were weak.

From these experiments we conclude that the accuracy of lexicon-based sentiment analysis when applied to political conversation is quite poor: it just slightly outperforms a random classifier; it fails to flag disinformation and misleading propaganda; and, it's a far cry from being able to predict political orientation of the users.

Limits of Predictions using Social Media Data

Given the negative results we report above, one might suggest that "you would have done better if you did a different kind of analysis". However, recall that we did not invent new techniques of analysis that we used: We simply tried to repeat what others have done and found that the results were not repeatable. We argue that the problem is that, in the past, some researchers have felt comfortable treating social media as a black box: It may give you the right answer, even though you may not know why. We believe that there is an

[3] ADA (Americans for Democratic Action) is a liberal, political think-tank that publishes scores–ranging from 0 (most conservative) to 100 (most liberal)— for each member of the US Congress according to their voting record in key progressive issues.

opportunity for intellectual contribution if their methods are accompanied with at least a basic reasonable model on why they predict correctly. Below we argue why it should not be surprising to find weak or no correlations between social media data and electoral predictions.

Predicting elections is something that professional pollsters have been doing for the last 80 years, a mathematically proven application of correctly identifying likely voters and getting an un-biased representative sample of them. Today's social media do not seem fit to do this. To make this point clear, two arguments are outlined (due to lack of space):

First, we note that the complexity of professional polling cannot be duplicated by sampling social media data. Professional pollsters sample "likely voters" (those who voted in the previous elections), because it has been observed consistently that not every adult who has the right to vote will exercise it. In addition, sample results are age-adjusted because not every age group votes in the same proportion (Blumenthal 2004). There are no means of collecting this information reliably through social media. Even if there were, a really random sample of likely voters is still unattainable, because only those who have decided to express their opinion can be observed.

Second, social media allow manipulation by spammers and propagandists. Fake accounts are easy to create and they can be used to amplify the spammers message polluting the data for any observer (Metaxas and Mustafaraj 2010).

Conclusions

This research has revealed that data from Twitter did no better than chance in predicting results in the last US congressional elections. We argue that this should be expected: So far, knowledge of the exact demographics of the people discussing elections in social media is scant, while according to the state-of-the-art polling techniques, correct predictions require the ability of sampling likely voters randomly and without bias. Moreover, answers to several pertinent questions are needed such as the actual nature of political conversation in social media, the relation between political conversation and electoral outcomes, and the way in which different ideological groups and activists engage and influence online social networks.

Further research is needed regarding the flaws of simple sentiment analysis methods when applied to political conversation. In this sense, it would be very interesting to understand the impact of different lexicons and, even more important, to go one step further by using machine learning, as in the work of (Asur and Huberman 2010); or looking for a deeper understanding of the dynamics of political conversation in social media following the work of (Somasundaran and Wiebe 2010).

References

Asur, S., and Huberman, B. A. 2010. Predicting the future with social media. *CoRR* abs/1003.5699. http://arxiv.org/abs/1003.5699.

Blumenthal, M. 2004. The why and how of likely voters. http://bit.ly/dQ21Xj.

Bollen, J.; Mao, H.; and Zeng, X.-J. 2010. Twitter mood predicts the stock market. *CoRR* abs/1010.3003. http://arxiv.org/abs/1010.3003.

Carr, A. 2010. Facebook, twitter election results prove remarkably accurate. *Fast Company*. http://bit.ly/dW5gxo.

Choi, H., and Varian, H. 2009. Predicting the present with google trends. *Official Google Research Blog*. http://bit.ly/h9RRdW.

Facebook. 2010. The day after election day (press release). http://on.fb.me/hNcIgZ.

Gayo-Avello, D. 2011. A warning against converting social media into the next literary digest. In *CACM (to appear)*.

Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S.; and Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–4. http://1.usa.gov/gEHbtH.

Golbeck, J., and Hansen, D. L. 2011. Computing political preference among twitter followers. In *Proc. of Human Factors in Computing Systems*.

Goldstein, P., and Rainey, J. 2010. The 2010 elections: Twitter isnt a very reliable prediction tool. http://lat.ms/fSXqZW.

Lampos, V.; Bie, T. D.; and Cristianini, N. 2010. Flu detector - tracking epidemics on twitter. *Machine Learning and Knowledge* 6323:599–602.

Lui, C.; Metaxas, P. T.; and Mustafaraj, E. 2011. On the predictability of the u.s. elections through search volume activity. In *e-Society Conference*. http://bit.ly/gJ6t8j.

Metaxas, P. T., and Mustafaraj, E. 2010. From obscurity to prominence in minutes: Political speech and real-time search. In *WebSci10: Extending the Frontiers of Society On-Line*. http://bit.ly/h3Mfld.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proc. of 4th ICWSM*, 122–129. AAAI Press.

Shimshoni, Y.; Efron, N.; and Matias, Y. 2009. On the predictability of search trends. http://doiop.com/googletrends.

Smith, A. 2011. Twitter and social networking in the 2010 midterm elections. *Pew Research*. http://bit.ly/heGpQX.

Somasundaran, S., and Wiebe, J. 2010. Recognizing stances in ideological on-line debates. In *CAAGET '10 Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*.

Tumasjan, A.; Sprenger, T.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proc. of 4th ICWSM*, 178–185. AAAI Press.

Williams, C. B., and Gulati, G. J. 2008. The political impact of facebook: Evidence from the 2006 midterm elections and 2008 nomination contest. *Politics & Technology Review* 1:11–21.