How (Not) To Predict Elections

Panagiotis T. Metaxas, Eni Mustafaraj
Department of Computer Science
Wellesley College
Wellesley, MA, USA
(pmetaxas, emustafa)@wellesley.edu

Daniel Gayo-Avello
Departamento de Informática
Universidad de Oviedo
Oviedo, Asturias, Spain
dani@uniovi.es

***Abstract*—Using social media for political discourse is increasingly becoming common practice, especially around election time. Arguably, one of the most interesting aspects of this trend is the possibility of "pulsing" the public's opinion in near real-time and, thus, it has attracted the interest of many researchers as well as news organizations. Recently, it has been reported that predicting electoral outcomes from social media data is feasible, in fact it is quite simple to compute. Positive results have been reported in a few occasions, but without an analysis on what principle enables them. This, however, should be surprising given the significant differences in the demographics between likely voters and users of online social networks.**

This work aims to test the predictive power of social media metrics against several Senate races of the two recent US Congressional elections. We review the findings of other researchers and we try to duplicate their findings both in terms of data volume and sentiment analysis. Our research aim is to shed light on why predictions of electoral (or other social events) using social media might or might not be feasible. In this paper, we offer two conclusions and a proposal: First, we find that electoral predictions using the published research methods on Twitter data are not better than chance. Second, we reveal some major challenges that limit the predictability of election results through data from social media. We propose a set of standards that any theory aiming to predict elections (or other social events) using social media should follow.

I. Introduction

In recent years, the use of social media for communication has dramatically increased. Research has shown that 22% of adult internet users were engaged with the political campaign on Twitter, Facebook and Myspace in the months leading up to the November 2010 US elections [1]. Empowered by the APIs that many social media companies make available, researchers are engaged in an effort to analyze and make sense of the data collected through these social communication channels. Theoretically, social media data, if used correctly, can lead to predictions of events in the near future influenced by human behavior. In fact, to describe this phenomenon, [2] talk about "predicting the future" while [3] have coined the term "predicting the present". In fact, researchers have reported that the volume of Twitter chat over time can be used to predict several kinds of consumer metrics such as the likelihood of success of new movies before their release [2] and the marketability of consumer goods [4]. These predictions are explained by the perceived ability of Twitter chat volume and Google Search Trends to monitor and record general social trends as they occur.

Being able to make predictions based on publicly available data would have numerous benefits in areas such as health (e.g. predictions of flu epidemics [5], [6]), business (e.g., prediction of box-office success of movies [7] and product marketability [4]), economics (e.g., predictions on stock market trends and housing market trends [3], [8], [9]), and politics (e.g., trends in public opinion [10]), to name a few.

However, there have also been reports on Twitter's ability to predict with amazing accuracy the voting results in the recent 2009 German elections [11] and in the 2010 US Congressional elections [12]. Given the significant differences in the demographics between likely voters and users of social networks [1] questions arise on what is the underlying operating principle enabling these predictions. Could it be simply a matter of coincidence or is there a reason why general trends are as accurate as specific demographics? Should we expect these methods to be accurate again in future elections? These are the questions we seek to address with our work.

The rest of this paper is organized as follows: The next section II reviews past research on electoral predictions using social media data. Section III describes a number of new experiments we conducted testing the predictability of the last two rounds of US elections based on Twitter volume and sentiment analysis. Section IV describes a set of standards that any methodology of electoral predictions should follow in order to be consistently competent against the statistical sampling methods employed by professional pollsters. The final section V has our conclusions and proposes new lines of research.

II. Predicting Past Elections

In the previous section we mentioned some of the attempts to use Twitter and Google Trends for predictions of real world outcomes and external market events. What about the important area of elections? One would expect that, following the previous research literature (e.g. [11], [12]), and given the high utilization that the Web and online social networks have in the US [1], Twitter volume should be have been able to predict consistently the outcomes of the US Congressional elections. Let us examine the instances and methods that have been used in the past in the claims of electoral results predictions and discuss their predictive power.

A. Claims that Social Media Data predicted elections

The word "prediction" means foreseeing the outcome of events that have not yet occurred. In this sense, the authors are not aware of any publications or claims that, using social media data, someone was able to propose a method that would predict correctly and consistently the results of elections *before* the elections happened. What has happened, however, is that on several occasions, *post processing* of social media data has resulted in claims that they might had been able to make correct electoral predictions. Such claims are discussed in the following subsection.

B. Claims that Social Media Data could have predicted elections

Probably due to the promising results achieved by many of the projects and studies discussed in the section I, there is a relatively high amount of hype surrounding the feasibility of predicting electoral results using social media. It must be noted that most of that hype is fueled by traditional media and blogs, usually bursting prior and after electoral events. For example, shortly after the recent 2010 elections in the US, flamboyant statements made it to the news media headlines. From those arguing that Twitter is not a reliable predictor (e.g. [13]) to those claiming just the opposite, that Twitter (and Facebook) was remarkably accurate (e.g. [14]). Moreover, the degree of accuracy of these "predictions" was usually assessed in terms of percentage of correctly guessed electoral races – e.g., the winners of 74% for the US House and 81% for the US Senate races were predicted [15] – without further qualification. Such qualifications are important since a few US races are won by very tight margins, while most of them are won with comfortable margins. These predictions were not compared against traditional ways of prediction, such as professional polling methods, or even trivial prediction methods based on *incumbency* (the fact that those who are already in office are far more likely to be re-elected in the US).

Compared to the media coverage, the number of scholarly works on the feasibility of predicting popular opinion and elections from social media is relatively small. Nevertheless, it does tend to support a positive opinion on the predictive power of social media as a promising line of research, while exposing some caveats of the methods. Thus, according to [16], the number of Facebook fans for election candidates had a measurable influence on their respective vote shares. These researchers assert that "social network support, on Facebook specifically, constitutes an indicator of candidate viability of significant importance [...] for both the general electorate and even more so for the youngest age demographic."

A study of a different kind was conducted by [10]. They analyzed the way in which simple sentiment analysis methods could be applied to tweets as a tool of automatically pulsing public opinion. These researchers correlated the output of such a tool with the temporal evolution of different indices such as the index of Consumer Sentiment, the index of Presidential Job Approval, and several pre-electoral polls for the US 2008 Presidential Race. The correlation with the first two indices was rather high but it was not significant for the pre-electoral polls, and they conclude that sentiment analysis on Twitter data seems to be a promising field of research to replace traditional polls although, they find, it's not quite there yet.

The work by [11] focuses directly on whether Twitter can serve as a predictor of electoral results. In that paper, a strong statement is made about predictability, 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."* In fact, they report a mean average error (MAE) of only 1.65%. Moreover, these researchers found that co-occurrence of political party mentions accurately reflected close political positions between political parties and plausible coalitions.

More recently, [12] used the Tweets sent by the *electoral candidates, not the general public*, and reported success in "building a model that predicts whether a candidate will win or lose with accuracy of 88.0%". While this concluding statement seems strong, a closer look in the claims reveals that they found their model to be less successful, as they admit that "applying this technique, we correctly predict 49 out of 63 (77.7%) of the races".

C. Claims that Social Media Data did not predict the elections

The previous subsection reveals some inconsistencies with electoral predictions in scholarly publications. While candidate counts of Twitter messages predicted with remarkable accuracy electoral results in Germany in 2009 [11], a more elaborated method did not correlate well with pre-electoral polls in the US 2008 Presidential elections [10]. Could it be that some of those results were just a matter of chance or the side-effect of technical problems? Who is right?

The work by [17] focuses on the use of Google search volume (not Twitter) as a predictor for the 2008 and 2010 US Congressional elections. They divided the electoral races in groups depending on the degree they were contested by the candidates, and they find that only a few groups of races were "predicted" above chance using Google Trends – in one case achieving 81% of correct results. However, they report that those promising results were achieved by chance: while the best group's predictions were good in 2008 (81%), for the same group the predictions were very poor in 2010 (34%).

Importantly, even when the predictions were better than chance, they were not competent compared to the trivial method of predicting through incumbency. For example, in 2008, 91.6% of the races were won by incumbents. Even in 2010, in elections with major public discontent, 84.5% of the races, were won by incumbents. Given that, historically, the incumbent candidate gets re-elected about 9 out of 10 times, the baseline for any competent predictor should be the incumbent re-election rate. According to such a baseline, Google search volume proves to be a poor electoral predictor. Compared to professional pollsters (e.g., The New York Times), the predictions were far worse; and, in some groups of races the predictions were even worse than chance!

In [18], the sentiment analysis methods of [10] and [11] are applied to tweets obtained during the US 2008 Presidential elections (Obama vs. McCain). [18] assigned a voting intention to every individual user in the dataset, along with the user's geographical location. Thus, electoral predictions were computed for different states instead of simply the whole of the US, and found that every method examined would have largely overestimated Obama's victory, predicting (incorrectly) that Obama would have won even in Texas. In addition, [18] provides some suggestions on the way in which such data could be filtered to improve prediction accuracy. In this sense, it points out that demographic bias in the user base of Twitter and other social media services is an important electoral factor and, therefore, bias in data should be corrected according to user demographic profiles.

Recently, [19] provided a thorough response to the work of [11] arguing that those authors relied on a number of arbitrary choices which make their method virtually useless for future elections. They point out that, by taking into account all of the parties running for the elections, the method by [11] would actually have predicted a victory for the Piratenpartei (Pirate Party) (which received 2% of the votes but no seats in the German parliament).

In this paper we decided to examine closer the claims of electoral predictions described in the previous subsection. Since we had collected data Twitter data from the US Congressional Elections in 2010, we were in a position to examine whether the methods proposed were as successful in instances other than the ones they were developed for. Moreover, we wanted to analyze *why* would electoral predictions using social media may (or may not) be possible. In the next section III we describe our computational experiments and in section IV we analyze the operating models behind electoral predictions.

III. New experiments on Twitter and elections

For our study, we used two data sets related to elections that took place in the US during 2010. Predictions were calculated based on Twitter chatter volume, as in [11], and then based on sentiment analysis of tweets, in ways similar to [10]. While we did not have comparable data to examine the methods of [12], we discuss some of its findings in the next section.

The first data set we used belongs to the 2010 US Senate special election in Massachusetts ("MAsen10"), a highly contested race between Martha Coakley (D) and Scott Brown (R). The data set contains 234,697 tweets contributed by 56,165 different Twitter accounts, collected with the use of Twitter streaming API, configured to retrieve near real-time tweets containing the names of any of the two candidates. The collection took place from January 13 to January 20, 2010, the day after the elections.

The second data set contains all the tweets provided by the Twitter "gardenhose" in the week from October 26 to November 1, the day before the general US Congressional elections in November 2, 2010 ("USsen10"). The gardenhose provides a uniform sampling of the Twitter data. The daily snapshots contained between 5.6 and 7.7 million tweets. Using the names of candidates for five highly contested races for the US Senate, 13,019 tweets were collected, contributed by 6,970 different Twitter accounts.

These two datasets are different. The MAsen10 is an almost complete set of tweets, while USsen10 provides a random sample, but because of its randomness, it should accurately represent the volume and nature of tweets during that pre-election week.

The first prediction method we examined is the one described by [11], which consists of counting the number of tweets mentioning each candidate. According to that study, the proportion of tweets mentioning each candidate should closely reflect the actual vote share in the election. Tweets containing the names of both candidates were not included, focusing only on tweets mentioning one candidate at a time.

The second prediction method extends the ideas from [10], which described a way to compute a sentiment score for a topic being discussed on Twitter. To that end, [10] relied on the subjectivity lexicon collected by [20] and labeled tweets containing any positive word as positive tweets, and the ones containing any negative word as negative tweets. Then, the sentiment score is defined to be the ratio between the number of positive and negative tweets. It must be noted that, according to [10], the number of polarized words in the tweet is not important, and tweets can be simultaneously considered as positive and negative. In addition, sentiment scores for topics with very different volumes of tweets are not easily comparable. Because of these issues, some changes had to be made to [10]'s approach in order to compute predicted vote shares. In our study, the lexicon employed is also [20], but tweets are considered either positive or negative but not both. Every tweet is labeled as positive, negative, or neutral, based on the sum of such labeled words (positive words contribute +1, while negative words contribute -1). A tweet might be labeled neutral when the sum of polarized words is 0, or when no contributing words appeared in it. Given the two-party nature of the races, the vote share is calculated with this formula:

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

where 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.

A. Results of Applying the Prediction Methods

For the MAsen10 data it was possible to make a more detailed analysis, since the data contained tweets before the election day (6 days of data), the election day (20 hours of data), and post-election (10 hours of data). The 47,368 tweets that mentioned both candidates were not used.

Table I shows the number of tweets mentioning each candidate and the election results predicted from the volume. The total count of tweets we collected (53.25% - 46.75% in favor of Brown) reflects closely the election outcome (Brown 51.9% - Coakley 47.1%). Correct prediction?

	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 I
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 51.9% OF THE VOTES).

	Coakley	Brown
Pre-election	46.5%	53.5%
Election-day	44.25%	55.8%
Post-election	27.2%	72.8%
All	41.0%	59.0%

TABLE II
PREDICTIONS BASED ON VOTE SHARE FOR MASEN10 DATA SET BASED ON SENTIMENT ANALYSIS. THE PRE-ELECTION PREDICTION CORRECTLY PREDICTS BROWN AS THE WINNER WITH A SMALL ERROR (1.1% FOR CORRECTED ELECTION RESULTS, ALSO SEE TABLE III).

We refrained from declaring victory in the predictive power of Twitter when we realized that the share volume for the *pre-election period*, actually predicted a win for Coakley, not Brown. Table I also shows how the number of tweets was affected by electoral events. Brown received 1/3 of all his mentions in the 10 hours post-election, when everyone started talking about his win, an important win that would have repercussions for the health care reform, a major issue at the time. Brown's win broke the filibuster-proof power of democrats in the US Senate and produced a lot of tweets.

While the simple Twitter share of pre-election tweets couldn't predict the result of the MAsen10 election, applying sentiment analysis to tweets and calculating the vote share with Equation (1), comes close to electoral results, as shown in Table II. For a second time in our research effort we refrained from declaring victory in Twitter's power in predicting elections, and decided to take a closer look in our data.

The two prediction methods were further applied to 5 other highly contested senate races from the USsen10 data set. The results of the 6 races are summarized in Table III. The actual results of the election don't always sum up to 100% because in a few races more than two candidates participated. So, in order to calculate the mean average error (MAE), the results were normalized to sum up to 100%. Using the values of the corrected election results, MAE values were calculated for both methods. The Twitter volume method had an error of 17.1%, while the sentiment analysis had an error of 7.6%. In other words, both MAE values are unacceptably high. Each method was able to correctly predict the winner in only half of the races.

B. Sentiment Analysis Accuracy

The result in Table III show that while both prediction methods are correct only half of the time, MAE is smaller for the sentiment analysis method. This difference was intriguing and we decided to study it closer. While a thorough evaluation of the accuracy of sentiment analysis regarding political conversation is out of the scope of this paper, some evidence on the issues affecting simple methods based on polarity lexicons is provided from three different angles:

	POS	NEG	NEUT	Accuracy
opposing Brown	124	76	150	21.71%
opposing Coakley	70	67	105	27.68%
supporting Brown	216	45	254	41.94%
supporting Coakley	240	72	213	45.71%
neutral	249	82	296	47.20%
				36.85%

TABLE IV
CONFUSION MATRIX FOR THE EVALUATION OF THE AUTOMATIC SENTIMENT ANALYSIS COMPUTED AGAINST A MANUALLY LABELED SET OF TWEETS.

1) Compared against manually labeled tweets: To evaluate the accuracy of the above described sentiment analysis method, a set of tweets were manually assigned to one of the following labels: opposing Brown, opposing Coakley, supporting Brown, supporting Coakley, or neutral. This set of tweets was chosen to reflect "one tweet, one vote": From the set of Twitter users that had indicated their location in the state of Massachusetts, we chose users with a single tweet in the corpus. This set contains 2,259 tweets. We read the tweets and manually assigned labels to them. Our labels were compared against those assigned by the automatic method, producing the confusion matrix in Table IV.

The results show that the accuracy of the sentiment analysis is only 36.85%, slightly better than a classifier randomly assigning the same three labels (positive, negative, and neutral).

2) Effect of misleading propaganda: A second evaluation was performed on a particular set of tweets, namely those included in a "Twitter bomb" targeted at Coakley [21] containing a series of tweets spreading misleading information about her. The corpus used in this study contained 925 tweets that were part of such the Twitter bomb. According to the automatic sentiment analysis, 369 of them were positive messages, 212 were neutral, and only 344 were negative. While all of these tweets were part of an orchestrated smearing campaign against Coakley, most of them were characterized as neutral or even positive by the automatic sentiment analysis.

Therefore, we conclude that by just relying on polarity lexicons the subtleties of propaganda and disinformation are not only missed but even wrongly interpreted.

3) Relation to presumed political leaning: Finally, 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. To derive the political preference from the tweets, for every user, the corresponding tweets were grouped together and their accumulated polarity score was attributed to the user. The presumed political orientation of a user was calculated following the approach described by [22]. This approach

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 III

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 84.5% OF ALL 2010 RACES.

makes use of the ADA scores, which range from 0 (most conservative) to 100 (most liberal). ADA (Americans for Democratic Action) is a liberal, political think-tank that publishes scores for each member of the US Congress according to their voting record in key progressive issues. Official Twitter accounts for 210 members of the House and 68 members of the Senate were collected. Then, the Twitter followers of all these accounts were collected, and every user received the average ADA score of the Congress members it was following. The number of Twitter users following the above mentioned 278 Congress members is roughly half a million. A little more than 14 thousand of them also appear in the MAsen10 dataset, and they are used in the following correlation analysis.

For each of these 14 thousand users four different scores are computed: their ADA score which, purportedly, would reflect their political leaning, their opinion on Brown, their opinion on Coakley, and their "voting orientation" for this particular election. The voting orientation is defined as the result of subtracting the opinion on Coakley from the opinion on Brown. Given the range of the ADA scores and the sign of the rest of the scores, the correlations between them should be as follows. The correlation between ADA score and the opinion on Brown should be negative; after all, republicans (closer to 0 in the ADA scale) should value Brown positively, and democrats (closer to 100) should value him negatively. The opposite should be true for Coakley and, thus, a positive correlation should be expected. With regards to the ADA score and the voting orientation they should also be negatively correlated for the same reasons as ADA score vs opinion on Brown.

Table V shows the results of this experiment. The different scores do correlate as expected. However, the correlations are very weak, showing that they are essentially orthogonal with each other.

Based on these three experiments, we claim that the accuracy of lexicon-based sentiment analysis when applied to political conversation is quite poor. When compared against manually labeled tweets it seems to just slightly outperform a random classifier; it fails to detect and correctly assign the intent behind disinformation and misleading propaganda; and, finally, it's a far cry from being able to predict political preference.

	Pearson's r
Opinion on Brown vs Avg. ADA scores	-0.150799848
Opinion on Coakley vs Avg. ADA scores	+0.09304417
Voting orientation vs Avg. ADA scores	-0.178902764

TABLE V

CORRELATION BETWEEN AVERAGED ADA SCORES (WHICH PURPORTEDLY REFLECT USERS POLITICAL PREFERENCE) AND THE OPINIONS ON THE TWO CANDIDATES AND THE VOTING ORIENTATION. THE CORRELATIONS FOUND ARE CONSISTENT WITH THE INITIAL HYPOTHESES BUT VERY WEAK TO BE USEFUL.

C. Could we had done better than that?

The previous subsection reviewed how the methods proposed to predict elections would have performed in several instances using data from the 2010 US Congressional elections. These experiments were important because a wider set of test cases was needed to base any claims of predictability of elections through Social Media.

Given the unsuccessful predictions we report, one might counter that *"you would have done better if you did a different kind of analysis"*. However, recall that we did not try to invent new techniques of analysis: We simply tried to repeat the (reportedly successful) methods that others have used in the past, and we found that the results were not repeatable.

IV. HOW TO PREDICT ELECTIONS

In the past, some research efforts have treated social media as a black box: it may give you the right answer, though you may not know why. We believe that there is an opportunity for intellectual contribution if research methods are accompanied with at least a basic reasonable model on why they would predict correctly. Next we discuss some standards that electoral predictions should obey in order to be repeatedly successful.

A. A method of prediction should be an algorithm.

This might seem as a trivial point, but it is not always easy to follow when dealing with social media. Of course, every election might seem different and adjustments in the data collection and analysis may be necessary. Nevertheless, these adjustments should be determinable before hand, because, as Duncan Watts [23] argues in his recent book, they all seem obvious afterwards.

More specifically, we propose that a method should clearly describe *before the elections*: (a) the way in which the Social

Media data are to be collected, including the dates of data collection, (b) the way in which the cleanup of the data is to be performed (e.g., the selection of keywords relevant to the election), (c) the algorithms to be applied on the data along with their input parameters, and (d) the semantics under which the results are to be interpreted.

The previous section observed that the currently available tools for analyzing large volumes of data are not always accurate. Sentiment analysis can get incorrect readings of sentiment, because the complexity of human communication cannot be easily described completely with a small set of non-contradicting rules. Hoping that the errors in sentiment analysis "somehow" cancel themselves out is not defensible.

B. Social Media Data are fundamentally different than Data from Natural Phenomena.

In particular, Social Media allow manipulation by those who have something to gain by manipulating them. Spammers and propagandists write programs that create lots of fake accounts and use them to tweet intensively, amplifying their message, and polluting the data for any observer. It's known that this has happened in the past (e.g., [21], [24]). It is reasonable that, if the image presented by social media is important to some (advertisers, spammers, propagandists), there will likely be people who will try to tamper with it.

This brings an important point in terms of selecting tools for analysis. Using on social media data the same analytical tools as one would use on data from natural phenomena may not result in repeatable predictions. For example, the social media metrics that *post processing* of candidates' tweets found to increase prediction rates [12], will not likely be the same in the next elections. The candidates in the next elections will certainly manipulate their tweets in a different manner and the metrics that will increase predictability in the next elections (if at all) will be different.

C. Form a testable theory on why and when it predicts.

Predicting elections with accuracy should not be supported without some clear understanding of why it works. If a theory to predict elections is to be identified, the research should be able to explain why this is the case in a testable way, and not treat it as a black box.

Related to this point is the establishment of a baseline for successful predictions. A success rate for elections that is close to chance is not an appropriate baseline, since they are trivial ways of prediction that are much better than that. For example, in 2008, incumbents won 91.6% of the races they run, and in 2010, at a time of reportedly major upsets, the incumbents still won in 84.55% of the races they run. Since in the US congressional elections about nine out of ten of the times the incumbent wins, incumbency success rate is an appropriate baseline (as also [12], [17] propose). Similarly, many electoral districts are known to be consistently electing candidates from the same party for years. Predictions performing below these trivial baselines should not be considered competent.

D. Learn from the professional pollsters.

This last point is not a necessary one, but it is *one way* through which predicting elections through social media could work. In particular, prediction can come through correctly identifying likely voters and getting an un-biased representative sample of them. That's what professional pollsters have been doing for the last 80 years, with mostly impressive results, but that's something that today's Social Media cannot do. Below we describe the complexity of professional polling and explain the reasons why their methods cannot be duplicated by unsophisticated sampling of Social Media data.

Professional polling is based on statistically reliable sampling and is able to prove why it is successful. There is a long history of electoral predictions, and every year significant effort is made all over the world to making sure that predictions are as close to electoral results as possible. Those involved in this endeavor enjoy high visibility, fame when successful and ridicule when not successful. All the experts in the field agree, however, that the most important aspect of correct prediction is the selection of a representative and unbiased sample of the population.

Professional pollsters need to obtain a random sample of the people who will actually vote, in order to achieve accurate predictions. To do that, one needs both a method for random sampling, and access to whoever the random sampling requires to sample. Since one cannot always achieve this, one has to be strive to come as close to this requirement as possible.

Since one cannot be sure about who will actually vote, the prediction can be approximated by sampling those who will likely vote. A typical approach considers the "likely voter", as one who has voted in the previous elections. This is so because not every adult who has the right to vote will exercise it. For example, in the 2000 presidential elections, if one sampled randomly the registered voters – 80% of who actually voted, one would be able to make far more accurate predictions than one that sampled just the eligible adults – 52% of whom actually voted [25]. A good random sampling method should turn out samples of equal number of people to be sampled by age group. However, the final calculation should not include the sample results of each age group without age adjustment. This is because in 2000, only 36% of citizens between the ages 18 and 24 voted compared to 50% of those between 25 and 34 and 68% of those over 35.

Consider then the unfiltered sample which can be obtained today from social media data, such as those provided by Twitter, Facebook, Myspace or other popular social networking services. To be comparable with the results of professional pollsters, a correct sample from Twitter should be able to identify the age range, voting eligibility and prior voting pattern of the tweeters. However, there are currently no means of collecting this information reliably, at least without intrusive methods that compromise privacy. But even then, a really random sample of the likely voters is still unattainable, because only those who have an active Twitter account and have decided to tweet about the election can be observed. Collecting

social media data today, is like going to a political rally and sampling the people gathered there, expecting that it will provide an accurate representation of the likely voters. Instead, a highly biased sample will be found. It would not help much to go to every political rally, because the large volume of voters who attend no rally will still be missing.

V. Conclusions

This research has revealed that data from social media did only slightly better than chance in predicting election results in the last US Congressional elections. We argue that this makes sense: So far, only a very rough estimation on the exact demographics of the people discussing elections in social media is known, while according to the state-of-the-art polling techniques, correct predictions requires 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.

In this paper we have also described three necessary standards that any theory aiming to predict competently and consistently elections using Social Media data should follow: The prediction theory should be an algorithm with carefully predetermined parameters, the data analysis should be aware of the difference between social media data and natural phenomena data, and it should contain some explanation on why it works. We argue that one way to do that, would be to establish a sampling method comparable to the ones used by professional pollsters, though there are many obstacles in doing so today.

In addition to that, 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 to go one step further by using machine learning techniques (such as in the work by [2]). Also, there is a need for a deeper understanding of the dynamics of political conversation in social media (following the work of [26]).

Finally, we point out that our results do not argue *against* having a strategy for involving social media in a candidate's election campaign. Instead, it argues that, just because a candidate is scoring high in some social media metrics (e.g., number of Facebook friends or Twitter followers), this performance does not guarantees electoral success.

Acknowledgment

The Twitter data for the November election was courtesy of the Center for Complex Networks and Systems Research at the Indiana University School of Informatics and Computing. The work of P. Metaxas and E. Mustafaraj was partially supported by NSF grant CNS-1117693..

References

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

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

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

[4] Y. Shimshoni, N. Efron, and Y. Matias, "On the predictability of search trends," *Google Research Blog*, 2009, http://doiop.com/googletrends.

[5] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–4, 2009, http://1.usa.gov/gEHbtH.

[6] V. Lampos, T. D. Bie, and N. Cristianini, "Flu detector - tracking epidemics on twitter," *Machine Learning and Knowledge*, vol. 6323, pp. 599–602, 2010.

[7] G. Mishne, "Predicting movie sales from blogger sentiment," in *In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, 2006.

[8] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 03/2011 2011.

[9] E. Gilbert and K. Karahalios, "Widespread worry and the stock market," in *Proc. of 4th ICWSM*, 2010. [Online]. Available: http://bit.ly/qoz4lh

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

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

[12] A. Livne, M. Simmons, E. Adar, and L. Adamic, "The party is over here: Structure and content in the 2010 election," in *Proc. of 5th ICWSM*, 2011. [Online]. Available: http://bit.ly/q9lSug

[13] P. Goldstein and J. Rainey, "The 2010 elections: Twitter isn't a very reliable prediction tool," *LA Times Blog*, 2010, http://lat.ms/fSXqZW.

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

[15] Facebook, "The day after election day (press release)," Facebook Notes, 2010, http://on.fb.me/hNcIgZ.

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

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

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

[19] A. Jungherr, P. Jürgens, and H. Schoen, "Why the pirate party won the german election of 2009 or the trouble with predictions: A response to "predicting elections with twitter: What 140 characters reveal about political sentiment"," *Social Science Computer Review*, 2011, http://bit.ly/nQU4Zx.

[20] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. of Human Lang. Tech. and Empir. Meth. in NLP*, ser. HLT '05. Stroudsburg, PA, USA: ACL, 2005, pp. 347–354. [Online]. Available: http://dx.doi.org/10.3115/1220575.1220619

[21] P. T. Metaxas and E. Mustafaraj, "From obscurity to prominence in minutes: Political speech and real-time search," in *WebSci10*, 2010. [Online]. Available: http://bit.ly/h3Mfld

[22] J. Golbeck and D. L. Hansen, "Computing political preference among twitter followers," in *Proc. of Human Factors in Comp. Sys.*, 2011.

[23] D. Watts, *Everything Is Obvious: Once You Know the Answer*. Crown Publishing Group, 2011. [Online]. Available: http://bit.ly/q2cUT6

[24] E. Mustafaraj, S. Finn, C. Whitlock, and P. Metaxas, "Vocal minority versus silent majority: Discovering the opionions of the long tail," in *Proc. of IEEE SocialCom*, 2011.

[25] M. Blumenthal, "The why and how of likely voters," Online Blog, 2004, http://bit.ly/dQ21Xj.

[26] S. Somasundaran and J. Wiebe, "Recognizing stances in ideological online debates," in *CAAGET '10*, 2010.