A meta-analysis of state-of-the-art electoral prediction from Twitter data

Daniel Gayo-Avello

University of Oviedo (Spain)

Abstract

Electoral prediction from Twitter data is an appealing research topic. It seems relatively straightforward and the prevailing view is overly optimistic. This is problematic because while simple approaches are assumed to be good enough, core problems are not addressed. Thus, this paper aims to (1) provide a balanced and critical review of the state of the art; (2) cast light on the presume predictive power of Twitter data; and (3) depict a roadmap to push forward the field. Hence, a scheme to characterize Twitter prediction methods is proposed. It covers every aspect from data collection to performance evaluation, through data processing and vote inference. Using that scheme, prior research is analyzed and organized to explain the main approaches taken up to date but also their weaknesses. This is the first meta-analysis of the whole body of research regarding electoral prediction from Twitter data. It reveals that its presumed predictive power regarding electoral prediction has been somewhat exaggerated: although social media may provide a glimpse on electoral outcomes current research does not provide strong evidence to support it can currently replace traditional polls. Finally, future lines of work are suggested.

Keywords

Twitter, Social Media, prediction, forecasting, politics, elections.

Introduction

Real world events have an impact in online systems and trails left by users on such systems have been used to perform early event detection in an automatic fashion. Indeed, a number of well-known studies have

shown that flu (Ginsberg *et al.*, 2009), unemployment rates (Choi & Varian, 2009a), or car sales (Choi & Varian, 2009b) can be forecasted on the basis of Web search queries.

Previously, but in a similar way, blog posts were correlated with book sales (Gruhl *et al.*, 2005) or movie gross incomes (Mishne & Glance, 2006). It must be noted that while queries are not amenable to sentiment analysis, blog posts are and, hence, by applying opinion mining to blogs, stronger correlations can be found –e.g. (Mishne & Glance 2006; Mishne & de Rijke, 2006).

When compared to blogs, microblogging is a recent phenomenon; still, there is a growing number of studies aiming to "take the pulse" of society using such data. The brevity of microposts; the fact that a single service, Twitter, is the *de facto* standard for such kind of publishing; and the ease of collecting data with public APIs; coupled with the promising results obtained in the past by mining query logs and blogs; are probably the main reasons driving the interest of researchers on Twitter data to predict both present and future events.

For instance, Asur and Huberman (2010) exploited Twitter data to predict box-office revenues for movies; O'Connor *et al.* (2010) rather successfully correlated tweets with several public opinion time series; Tumasjan *et al.* (2010) claimed to have predicted the outcome of German elections; Bollen *et al.* (2011) the stock market; and Lampos and Cristianini (2010) the evolution of flu pandemics.

A shallow reading of that growing body of literature could suggest that Twitter data offer remarkable and versatile predictive power; however, this has been already questioned a few times. For instance, Wong *et al.* (2012) raised some doubts on the predictability of box-office performance from Twitter data; and, in a similar way, Jungherr *et al.* (2011) rebutted the main thesis by Tumasjan *et al.*, namely, that tweet volume and votes are strongly correlated.

Those skeptical studies, although a minority, cast reasonable doubts on the feasibility of using Twitter data for predictive purposes or, at least, on the feasibility of applying naïve methods to that end. To start with, it must be acknowledged a strong bias permeating research: The tendency of researchers to report positive results while not reporting negative ones. This bias, the so-called "file drawer" effect (Fanelli, 2010), can be harmful if not accounted for. Firstly, if caution is not exerted, it can be assumed that published (and positive) results are the norm and they are to be expected in the future within similar scenarios to those described in the literature.

Secondly, it makes difficult to publish negative or contradictory results since the burden of proof tends to lie on those criticizing the prevailing (positive) point of view.

Moreover, most of the aforementioned predictions –namely, those involving box-office results, the stock market, epidemics, or opinion polls– are not predicting events but exhibiting correlation between two time series: one obtained from Twitter data and another one observed in the "off-line" world. In that regard, such methods are models and, therefore, they are not proven or disproven; instead, they can be more or less accurate and, hence, more or less useful.

Elections, however, are discrete events; it is true there exist a political continuum where candidates and voters are located but, eventually, voters choose a candidate and cast their vote. The possibility of using Twitter data to predict electoral outcomes has been the subject of a rather substantial amount of research. The aim of this work is to provide a high level picture of that research and some recommendations to push it forward.

As with other predictive areas, there are opposing points of view regarding the predictive power of Twitter data on elections. Most of the debate has centered on methodological issues regarding data collection, vote prediction, and performance evaluation. Hence, some studies claim that simple methods achieve remarkable performance due to the massive amount of data available in Twitter. This author, in contrast, argues that (1) baselines chosen to evaluate performance up to now are not realistic; (2) simple methods achieve inconsistent results when replicated; and (3) their presumed tolerance to noise should not be taken for granted but much better substantiated. Therefore, in this work prior reports are analyzed to:

1. Reach conclusions on the presumed power of state-of-the-art methods for Twitter-based electoral prediction.
2. Point out the main methodological weaknesses in such approaches.
3. Suggest important challenges that still require further work; and
4. Provide recommendations and conclusions for those interested in pushing forward this area.

All of this requires a thorough review of the literature. Nevertheless, more important than providing a bibliography it is to organize those different reports within a coherent conceptual scheme. That scheme is described in a later section. The following one provides a broad picture of current research on the analysis of

Twitter from a political science perspective. That will provide a context for the main topic of this work: namely, electoral forecasting using Twitter data.

Application of Twitter data to political science

As aforementioned, Twitter has attracted plenty of attention from researchers mainly because of its real-time nature and the easiness of data collection; those researchers are exploiting Twitter data to better understand society. However, it is out of the reach of this work surveying such broad line of research; in that regard, the report by O'Connor *et al.* (2011) provides a brief, but highly illustrative, perspective on the current spectrum of social science research working on Twitter data.

This work only focuses on the body of research dealing with electoral prediction from Twitter data. Nevertheless, not every research on the application of Twitter to political science is related to elections; and, in fact, most of the time, politically-charged tweets are not related with elections. Hence, this section provides some background on the application of Twitter data to political science. The list of cited works is representative but necessarily incomplete because, as said, the aim of this section is just to provide context for the field surveyed in this work. Studies on electoral prediction from Twitter data and those dealing with bias or noise in Twitter data are not covered here but in the Appendix.

Needless to say, Twitter has been analyzed from many different points of view within political science; however, the main lines of research relevant for a reader interested in electoral prediction would be the following: public opinion; political leaning and polarization; and impact of Twitter in political discourse and debate.

One of the first studies on the feasibility of exploiting tweets to understand public opinion is the work by O'Connor *et al.* (2010). That paper is later mentioned since, in addition to public opinion, its authors also tried to find a correlation between Twitter data and pre-electoral polls. With regards to public opinion they studied the relation between Twitter data and both consumer confidence, and presidential approval in the US. They applied a simple method of sentiment analysis to obtain a Twitter sentiment index for both consume and presidential approval, finding that both indices exhibited a positive significant correlation with the actual time series. Similar,

but more recent, approaches to obtain a proxy to public opinion from Twitter can be found in the works by Mejova *et al.* (2013), Mitchell & Hitlin (2013), and Ceron *et al.* (2013).

Mejova *et al.* analyzed the perception of candidates taking part in the Republican Party presidential primaries in 2012. They applied state of the art supervised machine learning methods to determine the sentiment of tweets finding that (1) such a method achieved low performance, (2) the prevailing sentiment in Twitter is mostly negative, (3) there is a huge amount of humor and sarcasm specially in negative tweets, and (4) there was little correlation between the evolution of sentiment in Twitter and the actual polls. To explain such results they hypothesized that Twitter users lean towards more liberal political options, or maybe they are younger and, hence, more liberal.

Mitchell & Hitlin, from Pew Research, published a report in which Twitter sentiment regarding different politically charged topics was compared against results from traditional polls and surveys conducted at the same time. Their conclusions were rather shocking: Twitter reaction does not usually agree with that of the general population and, in addition to that, the reaction is not consistent at all. That is, while sometimes Twitter reaction is more liberal than that of the population in other occasions the reaction is more conservative. They disregarded any issue with the proprietary sentiment analysis method they had used and, hence, they attributed those inconsistencies to the fact that there were different communities of users reacting to different topics. In other words, data collected in Twitter exhibits self-selection bias.

Finally, Ceron *et al.* (2013), in addition to trying to predict two different elections in France (more details on this in a later section), exploited Twitter data to rate the popularity of several Italian leaders along 2011. They achieved mixed results: Firstly, opinion in Twitter is more negative than that obtained from polls and, secondly, while for some leaders the correlation between Twitter sentiment and polls was positive (albeit not strong), for other leaders there was no noticeable correlation at all.

So, in short, Twitter has been exploited to infer public opinion with mixed results; methods to automatically infer the sentiment of users are still a problem and, on top of that, the demographic and self selection biases in Twitter data are significant issues too. Needless to say, any improvement in the analysis of public opinion from Twitter data could help to push forward electoral prediction using the same data source.

Another problem approached by different researchers is that of inferring the political leaning of Twitter users (not only individuals but also media outlets accounts). To the best of this author's knowledge Golbeck & Hansen (2011) and Conover *et al.* (2011a) were the first to work on this task. Recent works in the area have been conducted by Barberá (2012), Boutet *et al.* (2012), and Wong *et al.* (2013).

Golbeck & Hansen implemented a simple method that relied on the scores issued by ADA (Americans for Democratic Action). ADA is a think-tank that publishes scores for each member of the US Congress on the basis of their voting record. Those scores range from 0, the most conservative, to 100, the most liberal. Given that many members of the Congress have got an official Twitter account, other Twitter users were scored by averaging the ADA scores of those representatives they were following in Twitter. Hence, this method does not rely on the published tweets but in the relationships between individuals and their representatives. An obvious problem is that external information about the political leaning of the representatives is required.

Conover *et al.* analyzed both content based and graph based methods to infer the political leaning of Twitter users. They found that supervised methods trained over labeled tweets exhibited pretty high accuracy. However, these results were outperformed when community detection methods were applied to the retweet graph.

Barberá (2012) also relied on the user graph to solve the same problem. Assuming that relationships among users are mostly homophilic he studied whether the structure of the social network of users in Twitter can be used as a source of information about their political leaning. Hence, this method is not based on content but in the relationships among users and accounts of political actors in Twitter. After applying his method to Twitter data belonging to Spanish users Barberá found that parties' official accounts were correctly classified and that political leaning inferred for individual users also exhibited high accuracy.

Bouted *et al.* (2012) inferred users' political leaning on the basis of the degree of tweets and retweets each user publishes regarding a given party account (in this case in the UK). According to them, some of the methods achieved 86% accuracy without any training. The approach by Wong *et al.* (2013) was rather similar but, instead of relying just on tweets mentioning political parties, they exploited tweets focusing on many different events for which a sentiment score, and a political leaning depending on that score, were assigned. Hence, depending on the sentiment users expressed for those events they were assigned a given political leaning.

To sum up, works regarding political leaning of Twitter users generally report high accuracy. Hence, they should help to understand the degree of support a given party (or ideology) has got in Twitter but also the possible ideological bias in Twitter when compared against the whole population. It must be noted, however, that the fact that a given user exhibits a given political leaning does not imply a vote in a given election.

Closely related to these works are several studies on the polarization of Twitter users according to their political preferences; those by Conover *et al.* (2011b) or Feller *et al.* (2011) are a representative sample. According to these reports, users tend to follow and retweet like-minded people although they discuss with a much more diverse set of individuals. Such features, in addition to political leaning prediction, could help to improve electoral prediction methods based on Twitter data.

Needless to say, there are aspects of politics and Twitter not covered by this shallow survey; for instance, its usage by political actors; its impact during electoral campaigns; or its role in the public sphere, especially for grassroots initiatives such as those seen during the Arab Spring, the Indignants protests or the Occupy movement. These and other perspectives are important but not highly related to electoral forecasting; thus, they are not explored in this work.

Characterization of Twitter-based electoral prediction methods

As aforementioned, Twitter data has been mined to determine the public opinion on several topics and pre-electoral and electoral polls have been studied as part of that public opinion. The number of reports in this issue has grown considerably and most of them claim from promising to positive results. However, most of the evidence provided by such reports is rather weak and, therefore, this work aims to analyze and contrast such evidence to, firstly, point out the weaknesses in current research and, secondly, to suggest further lines of work.

Although unstated in most of the studies, it is commonly implied that any method to predict electoral results from Twitter data is an algorithm. Such algorithms are devised as a pipeline which starts with the collection of data from Twitter, goes on processing that data, and finishes with a prediction which needs to be evaluated against the actual results of the elections. Needless to say, the algorithms can be parameterized to adapt to different scenarios, and predictions can be more or less detailed (for instance, the algorithms can just provide

the winner or vote rates for different candidates). Thus, there are a number of features defining any method to predict electoral results from Twitter; namely:

1. Period and method of collection: i.e., the dates when tweets were collected, and the parameterization used to collect them.
2. Data cleansing measures:
    a. Purity: i.e., to guarantee that only tweets from prospective voters are used to make the prediction.
    b. Debiasing: i.e., to guarantee that any demographic bias in the Twitter user base is removed.
    c. Denoising: i.e., to remove tweets not dealing with voter opinions (e.g. spam or disinformation) or even users not corresponding to actual prospective voters (e.g. spammers, robots, or propagandists).
3. Prediction method and its nature:
    a. The method to infer voting intentions from tweets.
    b. The nature of the inference: i.e., whether the method predicts individual votes or aggregated vote rates.
    c. The nature of the prediction: i.e., whether the method predicts just a winner or vote rates for each candidate.
    d. Granularity: i.e., the level at which the prediction is made (e.g. district, state, or national).
4. Performance evaluation: i.e., the way in which the prediction is compared with the actual outcome of the election.

It must be noted that not every feature in the previous needs to appear in every method. For instance, up to now, all of the research on electoral prediction from Twitter data has worked on tweets related to the parties and candidates taking part in the elections; however, it could be conceivable to use tweets related to broader topics (e.g. health system, unemployment, or taxes), or even the whole Twitter stream for a given country or region to predict elections.

In a similar way, many researchers have not applied data cleansing measures arguing that, with enough data, Twitter-based methods should be noise tolerant. Such an approach does not imply, however, that data cleansing measures are to be removed from the conceptual scheme; indeed, not applying any of such measures

could be an option regarding that feature. In addition to that, it must be noted that, up to now, none of the methods using raw tweets without removing spam, robots, or astroturfers has provided strong proof of being noise tolerant. Certainly, for each of such methods there has been one report claiming positive results but, as it will be shown, successive papers replicating those methods have not achieved comparable results. Of course, this does not imply that noise tolerant methods are unfeasible but, instead, that noise tolerance should be proven through carefully devised experiments.

So, in short, any method to predict elections from Twitter data should be studied according to four different aspects: (1) the data collection approach –which could be focused on candidates and parties or could be much broader; (2) the approach taken to deal with noise –i.e. removing most of it or accepting its presence; (3) the method of prediction and the way in which that prediction is provided –e.g. giving the winner of the race, the vote share, or the number of seats achieved; and (4) the way in which the method is evaluated.

Consequently, all of the papers analyzed in this work have been characterized according to such a scheme. Table 2 shows that characterization and it includes descriptive information for each of the elections. Following sections provide further details for each of the methods in the literature with regards to each of the features in the scheme.

General characteristics of research conducted up to date

The first thing that Table 2 reveals is that literature regarding electoral prediction has not actually made any prediction given that all of the reports were written *post facto*. Therefore, those studies claiming positive or promising results are, in fact, describing how certain elections could have been predicted.

Moreover, authors describing positive results do not usually replicate or improve their method in successive studies. Fortunately, as it will be shown, only two different approaches to voting inference in Twitter have been widely used –namely, tweet counting and lexicon-based sentiment analysis– and, thus, a number of papers have evaluated and compared both. That will help to draw some conclusions about their respective performance.

In addition to that, a number of papers deal with the same elections. O'Connor *et al.* (2010) and Gayo-Avello (2011) covered the US presidential election in 2008. And Tumasjan *et al.* (2010) and Jungherr *et al.* (2011) covered the German federal election in 2009. By analyzing those papers it should be possible to determine if consistent results can be obtained from Twitter data or if, conversely, results depend on decisions made by the researchers when choosing the method's parameters.

The rest of papers correspond to single case scenarios; among such works the research conducted by Metaxas *et al.* (2011) deserves special attention since it covers six different races from the same elections in the United States. That paper hypothesizes that positive results achieved by the simplest methods used up to now could be due to mere chance.

Regarding the elections studied in the literature, the United States is the best covered scenario (4 papers), followed by Germany (2 papers); Ireland, Singapore, Netherlands, and France are covered by one paper each.

Method of collection

Twitter offers two methods for data collection: the Twitter Search API and the Twitter Streaming API. The first one is far less common nowadays although it was the only choice before the Streaming API was available. There are two main disadvantages when using the Search API: First, results are retrieved using a sliding window which means that it is not possible to retrieve data from any arbitrary date but just a few days before the current one. The second issue is that the Search API has been devised to power Twitter's search engine and, thus, it has a bias towards those users considered by the algorithm as more relevant.

Because of those problems and other technical issues, the Streaming API is the usual choice of researchers. It must be noted that the whole stream of tweets (the so-called Firehose) is not publicly available and that the tweets appearing in the public (and free) API are just a sample of the whole stream. According to Twitter the sample is statistically representative but this has been questioned (Morstatter *et al.* 2013). It must be noted that the Streaming API provides content in real time, not historical data.

No matter the API of choice, there are some parameters that need to be set before collecting any data such as keywords or geographical coordinates. The later are optional and allow the API user to delimit a geographical

area from which tweets would be obtained, this would be obviously an advantage in order to predict electoral results but, as it will be later discussed, it is seldom used. Keywords, however, are not optional and, in fact, they are crucial to determine the precision and recall of the collected dataset. Most of electoral predictions using Twitter data have employed names of candidates and parties as keywords. It must be noted that, in general, researchers provide very little information about the keywords chosen to compile their datasets. In addition to that, very little attention is paid to the precision of the collected data or to alternative ways to improve recall.

Period of collection

Table 2 reveals important differences among studies with regards to some of the features in the scheme. For instance, the period of data collection varies widely. Table 1 shows that some studies collected data just one week before the elections while others collected data for weeks, months or even years before the event.

There is consensus, however, about the ending point for the period of collection: the day before elections. The only paper not following this convention is that by Tumasjan *et al.* (2010) and, as it will be shown, this was a matter of later criticism by other researchers (Jungherr *et al.*, 2011).

It must be noted that it is unclear the impact of the period of collection in the predictions: Jungherr *et al.* (2011) showed that by using different time windows performance underwent significant variations. At the same time, Metaxas *et al.* (2011), using just one week of data, obtained both correct and incorrect predictions. Yet, none of those researchers draw any conclusion regarding the nature of the impact exerted by the size of the time window chosen for the data collection. Hence, further research in this matter is needed to find compelling criteria to choose an appropriate period of collection. In the absence of such criteria researchers should refer to literature in traditional electoral forecasting (cf. Lewis-Beck & Rice 1992 and Campbell & Garand 2000).

<< Table 1 about here >>

<< Table 2 about here >>

Data cleansing

Data cleansing refers to measures adopted to filter out or weight tweets in order to improve the accuracy of the prediction. Needless to say, depending on the properties of such filtering there would be a continuum of

methods. At one extreme there would be those methods applying no filtering or weighting; such methods usually claim to be tolerant to both noise due to spam, astroturfing or misleading propaganda, and to bias in the Twitter user base. As it was aforementioned, besides the published positive (or promising) results in single cases, there is little proof in the literature supporting the idea that such methods are indeed noise and bias tolerant; especially, given that attempts to replicate such methods have usually achieved poorer results than those claimed in the original papers. However, this does not imply that noise and bias tolerant methods should not be studied in the future.

At the other extreme there would be those methods trying to replicate as close as possible traditional polling methods. Such methods would adopt measures to ensure (1) data purity, i.e. the collected data only comprises those users who are prospective voters, (2) noise removal, i.e. the collected data only comprises those tweets dealing with the electoral process; and (3) bias correction; i.e. any demographic bias in the Twitter user base is weighted accordingly to the voting population.

The first measure involves those decisions adopted to select Twitter users that are likely voters in the election of interest. It must be noted that such information is unavailable and, hence, researchers can, at best, limit the data collection to those users located in the area of interest for a given election. Such a method does not take into account those users not providing a valid location or emigrants eligible to vote; however, it certainly filters out those users expressing their views on the campaign without being eligible to vote. This approach to data purity can be implemented by collecting just geolocated tweets or by checking the location string of the users in the collection.

It must be noted that, despite its simplicity, only two studies in the literature have applied such a measure: Gayo-Avello (2011) by relying on tweets geolocated in counties of interest, and Skoric *et al.* (2012) by limiting the dataset to those users located in Singapore. It is possible that the fact that many Twitter users do not geolocate their tweets or provide a valid location (Hecht *et al.* 2011) have led researchers to not apply this method so they can obtain larger collections.

Needless to say, there are certain scenarios in which geolocation can be obtained by indirect means: such as those cases in which voters speak a non global language. For instance, (Tumasjan *et al.*, 2010; Jungherr *et al.*,

2011; Tjong Kim Sang & Bos, 2012) ensured to a certain extent the purity of their collections on the basis of language use. The first two papers deal with German elections and, thus, it is very likely that tweets about German parties and politicians written in German are produced by German users and not Austrian or Swiss users. In the same way, the third study deals with Dutch elections and the data is probably originated in Netherlands and not Belgium. Unfortunately, when considering more global languages (as English, for instance) or elections of worldwide interest (e.g. elections in the U.S., France or U.K.) such an approach is no longer valid and geolocation should be incorporated.

The second measure to clean the data is denoising. This includes any post-processing of the dataset to remove tweets or users not dealing with the electoral process. In other words, it implies the removal of spam, rumors, propaganda, disinformation and users mainly producing noisy tweets. Table 2 reveals that no paper in the literature has adopted such measures although a few of them acknowledge the problem. For instance, Metaxas *et al.* (2011) made this warning:

"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. 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."

In addition to that, these researchers conducted an experiment to check the robustness against such kind of manipulation of the sentiment analysis methods most common in this kind of studies. They found that:

"[B]y just relying on polarity lexicons the subtleties of propaganda and disinformation are not only missed but even wrongly interpreted."

At this time, a key issue must be introduced: most of the research regarding Twitter-based electoral predictions is not using state of the art sentiment analysis. When applied, sentiment analysis methods are rather primitive and, hence, results are extremely sensible to noise. Needleless to say, better methods of sentiment analysis should be applied in the future but, anyway, it should be checked if they are tolerant to the noise in political tweets or measures to remove such noise would be required.

The third and last data cleansing measure is debiasing. Twitter's user base is not a representative sample of the population, and that problem can be tackled with by determining the demographic strata users belong to and weighting their tweets accordingly. The low representativeness of Twitter has been widely discussed. For instance, (danah boyd, 2010) wrote the following:

"Big Data presents new opportunities for understanding social practice. Of course the next statement must begin with a 'but.' And that 'but' is simple: […] just because you have a big N doesn't mean that it's representative or generalizable."

Besides, even in the United States Twitter use is minor (11% of Americans) and their users are "overwhelmingly young" (Lenhart & Fox, 2009). This low representativeness is a major problem because dominating demographic groups may tilt toward a few selected political options (Smith & Rainie, 2008), and such a leaning heavily distort results (Gayo-Avello, 2011).

Again, two approaches are possible: acknowledging the bias in the user base but arguing the method is bias-tolerant, or assuming that bias-tolerant methods are not feasible and trying to weight each user and tweet according to the demographic composition of the voting population. As with claims about noise tolerance there is little evidence in the literature that methods applied to Twitter data are tolerant to bias. The main argument in favor of such tolerance is the achievement of positive results even when ignoring bias; however, as it has been said, such an argument is rather weak and is based on single case results.

Unfortunately, results obtained when debiasing Twitter data according to the actual population are not really conclusive, mainly because, up to now, only two studies have attempted such an approach. The way in which such debiasing methods have been applied is as follows: First, demographic information about Twitter users is obtained; then, tweets from each demographic group are weighted according to prior knowledge about their electoral involvement. Needless to say, the first task is not simple; unlike other services, such as Facebook, Twitter profiles do not include structured information: There is no way to indicate the user's sex or age and, instead, profiles consist of free text fields for name, location, website, and biography. Nonetheless, it is not

unsolvable and in the final section some references on this matter are commented. Indeed, as it has been aforementioned, two papers have tried some kind of debiasing (see Table 2).

Gayo-Avello (2011) was able to obtain the age for about 2,500 users in his dataset by crossing their full names and county of residence with online public records. This way he found that the dataset was dominated by users in the 18-44 age interval. Then, by weighting their tweets according to age participation in the 2004 elections he was able to reduce the error from 13.10% to 11.61% –a significant boost in performance.

Tjong Kim Sang & Bos (2012) tried a different approach: debiasing the data according to the presumed political leaning of the population. Certainly, such information is extremely relevant, especially if a "shy-Tory" effect[i] is suspected. Unfortunately, their results were not conclusive since (1) the authors had to rely on pre-electoral polling data which could be seen as overfitting; and (2) the performance of the method when debiasing was no better than a simpler method based on tweet counts.

Hence, on one hand, those methods ignoring bias in Twitter data have not actually proven they are bias-tolerant and further research is needed in that line of work. On another hand, just two attempts to partially debias Twitter data have been conducted with inconclusive results. In the first case, correcting the data according to the age of the users improved the accuracy. In the second case, on the contrary, correcting the data according to the presumed political leaning of users did not improve the prediction.

So, in short, most of the methods attempting to predict elections based on Twitter data have not clean the collected tweets at all. Researchers working on raw unfiltered data acknowledge the presence of noise and biases in their datasets but claim that, given their positive or promising results, the simple methods they applied are noise and bias tolerant. In contrast, other researchers have shown that it is possible to limit the data collection to those regions of interest for a given election, that primitive sentiment analysis methods are very sensitive to noise and that, although somewhat feasible and sensible from a theoretical point of view, removing bias from Twitter data has only partially demonstrated its usefulness.

Method of prediction

Inferring votes by counting tweets

Up to now, two main methods have been used to infer votes from tweets. The first one, originally proposed by Tumasjan *et al.* (2010), consists of merely counting the tweets mentioning a given candidate or party: the larger the number of tweets, the larger the vote rate. That method is appealing for many reasons: it is simple to implement, it can be applied in near real-time, and it can be used both to obtain aggregated vote rates and to infer voting intentions for individuals (i.e. the candidate a user is mentioning the most would be his or her chose). Besides, Tumasjan *et al.* (2010) claimed the method exhibited very good performance:

"The mere number of tweets reflects voter preferences and comes close to traditional election polls."

Indeed, Table 2 shows they reported an error of 1.65% for the German federal elections in 2009. Jungherr *et al.* (2011) later criticized some of the decisions taken by Tumasjan *et al.* (2010), especially those regarding the selection of parties, and the period of data collection. As it has been discussed, the selected time window has an impact on predicted results; using the same time window employed by Tumasjan *et al.*, Jungherr *et al.* (2011) found error values that were in the order of that reported by Tumasjan *et al.* (2010). However, when they used a time window ending at the election day (the most plausible decision and the most common in the literature) they found an error of 2.13% which is substantially larger than both the original report of Tumasjan *et al.* (2010) and traditional polls.

It has been already stated that further research is needed with regards to periods of data collection but, anyway, there are no reasons to suppose that tweet counting can be more sensitive to this issue than methods relying on sentiment analysis. Arguably, it is the selection of the candidates and parties to monitor the key parameter of methods based on tweet counts. The analysis of the Pirate Party case by Jungherr *et al.* (2011) showed that considering all the parties running for election even though they have not prior representatives achieves very different results. Unfortunately, there are no more studies in this regard: all of the papers applying tweet counting have monitored only major parties.

At this point it would seem that there are no conclusive answers to the question of whether tweet counts are a good predictor of voting intention for major parties or not; after all, half of the papers using such simple method have correctly predicted the winner of the elections. However, that global performance alone is not highly impressive and, moreover, common arguments in favor of those methods are rather weak: the most frequent is summarized in the phrase *"there is no such thing as bad publicity"*, while the other is the usual argument about those methods being noise tolerant. Asserting that polarity in political opinions does not matter at all or that, in some way, positive and negative opinions about every party and candidate cancel each other so that only raw counts matter are too simplistic reasons to be accepted at face value.

In this regard, an experiment conducted by Gayo-Avello (2011) is quite informative. He collected data from an informal opinion-poll conducted during the US presidential election in Twitter. A website called TwitVote[ii] asked users to declare their votes with a tweet tagged with the hashtag *#twitvote*. By collecting those tweets published in election day, he was able to find the actual votes for a number of users. With such data he was able to compute the precision of both analyzers (i.e. the tweet counting method and the lexicon-based sentiment analyzer) when inferring votes from tweets published by those users whose eventual vote was known. In addition to that, both methods were compared against a perfectly informed random classifier: one assigning voting intention with regards to the proportion of "votes" according to TwitVote.

This way, he found not only that the lexicon-based method outperformed the tweet counting method, but also that tweet counting underperformed the random classifier for both candidates: slightly for Obama and by a huge margin for McCain. The lexicon-based sentiment classifier, however, outperformed the random classifier with regards to both candidates.

Certainly, it could be argued that users in the dataset by Gayo-Avello and those in TwitVote are different enough to justify the apparent underperformance of the tweet counting method when compared against the random classifier. In this regard it must be noted that, according to Gayo-Avello, (1) when taking into account those users appearing in both datasets the performance of the tweet counting method is still below that of a lexicon-based approach, and (2) the vote share in both datasets is not that different: 86.6% of users in both datasets "voted" for Obama compared to 85.9% of users in TwitVote who "voted" for him. Therefore, it seems

rather plausible to consider vote sharing in TwitVote as a reasonable baseline against which to compare voting inferring methods for that election.

So, in short, tweet counting does not seem to outperform the simplest of sentiment analysis methods nor a random classifier. Moreover, only half of the races predicted by counting tweets achieved positive results, thus, making the hypothesis by Metaxas *et al.* (2011) that such results are due to mere chance rather plausible.

Inferring votes with sentiment analysis

The other popular method to infer voting intentions from tweets is sentiment analysis. The name is misleading because despite the extensive research conducted in that field (cf. Pang & Lee 2008 or Liu 2012) most of the studies on electoral prediction have relied on the simplest methods.

Except for Bermingham & Smeaton (2011), Tjong Kim Sang & Bos (2012), and Ceron *et al.* (2013) who applied supervised sentiment analysis –with mixed results, it must be said– the rest of studies have relied on lexicons to determine the polarity of tweets. O'Connor *et al.* (2010) were the first using that method. They relied on the lexicon by Wilson *et al.* (2005) which consists of a list of terms labeled as positive or negative. Thus, tweets can be scored one way or the other, or even assigned both scores. Like the method based on tweet counts this one is also appealing because of its simplicity; but like the previous method it is also unsatisfactory. O'Connor *et al.* already found many examples of incorrectly detected sentiment although they argued that:

"With a fairly large number of measurements, these errors will cancel out relative to the quantity we are interested in estimating, aggregate public opinion."

Using a noisy measurement instrument is not a problem *per se* but it can be problematic if the classifier is producing different amounts of errors for each candidate. In this regard, the experiment on TwitVote described in previous subsection is again of interest. In that experiment it was found that the performance of a lexicon-based classifier similar to that used by O'Connor *et al.* (2010) was very different depending on the candidate: precision for Obama was rather high (88.8%) but very poor for McCain (17.7%). Hence, although the method outperforms a random classifier it is highly unbalanced and, therefore, it is not very realistic to simply expect errors to cancel out when aggregating results.

Additional experiments with lexicon-based methods were conducted by Metaxas *et al.* (2011) confirming, first, that their performance is only slightly better than that of a random classifier; and second, that misleading information and propaganda are missed or wrongly interpreted as candidate support.

In short, polarity based methods employed up to date:

1. Miss the subtleties of political language.
2. Exhibit very poor performance and,
3. Produce unbalanced results making unrealistic to accept that errors will cancel out when aggregating data.

At this point it must be noted again that the main problem is that, up to now, sentiment approaches adopted when predicting elections from Twitter data have been unnecessarily simple and not representative of the state of the art in sentiment analysis. In other words, most of the research up to this day has been approached in a tentative way and, therefore, the sentiment analysis applied was rudimentary and achieved subpar accuracy. Arguably, applying state of the art methods much better predictions could be achieved and, hence, the adaptation of better methods of sentiment analysis is still an open challenge in this field of research.

Performance evaluation

Evaluation measures

The final question regarding electoral prediction methods is how to evaluate them. Needless to say, the actual outcome of the elections is needed but there is little consensus in the literature about which information to consider as such outcome. Some researchers have only considered the winner of the election without any other consideration; others have considered the number of seats achieved in congress or senate, while others have considered the actual vote sharing.

Moreover, except for local elections, predictions can be computed and evaluated at different levels (e.g. national, state, or district) and, besides, depending on the subtleties of the electoral system the popular vote may differ widely from the number of seats achieved or even the final results of the elections (think for instance of the Electoral College in the US or the impact of the D'Hondt method).

Current research has produced predictions mainly at national level (Tumasjan *et al.*, 2010; Jungherr *et al.*, 2010; Bermingham & Smeaton, 2011; Skoric *et al.*, 2012; Tjong Kim Sang & Bos, 2012; Ceron *et al.* 2013) with a couple of papers focusing at state level (Gayo-Avello, 2011; Metaxas *et al.*, 2011).

Those predictions have been evaluated against vote rates (Gayo-Avello, 2011; Tumasjan *et al.*, 2010; Jungherr *et al.*, 2011; Metaxas *et al.*, 2011; Bermingham & Smeaton, 2011; Skoric *et al.*, 2012; Ceron *et al.* 2013); against number of seats (Tjong Kim Sang & Bos, 2012); and also as dichotomous decisions (Metaxas *et al.*, 2011).

When predicting vote rates MAE (Mean Absolute Error) is normally used to measure accuracy. This is not necessarily the best option, especially since MAE values are not comparable across different elections. For instance, the Senate election in Kentucky was correctly predicted with a MAE of 39.6% while a MAE of 6.3% produced an incorrect prediction in California (see Table 2). Despite of this obvious problem, MAE is a pretty well known measure which allows researchers to compare their method's performance against that of pre-electoral polls for a given election. Future research should look for better measures than MAE to evaluate performance of this kind of methods, especially across different elections. In this regard, researchers could start with measures such as R-squared, Standard Error of Estimate (SEE), or Root Mean Squared Error (cf. Lewis-Beck 2005).

Another common way to measure of performance is the number of correctly guessed races or, simply put, predicting the winner. This approach is appealing because of its simplicity but it can be misleading since no information is provided about how far or close the prediction was from the actual results; besides, it greatly depends on the granularity of the prediction. For instance, Gayo-Avello (2011) predicted a Obama victory in the US which, according to this performance criterion, was correct; however, that prediction included a victory in Texas, which was obviously wrong. By just changing the granularity level and the evaluation method, a negative result could have been turned into a promising one; it seems reasonable to argue that future research requires stronger evaluation criteria so that the same results cannot admit two opposing readings.

In addition to that, in the following subsection it will be discussed that predicting the winner is, most of the times, not a great achievement on itself and that a credible baseline must be provided for comparison.

Moreover, experts in traditional electoral forecasting also advise against using the winner as a way to judge a given prediction (Campbell 2004); indeed, the desirable outcome for any electoral prediction should be the vote percentage.

On the appropriateness of baselines

The two usual ways of reporting performance of a Twitter-based electoral prediction method have a common issue: the lack of reasonable baselines against which to compare in order to report relative performance. This was firstly stated by Metaxas *et al.* (2011) that, with regards to winner prediction, suggested that:

"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."

Under the light of such assumption, it seems clear that a method that correctly guesses 50% of the races in one election is quite far from competent. With regards to vote rate prediction this author is not aware of any baseline akin to the previous one; nevertheless, it seems plausible to use the results of the immediately prior election as a prediction. Clearly, this has got some issues: e.g. there are no prior results for new parties running for election, or for coalitions created or dismantled between elections. Anyway, that baseline is simple to implement and it can provide an easily understandable measure of the relative performance of a given method. In addition to that, even when using a measure with known issues as MAE, it is possible to check if a given prediction method is outperforming or underperforming the baseline. Therefore, such a baseline has been used to produce the data in Table 3, and to discuss the performance of the surveyed methods when predicting different elections. Both the table and the discussion appear in the following subsection.

On the feasibility of predicting elections from Twitter data

Table 2 describes 16 different attempts to predict elections based on Twitter data[iii]. About half of them were successful, one was the result reported by Tumasjan *et al.* (2010) –strongly contested by Jungherr *et al.* (2011)– others correspond to predictions in the paper by Metaxas *et al.* who, in turn, were able to predict just half of the races; and finally, Ceron *et al.*(2013) were able to predict with reasonable accuracy both the second round of the French presidential elections and the first round of the French legislative elections. The rest of studies failed

their predictions by wide margins. Taking into account that in most of the studied elections there were only two parties (or candidates) with actual possibilities of winning the race, a fifty percent performance seems close to mere chance.

When any measure of performance was reported in those papers it was MAE that, as said, does not allow for comparison neither across papers nor races in the same election. Hence, to better understand the actual performance of published methods and to make comparisons among them, a new baseline was suggested in previous section: assuming that past vote rates from the immediately prior election would happen again. This way, any method underperforming the performance (e.g. applying MAE) of such a baseline should be considered unsuccessful. Thus, performance measures were computed for each of the elections studied in the literature for (1) the proposed baseline, (2) Twitter predictions based on tweet counts, and (3) Twitter predictions based on sentiment analysis where available. Fortunately, all of the papers, except for (Tjon Kim Sang & Bos, 2012) directly provide such information. In the later case it was computed from the results reported by the authors (seats in the Senate). Table 3 shows those performance measures which have been used to draw some conclusions about the predictive power of currently applied methods.

On the feasibility of counting tweets to predict elections

There are three reports where predictions based on tweet counts outperform the baseline: (Tumasjan *et al.*, 2010; Bermingham & Smeaton, 2011; Tjon Kim Sang & Bos, 2012). Other three reports reveal that such a method underperforms the baseline: (Gayo-Avello, 2011; Metaxas *et al.*, 2011; Skoric *et al.*, 2012).

Those results are not very conclusive and it could be argued that positive results are an artifact due to mere chance. On top of that, Jungherr *et al.* (2011) showed that (1) prior important decisions regarding which parties to monitor are to be taken; and (2) performance strongly depends on the time window employed for data collection. In fact, by including the Pirate Party, Jungherr *et al.* showed that predicted results were very different. Actually, all the researchers who have replicated the method by Tumasjan *et al.* (2010) discharged minor parties – including those teams which underperformed the proposed baseline.

So, in short, Twitter prediction based on tweet counts:

1. Is too dependent on arbitrary decisions such as the parties or candidates to be considered, or the selection of a period to collect the data.
2. Its performance is too unstable because it strongly depends on such parameterizations, and
3. Considering the reported results as a whole we should not disregard the hypothesis that positive results could have been due to chance or, even, to unintentional data dredging due to *post hoc* analysis.

Under the light of current research there is no strong evidence to consider tweet counting as a reliable prediction method. However, given its simplicity it would be wise to apply it to as many elections and scenarios as possible to definitely check if the method can or cannot consistently outperform the baselines.

On the feasibility of using sentiment analysis to predict elections

It is unclear the impact that sentiment analysis has in Twitter-based predictions. The studies applying that technique are much fewer than those counting tweets and the picture they convey is confusing to say the least. On top of that, as it was aforementioned, all of the so-called sentiment analysis methods applied are not representative of state of the art techniques in the field and, thus, they achieve subpar accuracy.

According to Gayo-Avello (2011) lexicon-based sentiment analysis outperforms tweet counting but it still underperforms the baseline. Metaxas *et al.* (2011) found again that the same method outperforms raw tweet counts, but also the baseline. However, they also found that lexicon-based methods are close to random classifiers and that the proportion of correctly guessed races is no better than chance.

Results in (Bermingham & Smeaton, 2011; Tjon Kim Sang & Bos, 2012) introduce some contradictions. The former found that sentiment analysis outperforms the baseline but their method was somewhat overfitted since they incorporated data from pre-electoral polls. Results by the second authors reveal that their most complex method (involving not only sentiment analysis but also political leaning information derived from pre-electoral polls) outperforms the baseline but, at the same time, underperforms raw tweet counts.

Finally, results by Ceron *et al.* (2013) regarding the French legislative and presidential elections are rather promising. Certainly, their method still underperforms traditional polls but the error is comparable to them; the fact that vote prediction for left parties was overestimated, while for far right parties it was overestimated seems

to point out the method is sensible to ideological or self-selection biases in Twitter data. Anyway, it is one of the first studies successfully applying a state of the art method of sentiment analysis to the problem.

In short, results are still scarce and some of them even contradictory. However, most of the initial studies relied on naïve methods of sentiment analysis and, in all probability, that could explain their wide errors. More recent works applying better methods of sentiment analysis seem to be improving the accuracy of the predictions and, hence, it is clear that further research regarding the application of state of the art sentiment analysis to this field is needed. This work concludes with some recommendations in that sense.

<< Table 3 about here >>

Conclusions

This paper concludes enumerating the main weaknesses of current research regarding political prediction with Twitter data. Then a list of challenges and recommendations for future research are provided.

Weaknesses in current research

The previous analysis of the literature revealed that the predictive power of Twitter regarding political elections has been somewhat overstated. Certainly, no single paper has claimed to completely solve the problem; however, a few of them have made flamboyant claims supported by thin evidence, and the research community has many times taken out of context positive findings in single case scenarios, while forgetting the many remaining issues. Moreover, it is almost unavoidable to attribute some of the reported positive results to mere chance or involuntary data dredging since all of the results were obtained *post facto*. Finally, it has been shown that simple baselines can achieve better accuracy in many cases. None of this is surprising given that approaches up to date suffer from a number of weaknesses that should be tackled with in the future:

1. All of them are *post hoc* analysis. Even if no data dredging occurred in any of the surveyed studies the only way to dispel such a concern is by conducting the research and publishing the prediction before the elections take place. It must be noted, besides, that this condition has been emphasized by experts in traditional electoral forecasting; according to them, models must have lead, that is, *"the forecast must be made before the event. The farther in advance […] the better"* (Lewis-Beck 2005).

2. Performance should be compared against reasonable baselines. It has been shown that commonly reported measures of performance are either inadequate (winner prediction) or not comparable across elections (MAE). Hence, different measures should be studied and, in addition to that, better baselines are needed to compare new methods against them. It has been shown that incumbency is a reasonable one when predicting a number of races, and that using results from the previous election is a sensible choice when predicting vote rates for a single election. Nevertheless, additional baselines should be explored.
3. Sentiment analysis is applied with naïveté. Commonly used methods are slightly better than random classifiers and fail to catch the subtleties of political discourse. State of the art methods in sentiment analysis should be applied if the prediction method claims to be based on voter sentiment; humor and sarcasm should be taken also into consideration. On another hand, if the method claims to be noise tolerant and, thus, it ignores sentiment or polarity in the tweets it should demonstrate that it consistently works and outperforms the baseline in different elections. In other words, single case studies are not useful to proof or disproof the feasibility of a given method.
4. All of the tweets are assumed to be trustworthy when it is not the case. This issue is highly related with the previous one: noise tolerance cannot be assumed at face value but demonstrated with consistent predictions across different elections; on another hand if the method applies sentiment analysis then spam, misleading propaganda and astroturfing should be detected and filtered out, or the method should be tolerant to that noise.
5. Demographics bias is ignored even when it is well known that social media is not a random sample of the population. As it has been shown some of the methods claiming to be noise tolerant implicitly include bias in the data as a kind of noise; bias tolerance could be much more difficult to proof but, as with noise tolerance, single case positive reports would be weak arguments to support such a claim. Furthermore, those researchers aiming to replicate traditional polling should show that material improvements can be achieved when incorporating demographic information and debiasing methods.
6. Self-selection bias is simply ignored. People tweet on a voluntary basis and, therefore, data is produced by those politically active. This issue is slightly related to the previous one and, hence, those researchers

claiming their methods are noise tolerant should do their best to proof their methods are consistently immune to this additional kind of bias. Conversely, advocates of replicating traditional polls with Twitter data should be able to offer ways to circumvent this problem, and not simply use it as a blanked argument against the general feasibility of predicting elections using Twitter data.

Challenges and lines for future work

Probably the main challenge that this area of research should face is to settle down whether prediction based on Twitter data can actually be noise and bias tolerant or if, on the contrary, it should replicate actual polls and, hence, attention should be paid to geographical and demographical issues, in addition to other problems, specific to social media, such as spam, astroturfing and self-selection. As it has been shown, researchers claiming Twitter based methods can be noise and bias tolerant do not provide strong proof to support such claims but just single case positive results (that most of the time cannot be replicated in different elections). Nevertheless, this fact does not disproof the general idea of devising a noise and bias tolerant method.

Either way, it is simply not possible to ignore any more those kinds of noise and bias and, instead, anyone claiming a method is tolerant to both issues should (1) apply the proposed method to different scenarios where the nature of noise and bias is completely different (e.g. scenarios where the overrepresented ideologies differ; where the prevailing demographic groups are quite different in terms of age, education, race or income; or where the astroturfing is produced by very different groups), (2) compare the collected data with the features of the actual population to reveal the leaning of Twitter users for each scenario, and (3) proof that the same method achieves accurate predictions when working on such different scenarios.

On the other hand, those researchers claiming that a faithful replication of traditional polls is required should clearly show the ways in which they tackle with (1) noise (i.e. spam, misleading propaganda and astroturfing), (2) self-selection bias (i.e. some users not revealing their opinion and different communities reacting to different events), (3) detection of likely voters, and (4) deweighting of demographical biases in the users appearing in their data. Those methods should consistently work across different scenarios and they should consistently outperform a similar method which ignored noise and biases in Twitter data. Needless to say

geographical and demographical information about the users should be inferred if not available (e.g. Mahmud *et al.* 2012, Pennacchiotti & Popescu 2011).

Each of those approaches starts from very different assumptions but, up to now, there is no strong evidence in favor of one over the other. Needless to say, it seems sensible to assume that one needs to outperform the other and, hence, this matter needs to be settled down in the short term.

It must be noted that the method of inferring voting intention from the data is reasonably independent from either approach. Both can apply from simple tweet counting to sophisticated sentiment analysis methods. As it has been already said, there is no strong support on favor of tweet counting as a prediction method; however, it would be wise to apply it in every case to eventually determine if it can be discarded as a prediction method. With regards to sentiment analysis, it is clear that lexicon-based methods should not be employed and that, instead, state of the art sentiment analysis should be employed with a focus on the many issues of political tweets such as spam, astroturfing, humor and sarcasm.

Finally, there is the matter of the nature of the prediction and the way to evaluate its accuracy. Winner prediction and number of correctly guessed races should be avoided in the future and, instead, predicted vote rates should be provided. This option should be preferred over number of seats or number of Electoral College votes given that, this way, subtle differences among electoral systems can be safely ignored. With regards to the granularity of the prediction researchers should aim to provide predictions at the highest possible resolution and, at least, one level below the national one. Regarding performance evaluation, MAE should be reported since it is a well-known and easily understandable measure; however, on the light of the aforementioned issues it exhibits other measures should be analyzed. In addition to that, MAE should not be reported on its own but instead compared to the MAE (or other measure of error) of a reasonable baseline. In this regard, using past vote rates could be an option but traditional forecasting methods could also be used as baselines –cf. Lewis-Beck & Rice (1992) and Campbell & Garand (2000). This way, Twitter based methods could proof if they are competitive with other well-established forecasting approaches.

In this regard it must be noted that traditional forecasting models are regression models obtained from historical data and, hence, the same approach can be applied when working with Twitter data. In fact, there are a

couple of recent works that have tried to obtain regression models by combining pre-electoral polls with Twitter data.

Shi *et al.* applied such a method to Republican primaries in 2011. Their model included aggregated features obtained from Twitter during the campaign such as number of tweets and retweets, or the number of followers a given candidate had got; they did not use the actual tweet contents, however. Their regression model was fit and tested against poll results in Realclearpolitics website; they used the first 90% of the poll time series for training and the remaining 10% for testing achieving rather reasonable results. It must be noted that these authors employed the same Twitter data to make predictions using both tweet counting and lexicon-based sentiment analysis achieving much poorer results than when using their regression model.

A similar work was later developed by Lampos *et al.* (2013). They employed results from voting intention polls as training material for their model which combined both the tweets contents with user information into one single model. According to these authors intention inferred by their method outperform forecasts made from the available polls.

Both reports are quite promising and they are opening a path for future research that would help to integrate it with traditional electoral forecasting. Needless to say, further work is needed. First, it would be of interest to work in a similar fashion to that of traditional forecasting models; that is, obtaining a model not from one single election but from data corresponding to a series of elections. Second, this kind of models should be generalized to be applied to different elections in different countries with comparable electoral systems and similar Twitter usage.

Final note

Electoral prediction from Twitter data should acknowledge self-selection bias (e.g. Mustafaraj *et al.* 2011 or Mitchell & Hitlin 2013), the presence of spam and astroturfing (e.g. Mustafaraj & Metaxas 2010 or Ratkiewicz *et al.* 2011) and the prevalence of negative comments plagued with humor and sarcasm (Mejova *et al.* 2013).

Thus, methods to infer the size of silent majorities should be explored (e.g. Venkataraman *et al.* 2012) in addition to those to filter out misleading information (e.g. Castillo *et al.* 2011). This should be taken into account

no matter the prediction method is claimed to be noise tolerant or not; in the first case, the level of noise and the degree of bias from the actual population should be assessed to proof whether the accuracy of the method is unaffected by them. In the second case, all of those methods could be used to filter out the noise and remove the sources of bias.

If sentiment analysis was applied, state of the art methods are required (cf. Pang & Lee 2008 or Liu 2012), and humor and sarcasm should be either filtered out (e.g. Reyes *et al.* 2012) or the method should be tolerant to their presence.

Finally, well-established forecasting methods from the political science literature (cf. Lewis-Beck & Rice 1992 and Campbell & Garand 2000) should be used as baselines in order to provide convincing evidence of the competitiveness of the new proposed methods. In this regard it could be wise to take an approach similar to that followed in political science and propose regression based methods able to be applied in many different scenarios instead of single case models. In fact, the final goal of this line of research should be to integrate Twitter based electoral prediction within the area of electoral forecasting.

Acknowledgements

The author would like to thank the anonymous reviewers for their valuable comments which were extremely helpful to enrich the final version of the paper.

Bio

Daniel Gayo-Avello is an associate professor in the Department of Computer Science at the University of Oviedo, Spain. His research interests include information retrieval, Web mining, in particular query log mining, and online social network analysis. Gayo-Avello has a PhD in computer science from the University of Oviedo. Contact him at dani@uniovi.es or via Twitter at @PFCdgayo.

References

ASUR, S., AND HUBERMAN, B.A. 2010, “Predicting the Future with Social Media”, in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society, Los Alamitos, CA, USA, 492–499.

BARBERÁ, P. 2012, "A New Measure of Party Identification in Twitter. Evidence from Spain", paper presented at the Second EPSA Conference, June 21-23, 2012, Berlin, Germany.

BERMINGHAM, A., AND SMEATON, A. 2011, "On Using Twitter to Monitor Political Sentiment and Predict Election Results", paper presented at the Workshop on Sentiment Analysis where AI meets Psychology, November 13, 2011, Chiang Mai, Thailand.

BOLLEN, J., MAO, H., ZENG, X.J. 2011, "Twitter mood predicts the stock market", J. Comput. Science, vol. 2, no. 1, 1–8.

BOUTET, A., KIM, H., AND YONEKI, E., 2012, "What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election", Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), July 2012.

BOYD, D. 2010, "Big Data: Opportunities for Computational and Social Sciences", April 17, 2010, available at: http://www.zephoria.org/thoughts/archives/2010/04/17/big-data-opportunities-for-computational-and-social-sciences.html (accessed 14 June 2012).

CAMBELL, J.E. 2004, "Introduction—The 2004 Presidential Election Forecasts", PS: Political Science & Politics, vol. 37, no. 4, pp. 733—736.

CAMPBELL, J., AND GARAND, J. (Eds.), 2000, Before the Vote: Forecasting American National Elections, Sage Publications, Inc. Thousand Oaks, California, US.

CASTILLO, C., MENDOZA, M., AND POBLETE, B. 2011, "Information credibility on twitter", in Sadagopan, S. *et al.* (Eds.) Proceedings of the 20th international conference on World Wide Web, ACM, New York, NY, USA, pp. 675–684.

CERON, A., CURINI, L., IACUS, S.M., AND PORRO, G., 2013, "Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France", New Media & Society, published online before print April 4, 2013, available at: http://nms.sagepub.com/content/early/2013/04/02/1461444813480466.abstract (accessed 15 May 2013).

CHOI, H., AND VARIAN, H. (2009a), "Predicting the Present with Google Trends", technical report, Google Inc. April 10, 2009, available at: http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf (accessed 14 June 2012).

CHOI, H., AND VARIAN, H. (2009b), "Predicting Initial Claims for Unemployment Benefits", technical report, Google Inc. July 5, 2009, available at: http://research.google.com/archive/papers/initialclaimsUS.pdf (accessed 14 June 2012).

CONOVER, M.D., GONÇALVES, B., RATKIEWICZ, J., FLAMMINI, A., AND MENCZER, F., 2011, "Predicting the political alignment of twitter users", In Proceedings of 3rd IEEE Conference on Social Computing (SocialCom).

CONOVER, M.D., RATKIEWICZ, J., FRANCISCO, M., GONÇALVES, B., FLAMMINI, A., AND MENCZER, F., 2011, "Political Polarization on Twitter". In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

FANELLI, D. 2010, "Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data", PLoS ONE, vol. 5, no. 4, available at: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0010271 (accessed 14 June 2012).

FELLER, A., KUHNERT, M., SPRENGER, T.O., AND WELPE, I.M., 2011, "Divided They Tweet: The Network Structure of Political Microbloggers and Discussion Topics". In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

GAYO-AVELLO, D. 2011, "Don't turn social media into another 'Literary Digest' poll", Communications of the ACM, vol. 54, no. 10, pp. 121–128.

GINSBER, 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, vol. 457, 1012–1014.

GRUHL, D., GUHA, R., KUMAR, R., NOVAK, J., AND TOMKINS, A. 2005, "The predictive power of online chatter", in Grossman, R. *et al.* (Eds.), Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM, New York, NY, USA, 78–87.

HECHT, B., HONG, L., SUH, B., AND CHI, E.H., 2011, "Tweets from Justin Bieber's Heart: The Dynamics of the 'Location' Field in User Profiles", paper presented in CHI 2011, May 7-12, 2011, Vancouver, Canada.

HOPKINS, D.J., AND KING, G., 2010, "A method of automated nonparametric content analysis for social science", American Journal of Political Science, vol. 54, no. 1, pp. 229-247.

JUNGHERR, A., JÜRGENS, P., AND SCHOEN, H. 2011, "Why the Pirate Party won the German election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T.O., Sander, P.G., & Welpe, I.M. "Predicting elections with Twitter: What 140 characters reveal about political sentiment", Social Science Computer Review, published online before print April 25, 2011, available at: http://ssc.sagepub.com/content/early/2011/04/05/0894439311404119.abstract (accessed 14 June 2012).

LAMPOS, V., PREOTIUC-PIETRO, D., AND COHN, T. 2013, "A user-centric model of voting intention from Social Media", in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013).

LENHART, A., AND FOX, S. 2009, “Twitter and status updating”, report, Pew Internet and American Life, February 12 2009, available at: http://www.pewinternet.org/Reports/2009/Twitter-and-status-updating.aspx (accessed 14 June 2012).

LEWIS-BECK, M.S. 2005, “Election Forecasting: Principles and Practice”, The British Journal of Politics & International Relations, vol. 7, pp. 145—164.

LEWIS-BECK, M.S., AND RICE, T.W. 1992, Forecasting Elections, Congressional Quarterly Press, Washington D.C., US.

LIU, B. 2012, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers.

LIVNE, A., SIMMONS, M.P., ADAR, E., AND ADAMIC, L.A. 2011, “The Party Is Over Here: Structure and Content in the 2010 Election”, paper presented at the 5th International AAAI Conference on Weblogs and Social Media, July 17-21, 2011, Barcelona, Spain.

MAHMUD, J., NICHOLS, J., AND DREWS, C. 2012, “Where Is This Tweet From? Inferring Home Locations of Twitter Users”, in Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media.

MEJOVA, Y., SRINIVASAN, P., AND BOYNTON, B. 2013, “GOP Primary Season on Twitter: ‘Popular’ Political Sentiment in Social Media”, paper presented at WSDM’13, February 4-8, 2013, Rome, Italy.

METAXAS, P.T., MUSTAFARAJ, E., AND GAYO-AVELLO, D. 2011, “How (Not) to Predict Elections", in Proceedings of PASSAT/SocialCom 2011, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), IEEE Computer Society, Los Alamitos, CA, USA, 165–171.

MING FAI WONG, F., SEN, S., AND CHIANG, M. 2012, “Whay Watching Movie Tweets Won’t Tell the Whole Story?”, arXiv preprint, Princeton University, March 21, 2012, available at: http://arxiv.org/abs/1203.4642 (accessed 14 June 2012).

MISHNE, G., AND DE RIJKE, M. 2006, “Capturing global mood levels using blog posts”, paper presented at AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, March 27-29, 2006, California, USA.

MISHNE, G., AND GLANCE, N. 2006, “Predicting movie sales from blogger sentiment”, paper presented at AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs, March 27-29, 2006, California, USA.

MISLOVE, A., LEHMANN, S., AHN, Y.Y., ONNELA, J.P., AND ROSENQUIST, J.N. 2011, “Understanding the Demographics of Twitter Users”, paper presented at the 5th International AAAI Conference on Weblogs and Social Media, July 17-21, 2011, Barcelona, Spain.

MITCHELL, A., AND HITLIN, P. 2013, “Twitter Reaction to Events Often at Odds with Overall Public Opinion”, report, Pew Internet and American Life, March 4, 2013 available at: http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/ (accessed May 15 2013).

MORRIS, M.R., COUNTS, S., ROSEWAY, A., HOFF, A., AND SCHWARZ, J. 2012, “Tweeting is believing?: understanding microblog credibility perceptions”, in Poltrock, S. *et al.* (Eds.) Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, ACM, New York, NY, USA, 441–450.

MORSTATTER, F., PFEFFER, J., LIU, H., AND CARLEY, K.M., 2013, “Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose”, in Proceedings the Seventh International AAAI Conference on Weblogs and Social Media.

MUSTAFARAJ, E., AND METAXAS, P.T. 2010, “From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search”, paper presented at WebSci10: Extending the Frontiers of Society On-Line, April 26-27, 2010, Raleigh, NC, USA.

MUSTAFARAJ, E., FINN, S., WHITLOCK, C., AND METAXAS, P.T. 2011, “Vocal Minority Versus Silent Majority: Discovering the Opionions of the Long Tail”, in Proceedings of PASSAT/SocialCom 2011, 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), IEEE Computer Society, Los Alamitos, CA, USA, 103–110.

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”, paper presented at the 4th International AAAI Conference on Weblogs and Social Media, May 23-26, 2010, Washington, USA.

O’CONNOR, B., BAMMAN, D., AND SMITH, N.A. 2011, “Computational Text Analysis for Social Science: Model Assumptions and Complexity”, in Proceedings of Second Workshop on Computational Social Science and the Wisdom of Crowds, December 17, 2011, Granada, Spain.

PANG, B., AND LEE, L. 2008, Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, vol. 2, no. 1-2.

PENNACCHIOTTI, M., AND POPESCU, A.M. 2011, “A Machine Learning Approach to Twitter User Classification”, in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

RATKIEWICZ, J., CONOVER, M., MEISS, M., GONÇALVES, B., FLAMMINI, A., AND MENCZER, F. 2011, "Detecting and Tracking Political Abuse in Social Media", paper presented at the 5th International AAAI Conference on Weblogs and Social Media, July 17-21, 2011, Barcelona, Spain.

REYES, A., ROSSO, P., AND BUSCALDI, D. 2012, "From humor recognition to irony detection: The figurative language of social media", Data & Knowledge Engineering 74 (2012) 1–12

SHI, L., AGARWAL, N., AGRAWAL, A., GARG, R., AND SPOELSTRA, J. 2012, "Predicting US Primary Elections with Twitter", in Proceedings of the workshop Social Network and Social Media Analysis: Methods, Models and Applications, December 7, 2012, Lake Tahoe, Nevada, US.

SKORIC, M., POOR, N., ACHANANUPARP, PL, LIM, E.P., AND JIANG, J. 2012, "Tweets and Votes: A Study of the 2011 Singapore General Election", in Proceedings of 45th Hawaii International International Conference on Systems Science (HICSS-45 2012), IEEE Computer Society, Los Alamitos, CA, USA, 2583–2591.

SMITH, A., AND RAINIE, L. 2008, "The Internet and the 2008 election", report, Pew Internet and American Life, June 15, 2008, available at: http://www.pewinternet.org/Reports/2008/The-Internet-and-the-2008-Election.aspx (accessed 14 June 2012).

TJONG KIM SANG, E., AND BOS, J. 2012, "Predicting the 2011 Dutch Senate Election Results with Twitter", paper presented at the 13th Conference of the European Chapter of the Association for Computational Linguistics, April 23-27, 2012. Avignon, France.

TUMASJAN, A., SPRENGER, T.O., SANDNER, P., AND WELPE, I.M. 2010, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", paper presented at the 4th International AAAI Conference on Weblogs and Social Media, May 23-26, 2010, Washington, USA.

TUMASJAN, A., SPRENGER, T.O., SANDNER, P.G., AND WELPE, I.M. 2011, "Where There is a Sea There are Pirates: Response to Jungherr, Jürgens, and Schoen", Social Science Computer Review, published online before print May 18, 2011, available at: http://ssc.sagepub.com/content/30/2/235.abstract (accessed 14 June 2012).

VENKATARAMAN, M., SUBBALAKSHMI, K.P., AND CHANDRAMOULI, R. 2012, "Measuring and quantifying the silent majority on the Internet", in Proceedings of the 35th IEEE Sarnoff Symposium. 21-22 May, 2012, Newark, NJ, US.

WILSON, T., WIEBE, J., AND HOFFMANN, P. 2005, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", paper presented at the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, October 6-8, 2005, Vancouver, B.C., Canada.

WONG, F.M.F., TAN, C.W., SEN, S., AND CHIANG, M. 2013, "Quantifying Political Leaning from Tweets and Retweets", in Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM), July 2013.

Appendix: An annotated bibliography

This section covers the papers on electoral prediction analyzed above and, additionally, works on related areas such as credibility, rumors, and Twitter demographics.

Electoral prediction from Twitter data

O'Connor *et al.* 2010

This is the earliest paper discussing the feasibility of using Twitter data as a substitute for traditional polls although it does not describe any prediction method.

They employed a subjectivity lexicon to determine a positive and a negative score for each tweet in their dataset. Then, raw numbers of positive and negative tweets are used to compute a sentiment score.

Using that method, sentiment time series were prepared for a number of topics (namely, consumer confidence, presidential approval, and US 2008 Presidential elections). Both consumer confidence and presidential approval polls exhibited correlation with Twitter sentiment data, but no correlation was found with electoral polls.

Tumasjan *et al.* 2010, Jungherr *et al.* 2011, and Tumasjan *et al.* 2011

The first paper started the whole line of research analyzed in this work and proposed MAE as a measure of performance. It has two distinct parts: In the first one LIWC (Linguistic Inquiry and Word Count) is used to perform an analysis of the tweets related to different parties running for the German 2009 Federal election. In the second part, the authors state that the mere count of tweets mentioning a party accurately reflected the election results with a performance close to that of actual polls.

That claim was rebutted by Jungherr *et al.* (2011) who pointed out that the method required arbitrary choices (e.g. not taking into account all the parties running for the elections) and that its results depended on the selected time window.

Tumasjan *et al.* (2011) tried to dispel those doubts. Unfortunately, their counterarguments are not compelling enough and, besides, they toned down their previous conclusions: saying that Twitter data is not to replace polls but to complement them; or stating that the prediction method was not their main contribution[iv].

Metaxas *et al.* 2011.

With (Jungherr *et al.*, 2011) this is one of the few papers casting doubts on the predictive powers of Twitter. After analyzing results from a number of elections, they concluded that Twitter data is slightly better than chance when predicting elections. Hence, they suggested the use of incumbency as a baseline.

They also described three necessary standards for any method claiming predictive power: (1) it should be an algorithm, (2) it should take into account the demographic bias in Twitter's user base, and (3) it should be "explainable", i.e. black-box approaches should be avoided.

Livne *et al.* 2011

This is not a prediction method since it does not rely on users' tweets but in those by candidates plus their social graph. They also incorporate additional information such as the party a candidate belongs to, or incumbency.

They claim 88% precision when incorporating Twitter data (both tweets and graph) versus 81% precision without such data; the improvement is not substantial although noticeable.

Finally, it must be noted that elections are modeled as binary processes so important information is missed (such as in tight elections, or scenarios with coalitions).

Bermingham and Smeaton 2011

This paper discusses different approaches to incorporate sentiment analysis to a predictive method. The method was put to test with the 2011 Irish General Election revealing it was not competitive against traditional polls.

Gayo-Avello 2011

This paper describes how different methods failed to predict the 2008 US Presidential Election since they predicted an Obama victory in every state, including Texas. The methods were those proposed in (Tumasjan *et al.*, 2010; O'Connor *et al.*, 2010) and a post-mortem on the reasons for such a failure is provided.

A number of problems are suggested: (1) The "file-drawer" effect; (2) Twitter data is biased and is unrepresentative; and (3) the sentiment analyzers commonly used are only slightly better than random classifiers.

Tjong Kim Sang and Bos 2012

In this paper Twitter data regarding the 2011 Dutch Senate elections was analyzed. They found that tweet counting is a bad predictor and that sentiment analysis can improve performance.

Nevertheless, performance is below that of traditional polls and, moreover, the method relies to some extent on that polling data to correct for demographic bias.

Skoric *et al.* 2012

Similarly to previous paper, this also shows that there is certain correlation between Twitter chatter and votes but not enough to make accurate predictions –they found performance was much worse than that reported in (Tumasjan *et al.*, 2010).

They argue that Twitter can provide a somewhat reasonable glimpse on national results but it fails when focusing on local levels. Hence, in addition to the technical caveats they discuss additional problems such as democratic maturity of the country, competitiveness of the election, and media freedom.

Ceron *et al.* 2013

This paper applied supervised sentiment analysis to determine the popularity of Italian political leaders during 2011 and to predict the outcome of the French presidential and legislative elections. In the later two cases the authors collected tweets related to the candidates and parties running for election during one week before the election; then, they applied to those tweets the sentiment analysis method devised by Hopkins & King (2010) to find the vote rate for each candidate or party. Their results were worse than those of traditional polls in terms of MAE but still pretty reasonable. An issue with their method is that seems to be sensible to ideological or self-selection biases in Twitter since the vote rate for far left parties was overestimated while the vote rate for far right parties was underestimated.

Sources of bias in Twitter

Mislove *et al.* 2011

This paper analyzes a sample of Twitter users in the US along three different axes, namely, geography, gender and race/ethnicity.

The methods applied are simple albeit compelling. All of the data was inferred from the user profiles: geographical information was obtained from the self-reported location; gender was determined using the first name and statistical data from the US Social Security Administration; and the last name and data from the US Census was used to derive race/ethnicity.

Cleary, such methods are prone to error but it is probably rather tolerable and the conclusions of the study are sensible: highly populated counties are overrepresented, users are predominantly male, and Twitter is a non-random sample with regards to race/ethnicity.

They concluded that post-hoc corrections based on the over- and under-representation of different groups could be applied to improve predictions based on Twitter data.

Mustafaraj *et al.* 2011

This paper provides compelling evidence on the existence of two extremely different behaviors in social media: on one hand there is a minority of users producing most of the content (vocal minority) and on the other there is a majority of users who hardly produce any content (silent majority).

These two groups are clearly separated and the vocal minority behaves as a resonance chamber spreading information aligned with their own opinions. Thus, they suggest extreme caution when building predictive models based on social media.

Mitchell & Hitlin 2013

This report reveals that the reaction of Twitter user does not correlate with public opinion (in the US) and, on top of that, it is not consistent. That is, for some events the reaction of Twitter users denotes a more liberal leaning that the population as a whole, while for other events the reaction is much more conservative. The report emphasizes the fact that Twitter reach is modest among the population and, moreover, Twitter users are not representative of the public: they are younger and more liberal. In addition to that, the report attributes part of the inconsistency on Twitter reactions to self-selection bias: i.e. users reacting to some topics are not sharing their views with regards to other topics.

Denoising Twitter data

Mustafaraj and Metaxas 2010

This paper introduces the concept of "Twitter-bomb": the use of fake accounts in Twitter to spread disinformation by "bombing" targeted users who, in turn, would retweet the message achieving viral diffusion.

They describe a smear campaign orchestrated by a Republican group against Democrat candidate Martha Coakley and how it was detected and aborted.

Ratkiewicz *et al.* 2011

This paper describes the Truthy project inspired by the previous paper. Truthy is a system to detect astroturf political campaigns either to simulate widespread support for a candidate or to spread disinformation. The system is described in detail and a number of cases and performance analysis are provided.

Castillo *et al.* 2011, and Morris *et al.* 2012

The first paper is, to the best of my knowledge, the first one describing a method to separate credible from not credible tweets. It describes in detail which features to extract from the tweets to then train a classifier.

Morris *et al.* (2010) did not develop an automatic method to filter tweets but they conducted a survey to find the features that make users to perceive a tweet as credible. Content alone was found to be not enough to assess truthfulness so users rely on additional heuristics. These can be manipulated by the authors of tweets and, therefore, can affect credibility perceptions.

---

[i] This effect refers to those conservative voters not disclosing their intentions in polls and, thus, biasing the corresponding predictions for conservative vote.

[ii] http://twitvote.twitmarks.com/

[iii] As it has been shown in the commented bibliography [O'Connor *et al.* 2010; Livne *et al.* 2011] were not properly predictions.

[iv] Even when the phrase "predicting elections with Twitter" prominently appears in the title of that paper.

Authors	Start of collection	End of collection
Livne *et al.* 2011	3 years before election	election day
O'Connor *et al.* 2010	10 months before election	election day
Gayo-Avello 2011	5 months before election (candidate nomination)	election day
Tumasjan *et al.* 2010	7 weeks before election	one week before election day
Jungherr *et al.* 2011	7 weeks before election (to replicate findings by Tumasjan *et al.*)	election day
Skoric *et al.* 2012	1 month before election	election day
Bermingham & Smeaton 2011	3 weeks before election	election day
Metaxas *et al.* 2011	1 week before election	election day
Tjong Kim Sang & Bos 2012	1 week before election	election day
Ceron *et al.* 2013	1 week before election	election day

Table 1. Different periods of data collection used in the literature ordered by decreasing time length. All of the studies, except for the one by Tumasjan *et al.* finished data collection the day before elections.

Authors	Election	Period & Method of collection	Data cleansing	Prediction method	Performance evaluation	Reported results
O'Connor *et al.* 2010	US presidential election, 2008 (Nov. 4, 2008)	February to November 2008. Candidate names used as keywords.	No cleansing at all.	Lexicon-based sentiment analysis. Aggregated results at national level. No prediction attempted.	Correlation against pre-electoral polls.	No significant correlation found.
Gayo-Avello 2011		June 1 to November 3, 2008. Presidential and vice-presidential candidate names.	Geolocated tweets at county level. Attempt to debias data according to user age.	Lexicon-based sentiment analysis. Individual votes. Aggregated results at state level. Vote rates.	MAE against actual electoral results.	MAE 13.10% (uncompetitive with traditional polls).
Tumasjan *et al.* 2010	German federal election, 2009 (Sept. 27, 2009)	August 13 to September 19, 2009. Parties present in the *Bundestag* and politicians from those parties.	No cleansing at all.	Number of tweets. Aggregated results at national level. Vote rates.		MAE 1.65% (comparable with traditional polls, although larger).
Jungherr *et al.* 2011		Different time windows from August 13 to September 27, 2009. Parties running for election.				Unstable MAE depending on time window but larger than MAE reported by Tumasjan *et al.* 2010. Incorrect prediction when taking into account all parties running for election.

Table 2. Different studies on the feasibility of predicting elections with Twitter data characterized according to the scheme proposed above. Reports are ordered according to date of election and not of publication. Results appear on the right; those positive are shaded.

Metaxas *et al.* 2011	US Senate special election in MA, 2010 (Jan. 19, 2010)	January 13 to 20, 2010. Candidate names.	No cleansing at all.	Number of tweets. Aggregated results at state level. Vote rates.	Winner prediction and MAE	Incorrect prediction. MAE 6.3%.
				Lexicon-based sentiment analysis and vote share. Aggregated results at state level. Vote rates.		Correct prediction. MAE 1.2%
	US elections in CO, 2010 (Nov. 2, 2010)	October 26 to November 1, 2010. Candidate names used to filter a *gardenhose* dataset.		Number of tweets. Aggregated results at state level. Vote rates.		Incorrect prediction. MAE 24.6%
				Lexicon-based sentiment analysis and vote share. Aggregated results at state level. Vote rates.		Correct prediction. MAE 12.4%
	US elections in NV, 2010 (Nov. 2, 2010)			Number of tweets. Aggregated results at state level. Vote rates.		Correct prediction MAE 2.1%
				Lexicon-based sentiment analysis and vote share. Aggregated results at state level. Vote rates.		Incorrect prediction MAE 4.7%

Table 2 (continuation). Results by Metaxas *et al.* (2011) when trying to replicate results using methods analogous to those by (O'Connor *et al.*, 2010; Tumasjan *et al.*, 2010). As it can be seen, results are inconclusive and there is no clear relation between MAE and accuracy of prediction.

Metaxas *et al.* 2011	US elections in CA, 2010 (Nov. 2, 2010)	October 26 to November 1, 2010. Candidate names used to filter a *gardenhose* dataset.	No cleansing at all.	Number of tweets. Aggregated results at state level. Vote rates.	Winner prediction and MAE	Correct prediction. MAE 3.8%
				Lexicon-based sentiment analysis and vote share. Aggregated results at state level. Vote rates.		Incorrect prediction. MAE 6.3%
	US elections in KY, 2010 (Nov. 2, 2010)			Number of tweets. Aggregated results at state level. Vote rates.		Correct prediction. MAE 39.6%
				Lexicon-based sentiment analysis and vote share. Aggregated results at state level. Vote rates.		Correct prediction. 1.2%
	US elections in DE, 2010 (Nov. 2, 2010)			Number of tweets. Aggregated results at state level. Vote rates.		Incorrect prediction. MAE 26.5%
				Lexicon-based sentiment analysis and vote share. Aggregated results at state level. Vote rates.		Incorrect prediction MAE 19.8%

Table 2 (continuation). Results obtained by Metaxas *et al.* (2011) when trying to replicate results using methods analogous to those by (O'Connor *et al.*, 2010; Tumasjan *et al.*, 2010). As it can be seen, results are inconclusive and there is no clear relation between MAE and accuracy of prediction.

Livne *et al.* 2011	US elections, 2010 (Nov. 2, 2010)	March 25, 2007 to November 1, 2010. Tweets and social graph for 700 candidates.	Not applicable. This method did not employ potential voter tweets but candidate data.	Regression models for binary results of races which included external data. Aggregated results at state level. Winner prediction.	Winner prediction	81.5% accuracy when using external data alone. 83.8% accuracy when incorporating tweets (but not graph data). Not noticeable improvement.
Bermingham & Smeaton, 2011	Irish general election, 2011 (Feb. 25, 2011)	February 8 to 25, 2011. Major parties.	No cleansing at all.	Number of tweets (different samples tested). Aggregated results at national level. Vote rates.	MAE against actual electoral results.	MAE 5.58% (uncompetitive with traditional polls).
				ML-based sentiment analysis. Aggregated results at national level. Vote rates.		MAE 3.67% (uncompetitive with traditional polls even after overfitting for using poll data for training).
Skoric *et al.*, 2012	Singaporean general election, 2011 (May 7, 2011)	April 1 to May 7, 2011. Tweets by 13,000 Singaporean political engaged users. Parties and candidates names were used to filter the tweets.	Only data produced by users located at Singapore was used.	Number of tweets. Aggregated results at national level. Vote rates.		MAE 5.23% Inconclusive since pre-electoral polls are banned in Singapore.

Table 2 (continuation). Studies by (Livne *et al.*, 2011; Bermingham & Smeaton, 2011; and Skoric *et al.*, 2012). The former is not properly a prediction method.

Tjong Kim Sang & Bos, 2012	Dutch senate election, 2011 (May 23, 2011)	February 23 to March 1, 2011. Major parties.	No cleansing at all.	Number of tweets. Aggregated results at national level. Number of Senate seats.	Offset in number of seats as compared with actual results.	MAE (computed by this author) 1.33% Competitive with traditional polls.
			Attempt to debias data according to political leaning by using pre-electoral polling data	Sentiment analysis. Aggregated results at national level. Number of Senate seats.		MAE (computed by this author) 2% Comparable to traditional polls although larger.
Ceron *et al.* 2013	French presidential election, 2nd round, 2012 (May 6, 2012)	April 27 to May 5, 2012. Candidates for the second round.	No cleansing at all.	Sentiment analysis. Aggregated results at national level. Vote rates.	Winner prediction.	Correct prediction. MAE (computed by this author) 3.26% Comparable to traditional polls although larger.
	French legislative elections, 1st round, 2012 (June 10, 2012)	One week before the election. Candidates and parties taking part into the elections.	No cleansing at all.	Sentiment analysis. Aggregated results at national level. Vote rates.	MAE against actual electoral results.	MAE 2.38% Comparable to traditional polls although larger.

Table 2 (ending). Study by Tjong Kim Sang & Bos (2012). MAE (computed by this author) is rather competitive with traditional polls. Interestingly, the more complex method incorporating sentiment analysis and some pre-election poll data underperforms the simpler relying on tweet counts. The work by Ceron *et al.* regarding the French presidential and legislative elections achieved results comparable to those of traditional polls although with a larger of error.

Election	Baseline	Twitter prediction (tweet counts)	Twitter prediction (sentiment analysis)
US presidential election, 2008	5.86% (states analyzed in Gayo-Avello 2011)	15.87% (computed for this paper from data by Gayo-Avello 2011)	13.10% 11.61% when debiasing according to age (Gayo-Avello 2011)
German federal election, 2009	3.75%	1.65% (Tumasjan *et al.* 2010) From 1.51% to 3.34% (Jungherr *et al.* 2011)	Not available
US elections, 2010	8.85% (states analyzed in Metaxas *et al.* 2011)	17.12% (Metaxas *et al.* 2011)	7.58% (Metaxas *et al.* 2011)
Irish general election, 2011	6.26%	5.58% (Bermingham and Smeaton 2011)	3.67% includes polling data (Bermingham and Smeaton 2011)
Singaporean general election, 2011	3.36%	5.23% (Skoric *et al.* 2012)	Not available
Dutch senate election, 2011	2.38%	0.89% raw counts 1.33% normalized counts (computed for this paper by the author from data by Tjong Kim Sang and Bos 2012)	2% normalized counts plus debiasing (computed for this paper by the author from data by Tjon Kim Sang and Bos 2012)
French presidential election, 2nd round, 2012	4.7%	Not available	3.26% (Ceron *et al.* 2013)
French legislative elections, 1st round, 2012	4.68%	Not available	2.38% (Ceron *et al.* 2013)

Table 3. Performance measured as MAE for a naïve baseline predicting past vote rates will happen again and the two different kinds of Twitter predictions (i.e. those based on tweet counts and those relying on sentiment analysis). Those methods outperforming the baseline appear shaded.