http://doi.acm.org/10.1145/2001269.2001297

Don't Turn Social Media Into Another 'Literary Digest' Poll*

Daniel Gayo-Avello†

September 26, 2011

Department of Computer Science, University of Oviedo (SPAIN)

Abstract

User generated content has experienced an explosive growth both in the diversity of applications and the volume of topics covered by its users. Content published in micro-blogging systems like Twitter is thought to be feasibly data-mined in order to "take the pulse" of society. Recently, a number of positive studies have been published praising the goodness of relatively simple approaches to sampling, opinion mining, and sentiment analysis. This paper will attempt to play devil's advocate by detailing a study in which such simple approaches largely overestimated Obama's victory in the 2008 U.S. Presidential Elections. A thorough post-mortem of that experiment has been conducted and several important lessons have been extracted.

1 Introduction

Twitter is a micro-blogging service, i.e. a system to publish short text messages, or *tweets*, which are shown to users who are following the author. Many Twitter users decide not to protect their tweets and, hence, they appear in the so-called *public timeline*. Such tweets are accessible by means of Twitter's API and, thus, they are easy to collect.

Twitter's original slogan –*What are you doing?*– encouraged users to simply share updates about their daily activities with their friends. Nevertheless, Twitter has since evolved into a complex information dissemination platform,

*The original title for this paper was *"A warning against converting Social Media into the next Literary Digest"*.

†Correspondence to: Daniel Gayo-Avello, Department of Computer Science (University of Oviedo) – Edificio de Ciencias, C/Calvo Sotelo s/n 33007 Oviedo (SPAIN), dani@uniovi.es

especially during mass convergence situations [17]. In short, under certain circumstances, Twitter users not only provide information about themselves but also publish real-time updates on current events[1].

Therefore, Twitter has become a source of information on current events updated in real-time by millions of users[2] who are reacting to those events. It was only a matter of time before the research community turned to Twitter to exploit such a rich source of information.

The aim of this paper is not to provide a comprehensive survey on this topic but rather to focus on one of its most appealing applications: predicting present[3] and future events by using Twitter data.

Such an application seems quite natural in light of the excellent results that have been obtained by mining query logs (e.g. [3, 5]) and a number of studies have already been conducted on the topic. Asur and Huberman [1], for instance, exploited Twitter data to predict box-office revenues for movies; O'Connor *et al.* [10] rather successfully correlated Twitter data with several public opinion time series[4]; and Tumasjan *et al.* [14] claimed to have predicted the outcome of German elections by merely counting the number of mentions each candidate had received in Twitter.

Needless to say, such studies have been well received by the general public and the press and have created a fair amount of hype on the topic, particularly regarding the possibility of predicting elections. In fact, a couple of informal experiments claimed that last electoral outcomes in the United Kingdom and Belgium were accurately predicted by using Twitter data[5].

Such reports seem to imply that predicting future events from Twitter is fairly straightforward. Nevertheless, as this paper will show, that is not the case.

2 You can't (always) predict elections from Twitter

As of December 2008, 11% of American adults online were using Twitter or analogous services [9]. While that is an important amount, the fact remains that the vast majority of Internet users, not to mention people in general, are not using Twitter. Thus, Twitter users are just a sample and, probably, a very biased one.

[1] The 2008 Mumbai attacks or the 2009 Iranian election protests are perhaps among the best-known cases were Twitter played such a role.

[2] By mid-2009 Twitter had 41.74 million users [7].

[3] Bill Tancer, from Hitwise, argues that "predicting" present events should not be defined as "prediction" but, rather, as data arbitrage [13].

[4] These authors found correlation between Twitter data and the consumer confidence indices and the presidential job approval ratings. However, there was no substantial correlation between Twitter data and that from polls for the 2008 U.S. Presidential Elections.

[5] `http://www.scribd.com/doc/31208748/Tweetminster-Predicts-Findings`, and `http://geekblog.eyeforit.be/component/content/article/18-news/20-twitter-analysis-belgian-2010-elections-party-with-most-twitter-coverage-also-wins-elections.html`

In addition, another kind of bias permeates research: the tendency of researchers to report positive results while suppressing the negative ones. This so-called *file drawer* effect can have a hurtful influence if people plainly assume that conclusions from a few selected positive experiments can be straightforwardly applied to any other conceivable scenario.

It has been over 70 years since the ill-fated 1936 Literary Digest Presidential poll which is still remembered for its dismal failure in predicting the presidential elections in the United States. Conducted among its own readers, people from the telephone directory, and a list of registered car owners, the now infamous poll concluded that the Republican candidate, Alf Landon, would beat F.D. Roosevelt by a landslide. In reality, Roosevelt won the election with a 61% of the popular vote[6].

By ignoring negative results, current research risks converting Social Media Analytics into the next Literary Digest poll. In this paper one such negative result is detailed: namely, an experiment involving a large collection of tweets published during the 2008 U.S. Presidential campaign which predicted Obama to win every battleground state and Texas.

As with the Literary Digest poll, that experiment could be dismissed without much further ado by attributing its failure to poor sampling methods, or defects in the system which assigned voting intentions to user tweets, or even recurring to prejudices and stereotypes regarding the political views of Twitter users.

Obviously, due to its nature the sampling was biased, but the truth is that every prediction inferred from social media –even those with positive results– exhibits analogous biases. The sentiment analysis performed in the study in question was naïve but even more simple systems have proved sound enough to achieve positive results. And finally, no matter how appealing ideological bias could be to explain this outcome, such a hypothesis should be tested.

The reader will recall that previous section of this paper referenced two reports [10, 14] dealing with quite the same topic and will perhaps wonder what the present study intends to contribute to the matter.

First, it should be noted that the findings of the aforementioned studies could seem to be in contradiction. After all, Tumasjan *et al.* claimed that the number of tweets mentioning a candidate was a reflection of vote share and that this had a predictive power close to traditional polls. O'Connor *et al.*, however, did not find any substantial correlation between a much more complex sentiment analysis performed on Twitter data and several polls conducted during the 2008 U.S. Presidential Elections.

Nevertheless, because both studies dealt with two very different political scenarios (Tumasjan *et al.* studied elections in Germany whereas O'Connor *et al.* dealt with elections in the U.S.) and both used different kinds of ground-truth data (Tumasjan *et al.* compared data with the election results –popular vote– while O'Connor *et al.* used pre-election polls and not the actual election results) it is quite difficult to say if such predictions are possible from Twitter

[6]Many have blamed the biased sample as the source of the flawed result; nevertheless, the analysis by Squire [12] on the actual issues with that poll is highly recommendable.

data. Moreover, even if they were possible, there are still serious questions as to what the required conditions would be in order to make them.

Therefore, this paper aims to provide a balanced view of the actual possibilities of Social Media Analytics. In order to do that, a much more detailed study and analysis of electoral prediction from Twitter data was conducted than in the aforementioned studies.

Furthermore, the aim of the study described in this paper was not to compare Twitter data with pre-election polls or the popular vote as had previously been done. Instead, the goal was to obtain predictions on a state by state basis. Additionally, unlike the other studies, the predictions were not to be made by aggregating Twitter data; quite the contrary, voting intention for every single user was detected from their individual tweets. In order to do this, four different sentiment analysis methods described in the most recent literature were applied, and their performance was carefully evaluated.

As it will be shown, the results for the 2008 U.S. Presidential Elections could not have been predicted from Twitter data by using commonly applied methods. While this is certainly consistent with some of the results obtained by O'Connor *et al.*, this study goes one step further by clarifying the nature of the failure (a large overestimation of the vote share for Obama) and will provide a thorough analysis of its possible causes (such as urbanization and age demographics or, even, a possible "Shy Republican" factor).

Hence, the lesson becomes clear: researchers must be cautious about simplistic assumptions regarding forecasting based on the so-called *Big Data* in general, and Twitter data in particular.

3 The 2008 U.S. Presidential election Twitter dataset

For the purposes of the present study, a collection of tweets was started shortly after the 2008 U.S. Presidential elections in order to check the feasibility of employing Twitter to predict future election outcomes. The Twitter Search API was used, employing one query for each candidacy: `obama OR biden` for the Democratic candidates, and `mccain OR palin` for the Republicans.

An API parameter to indicate a geographical area was used in order to only consider tweets published by U.S. residents, and, in addition to using these "geolocated" queries, another API parameter to indicate a temporal interval for the query was also employed. Thus, by issuing queries limited both geographical and temporally, it was possible to obtain 100 tweets per candidate, per county, per day.

Doing this for every county would have involved submitting a large number of HTTP requests to Twitter's servers. Obviously, the number of daily requests one IP address can submit is limited, and more importantly, the Twitter index does not contain all published tweets but rather those within a sliding time frame. This meant it was critical to get the data as soon as possible and, as a result, the collection was focused on a few selected states: one traditional stronghold state for both parties (California for the Democrats, and Texas for

State	# tweets	# users	Population	Margin of error @ 95%
California	94,298	7,420	36,961,664	1.46%
Florida	27,647	2,874	18,537,969	2.44%
Indiana	11,842	1,083	6,423,113	3.87%
Missouri	16,314	1,408	5,987,580	3.48%
Montana	**817**	**105**	**967,440**	**12.98%**
N. Carolina	21,012	1,683	9,380,884	3.07%
Ohio	23,549	2,266	11,542,645	2.80%
Texas	43,160	4,358	24,782,302	1.97%

Table 1: Number of tweets and unique users collected per state, in addition to the 2009 population estimate for each state, and the expected margin of error at 95% level of confidence for each of the samples (provided that they were actually random).

the Republicans) as well as the six swing states (Florida, Indiana, Missouri, Montana, North Carolina, and Ohio).

Using the API in this way it was possible to collect data back to September of 2008. To get tweets from as far back as early June, the feed for every user within the already collected data was crawled, saving tweets mentioning one of the candidacies. This meant that the final collection comprised 250,000 tweets, published by 20,000 users from June 1, 2008 to November 11, 2008.

The first thing to check was whether or not the dataset could be considered a statistical representative sample[7]. Thus, the number of tweets and unique users in each state were compared to their populations. In addition, sampling errors were computed on the assumption that the collection was close to a random sample. The correlation between population and both the numbers of tweets and users was almost perfect (Pearson's r coefficients were 0.9604 and 0.9897, respectively), and, as shown in Table 1, all of the samples except for Montana exhibited a fairly low sampling error. Thus, Montana was discarded.

After this preliminary analysis, a time series was plotted for each candidacy, in addition to a 7-day moving average for each (Figure 1). The plot exhibits peaks corresponding to relevant events: the presumptive nomination of Obama (June 3), Obama's acceptance of the Democratic nomination (August 28), Palin's nomination for vice-president (August 29), the presidential debates (September 26, October 7 and 15), the vice-presidential debate (October 2), and Election Day (November 4). Thus, the number of tweets from September to November seemed to be consistent with a reasonable sampling: the amount of "conversation" grew as the campaign progressed, it showed bursts during important events, and it dropped after election day.

Interestingly enough, the number of tweets related to Obama/Biden was

[7] It should be noted that the method of collection introduced a sample-selection bias: first, just a fraction of the Twitter's *firehose* is provided as search results and, second, not every user in Twitter provides a sensible location (about 50% of the profiles, according to our own data).

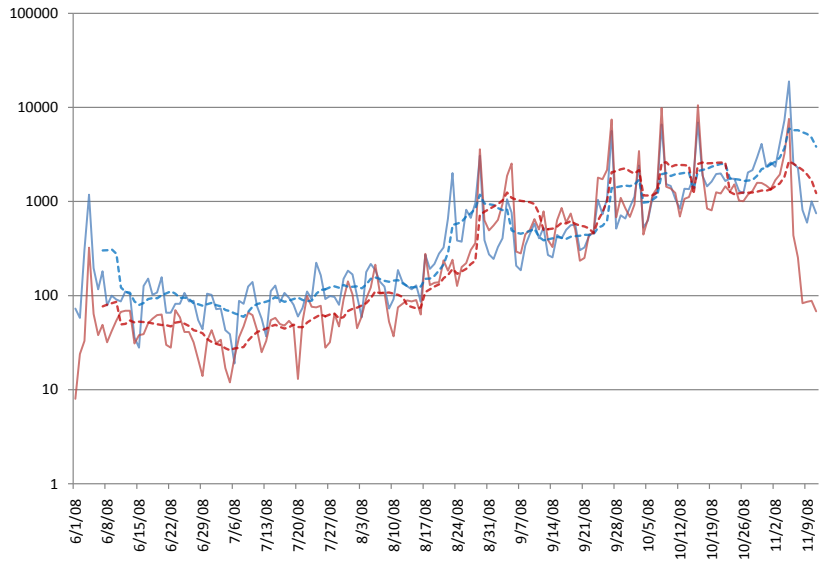

Figure 1: Time series depicting the volume of tweets regarding each of the main parties candidacies. Obama/Biden is shown in light blue and McCain/Palin in red. Dashed lines are 7-day moving averages.

consistently higher than those related to McCain until Palin was picked as the vice-presidential candidate. That "advantage" lasted only until the third presidential debate. As reflected in the moving averages, both parties' conventions produced almost the same volume of tweets; nevertheless, after Palin's nomination the number of tweets dealing with the Republican ticket outnumbered those dealing with the Democrats[8]. The same plots reveal how the difference between both candidates progressively reduced after each debate and, after the last one, tweets containing Obama or Biden once again outnumbered those referring to the McCain/Palin ticket.

Hence, this collection looked promising: it seemed to follow the evolution of national polls, and there was a strong correlation between the volume of users and tweets from each state and its population. All of this might seem to suggest an accurate sampling and, given that the number of people involved in this dataset was much larger than the samples employed in national polls, one might expect even greater accuracy (this would later prove to be wrong).

4 Inferring voting intention from tweets

Despite the extensive literature on automatic sentiment analysis (cf. [2, 4]), virtually all current research on micro-blogging analysis relies on rather simple

[8] This was also the first time that McCain lead the national polls.

methods. For the purpose of this study four different methods were applied. One was based on mention counts, two relied on polarity lexicons, and the last one was based on the *semantic orientation* method [15].

The idea underlying the first method was simple: to count the number of appearances of a candidate in the user's tweets assuming the one more frequently mentioned would be the one the user would vote for. This heuristic is pretty rough but, interestingly, it seemed to work in order to predict the outcome of elections in Germany [14]; leading Tumasjan *et al.* to remark:

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

The second method was based on the lexicon compiled by Wilson *et al.* [16] which consisted of a list of terms labeled as either positive or negative. Thus, a tweet was labeled positive if it contained more positive than negative terms and vice-versa. Because each tweet in the collection dealt with just one candidate it was possible to count, for each user, the number of positive and negative tweets for each candidate. It was therefore supposed that a user would vote for the candidate with the highest score. A similar method was employed by O'Connor *et al.* [10] with mixed results; who asserted:

> *"A high error rate merely implies the sentiment detector is a noisy measurement instrument. With a fairly large number of measurements, these errors will cancel out relative to the quantity we are interested in estimating aggregate public opinion."*

Another method relying on a polarity lexicon –Vote & Flip [4]– was implemented. This method basically consists of counting the number of positive, negative, neutral, and negation words appearing in a sentence to later apply a set of rules to infer its polarity.

Finally, *semantic orientation* [15] was adapted to this particular study. The original approach consisted of finding phrases with either a positive or negative polarity. Such a value was based on an estimation –found by means of a search engine– of the Pointwise Mutual Information between the phrase and the keywords *"poor"* and *"excellent"*. The implemented version, however, differed from the original in that it did not rely on either a search engine or on the pair *"poor/excellent"*. Instead, it relied on a subset of tweets published by users who had clearly stated their voting intentions[9].

Table 2 shows a few selected phrases which this method found to be either supporting or opposing each candidate. As expected, the patterns selected to build the subset appeared top-ranked but other interesting patterns were also found.

Obviously, in order to evaluate the performance of each of these methods the actual votes of the users were needed. During the elections an informal opinion-poll was conducted: a website called TwitVote[10] asked users to declare

[9] Tweets from users who had published anything including phrases such as **"I will vote for..."**, **"I'm not voting..."**, **"I'd vote..."**, etc., were employed.

[10] `http://twitvote.twitmarks.com/`

	Supporting phrases		Opposing phrases	
Obama	I'm voting	4.5433	Pelosi Reid	-6.0074
	Democrat Barack	4.3369	LA Times	-5.2705
	will vote	4.2214	Valerie Jarrett	-5.0640
	democratic presidential	4.1600	al-Mansour	-4.9485
	Obama leads	3.7239	Dohrn Ayers	-4.8230
	poll Obama	3.7239	Khalidi	-4.8230
	presidential nominee	3.3913	far left	-4.6855
	am voting	3.3369	Rashid Khalidi	-4.6855
	nominee Barack	3.0287	Ayers Klonsky	-4.6855
	30 reasons	2.9584	not vote	-4.5335
McCain	will vote	4.4790	republican presidential	-3.8064
	am voting	4.3636	McCain ad	-3.7239
	I'd vote	4.3636	sen. John	-3.3369
	I'm voting	4.2511	Palin campaign	-2.9584
	voting McCain	4.0265	knows how	-2.8063
	president and	3.9485	K. Michael	-2.7239
	I'm glad	3.9485	Paris Hilton	-2.6364
	a president	3.6855	is wrong	-2.4438
	I'll vote	3.6855	kill him	-2.2214
	our next	3.6855	Ashley Todd	-2.2214

Table 2: A selection of a few supportive and opposing phrases for both candidates obtained by means of semantic orientation.

their votes by publishing a tweet containing both their vote and the hashtag `#twitvote`. Thus, by collecting tweets published on November 4 and tagged as `#twitvote` it was possible to find the actual votes of a number of users.

Only two thousand users (9% of the dataset) used TwitVote. Among those who used it, 86.6% voted for Obama and the rest for McCain[11]. These results, so different from the actual popular vote, did not bode well for the study because they seemed to point to a large bias in Twitter users towards the Democratic Party. Nonetheless, the data was used to evaluate the performance of each of the methods to infer user voting intention, which proved quite inadequate (Table 3).

Precision when inferring users supporting Obama was rather high, but very poor with regard to McCain support. What's even more perplexing is that different methods achieved very similar results. Indeed, all of the methods seemed to drift towards a random classifier. This was bad on its own right but, in order to compare their relative performance, it seemed reasonable to compare all of the methods with a perfectly informed random classifier: one assigning voting intention with regards to the proportion of "votes" according to TwitVote.

As shown in Table 4, assuming the most frequently mentioned candidate

[11] This does not differ from the final results achieved by TwitVote: 85.9% Obama vs. 14.1% McCain.

Method	Precision Obama	Precision McCain	Accuracy
Most frequent candidate	82.4%	7.8%	50.7%
Polarity lexicon	88.8%	17.7%	61.9%
Vote & Flip	92.7%	10.7%	50.6%
Semantic Orientation	92.3%	15.6%	36.7%

Table 3: Performance results for each of the four automatic sentiment analysis methods employed to infer user voting intention.

would be the one chosen underperformed the random classifier. What is more intriguing is that the Vote & Flip method, which is more elaborate than the one which simply counts the number of polarized terms, underperformed it when it came to McCain. Finally, only two methods outperformed the random classifier with regard to precision: Polarity Lexicon and Semantic Orientation. The former is better with regards to estimating McCain support and global accuracy and, hence, it was chosen to infer votes for all of the users in the dataset.

Of course, no real application could rely on such poor classifiers; however, the study was continued in order to find what other lessons could be obtained. The first lesson was that Sentiment Analysis is a difficult challenge and one should exercise extreme caution when assuming a naïve classifier can do the work.

Method	ΔPrecision Obama	ΔPrecision McCain	ΔAccuracy
Most frequent candidate	-4.8%	-41.8%	-34%
Polarity lexicon	**2.5%**	**32.1%**	-19.4%
Vote & Flip	7%	-20.1%	-34.1%
Semantic Orientation	**6.6%**	**16.4%**	-52.2%

Table 4: Differences in performance when comparing each of the methods against a perfectly informed random classifier (i.e. one assigning a vote to Obama with a 0.866 probability, and to McCain with 0.134; such a method would achieve 86.6% and 13.4% precision for each candidate and 76.8% accuracy).

5 The 2008 U.S. Presidential Elections according to Twitter data

Table 5 reveals the failure when predicting the 2008 U.S. Presidential Elections from Twitter data. The Mean Absolute Error (MAE) is 13.10% for the

State	Actual % of Obama votes	% of Twitter "votes"	Twitter error	% of TwitVote "votes"	TwitVote error
California	62.28%	62.70%	0.42%	91.89%	29.61%
Florida	51.42%	66.20%	14.78%	81.32%	29.90%
Indiana	50.50%	64.70%	14.20%	87.88%	37.38%
Missouri	50.07%	68.10%	18.03%	83.61%	33.54%
N. Carolina	50.16%	66.60%	16.44%	98.38%	48.22%
Ohio	52.31%	59.80%	7.49%	86.57%	34.26%
Texas	44.06%	64.40%	20.34%	76.97%	32.91%
		MAE	13.10%	MAE	35.12%

Table 5: Prediction of the 2008 U.S. Presidential Elections according to data collected in Twitter and to the subset of users who issued a vote in TwitVote. The MAE is very large and a victory for Obama in Texas is predicted. Nevertheless, the prediction using tweets is substantially better than the direct-poll conducted by TwitVote.

prediction based on Twitter data, and 35.12% for TwitVote[12].

Something was decidedly wrong with this and, therefore, it deserved a thorough analysis. The error could of course be attributed to the collected data but this is probably not the case because the volume of tweets and users is highly correlated with the populations of the respective states, and the conversation exhibits bursts in key moments of the campaign. Indeed, given that the classifier largely overestimates McCain support and yet Obama leads the results, it seems quite reasonable to assume that *self-selection* bias was tainting the sample. Two plausible hypotheses could explain that bias in Twitter:

- Urbanites and young adults are more prone to use Twitter and they have a tendency towards liberal political opinions.
- Republican voters use Twitter less than Democratic voters or they are reluctant to publicly express their political opinions (the so-called "Shy Republican" factor).

To test the first hypothesis, the study relied on the number of users per county in addition to the counties' population and population density. In this way it was possible to look for any correlation between the percentage of users in a county and its population density. Using the actual results for the elections on each of those counties it was also possible to look for any correlation between densely populated areas and a tendency towards Democratic vote.

All of the states show a positive correlation between population density and Democratic vote in these particular elections (Table 6). Moreover, all of the states except for Missouri and Texas, exhibit a positive correlation between

[12]According to [6], 8 out of 17 national phone polls predicted the final margin for these elections with an error below 1% and most of the others below 3%. Thus, results achieved by exploiting Twitter data are still far less accurate than those achieved with traditional polling.

	Twitter users vs Population density	Democratic vote vs Population density
California	0.9452	0.4069
Florida	0.1768	0.4740
Indiana	0.2956	0.5452
Missouri	**-0.0079**	0.5239
N. Carolina	0.5425	0.3968
Ohio	0.6343	0.5676
Texas	**-0.0535**	0.4789

Table 6: Correlation (Pearson's r) between the percentage of users in a county and its population density, and between population density and Democratic vote in the 2008 U.S. Presidential Elections.

population density and Twitter usage. *Hence, it seems that the collected sample over-represents urban voters*[13] *who were more prone to vote for Obama.*

With regard to user age, it should be noted that Twitter does not record birth date. Nevertheless, using the users' names and both their county and location, it was possible to found the age of about 2,500 users in online public records. It was discovered that 18-29 year old users amounted for 23.7% of the total, and those in the 30-44 interval amounted for 54.5%. This contrasted with the age distribution in the elections where both groups amounted for 18% and 29%, respectively. Thus, it is clear that younger people are over-represented in Twitter, and in this particular case it can explain part of the error[14].

To test this possibility, a prediction was made using the users with a known age and weighting their votes accordingly to the participation of each age group in the 2004 and 2008 elections. The MAE for the age corrected predictions is 11.6% against the 13.1% of the original one (Table 7). Hence, although Twitter data overestimate the opinion of younger users, it is possible to correct that, provided that the actual age distribution was known.

With regards to a hypothetical different behavior in Republican voters (i.e. using Twitter less than Democratic voters or not discussing their political views), little can be said with the data at hand. Given the uneven support for Obama not only in the collected dataset but also in TwitVote, it seems pretty clear that Republicans, or at least McCain supporters, tweeted much less than Democratic voters during the 2008 elections. This is consistent with the findings of [11] who argued that, because of the prevalence of younger users and their tilt toward Democrats and Obama:

> *"Democrats and Obama backers are more in evidence on the Internet than backers of other candidates or parties."*

[13] This is consistent with the findings of [9], who report that 35% of Twitter users live in urban areas and only 9% live in rural areas.

[14] This is, again, consistent with the findings of [9] who said that *"Twitter users are overwhelmingly young"* although *"Twitter use is not dominated by the youngest of young adults"*. In addition to that, according to [11] *"young voters tilt toward Obama specifically and towards the Democrats generally"* and they *"stand out compared with their elders based on their creation of political commentary and writing"*. All of this could justify part of the bias.

State	Actual % of Obama votes	Twitter votes age-corrected according to 2004 participation	Error	Twitter votes age-corrected according to 2008 participation	Error
California	62.28%	62.5%	0.22%	62.5%	0.22%
Florida	51.42%	63.6%	12.18%	63.3%	11.88%
Indiana	50.50%	59.1%	8.6%	59.3%	8.8%
Missouri	50.07%	66.9%	16.83%	67.1%	17.03%
N. Carolina	50.16%	68.2%	18.04%	68.4%	18.24%
Ohio	52.31%	58.4%	6.09%	58.1%	5.79%
Texas	44.06%	63.4%	19.34%	63.5%	19.44%
		MAE	11.61%	MAE	11.63%

Table 7: Statistically correcting the age bias taking into account the users age and the different participation of each age group in the 2004 and 2008 elections.

6 Some lessons from the "fiasco"

In short, the 2008 U.S. Presidential Elections could not have been accurately predicted from Twitter by applying the most common current methods. This finding is consistent with that of [10] who did not find any substantial correlation between a sentiment analysis of tweets and several pre-election polls conducted during the campaign. In addition, the possible biases in the data are consistent with the findings of [9, 11].

Hence, the problem with predicting the outcomes of these elections was not in the data collection. Instead, the problem occurred in minimizing the importance of bias in Social Media data and by ignoring how such data differs from the actual population. Several lessons can be learned from this:

1. *The Big Data fallacy.* Social Media is very appealing because researchers can obtain large data collections to be mined. Nevertheless, *just being large does not make such collections statistically representative of the population as a whole.*

2. *Watch out for demographic bias.* In much the same vein as the first lesson: Social Media users tend to be relatively young and, depending on the population of interest, this can introduce an important bias. *To improve results it is necessary to know user age and try to correct for the bias in the data.*

3. *Beware of naïve sentiment analysis.* It is possible that some applications can achieve reasonable results by merely accounting topic frequency or using simple approaches to sentiment detection. Nevertheless, as shown in this paper, *noisy instruments should be avoided* and researchers must

always carefully check whether or not they are using a random classifier. *Moreover, texts of political nature are specially difficult to deal with* [17].

4. *Silence speaks volumes.* Non-responses can play an even more important role than the collected data. If the lack of information mostly affects only one group the results can considerably differ from reality. Needless to say, estimating the degree and nature of non-response is very difficult –if not completely impossible– and therefore researchers must be wary of the hazards that it involves.

5. *(A few) Past positive results do not guarantee generalization.* Researchers should always be aware of the *file drawer* effect, and should carefully evaluate positive reports before assuming the reported methods can be straightforwardly applied to any similar scenario with identical results. This becomes especially important if there are any counterexamples, like the one detailed in this study.

In summary, until Social Media becomes regularly used by the vast majority of people, its users cannot be considered a representative sample and, thus, forecasts from such data will be of questionable value at best and incorrect in many cases. Until then, if using such data it is necessary to identify the different strata of users –based on age, income, gender, race, etc.– in order to weight their opinions according to the proportion of each stratum in the population.

Acknowledgments This work was partially financed by grant UNOV-09-RENOV-MB-2 from the University of Oviedo. The author would like to thank Iván Menéndez-González for his work on a Twitter crawler prototype, in addition to Brendan O'Connor and Andranik Tumasjan for providing early copies of their papers. He is in debt to Miguel Fernández-Fernández, Rodrigo García-Suárez, and Tensi Fernández-Cuervo for their insightful discussions, and to the Associate Editor Abigail Sellen and the three anonymous reviewers for their valuable comments.

References

[1] Sitaram Asur, and Bernardo A. Huberman, 2010. Predicting the Future with Social Media. Technical report. Available at: `http://arxiv.org/abs/1003.5699v1`

[2] Erik Boiy, Pieter Hens, Koen Deschacht, and Marie-Francine Moens, 2007. Automatic Sentiment Analysis in On-line Text. In Proceedings of ELPUB2007 Conference on Electronic Publishing, pp. 349–360.

[3] Hyunyoung Choi, 2009. Predicting the Present with Google Trends. Technical report. Available at: `http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf`

[4] Yejin Choi, and Claire Cardie, 2009. Adapting a Polarity Lexicon using Integer Linear Programming for Domain-Specific Sentiment Classification. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 590–598.

[5] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant, 2009. Detecting influenza epidemics using search engine query data. Nature, vol. 457, pp. 1012–1014.

[6] Scott Keeter, Jocelyn Kiley, Leah Christian, and Michael Dimock, 2009. Perils of Polling in Election '08. Pew Internet and American Life. Available at: http://pewresearch.org/pubs/1266/polling-challenges-election-08-success-in-dealing-with

[7] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, 2010. What is Twitter, a Social Network or a News Media? In Proceedings of the 19th International World Wide Web (WWW) Conference, April 26-30, 2010, Raleigh NC (USA).

[8] Amanda Lee Hughes, and Leysia Palen, 2009. Twitter Adoption and Use in Mass Convergence and Emergency Events. In Proceedings of the 6th International ISCRAM Conference – Gothenburg, Sweden, May 2009.

[9] Amanda Lenhart, Susannah Fox, 2009. Twitter and status updating. Pew Internet and American Life. Available at: http://www.pewinternet.org/Reports/2009/Twitter-and-status-updating.aspx

[10] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith, 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media.

[11] Aaron Smith, and Lee Rainie, 2008. The Internet and the 2008 election. Pew Internet and American Life. Available at: http://www.pewinternet.org/Reports/2008/The-Internet-and-the-2008-Election.aspx

[12] Peverill Squire, 1988. Why the 1936 Literary Digest Poll Failed. Public Opinion Quarterly, vol. 52, no. 1, pp. 125–133.

[13] Bill Tancer, 2008. Click: What millions of people are doing online and why it matters. Hyperion.

[14] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe, 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In Proceedings

of the 4th International AAAI Conference on Weblogs and Social Media.

[15] Peter D. Turney, 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424.

[16] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of HLT-EMNLP-2005, pp. 347–354.

[17] Bei Yu, Stefan Kaufmann, and Daniel Diermeier, 2008. Exploring the characteristics of opinion expressions for political opinion classification. In Proceedings of the 2008 international conference on Digital government research, pp. 82–91.