DOI:10.1145/2001269.2001297

The power to predict outcomes based on Twitter data is greatly exaggerated, especially for political elections.

BY DANIEL GAYO-AVELLO

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

CONTENT PUBLISHED IN microblogging systems like Twitter can be data-mined to take the pulse of society, and a number of studies have praised the value of relatively simple approaches to sampling, opinion mining, and sentiment analysis. Here, I play devil's advocate, detailing a study I conducted late 2008/early 2009 in which such simple approaches largely overestimated President Barack Obama's victory in the 2008 U.S. presidential election. I conducted a thorough post-mortem of the analysis, extracting several important lessons.

Twitter is a microblogging service for publishing very short text messages (only 140 characters each), or tweets, to be shared with users following their author. Many Twitter users do not protect their tweets, which then appear in the so-called public timeline. They are accessible through Twitter's own API, so are easily accessed and collected.

Twitter's original slogan—"What are you doing?"—encouraged users to share updates about the minutia of their daily activities with their friends. Twitter has since evolved into a complex information-dissemination platform, especially during situations of mass convergence.[8] Under certain circumstances, Twitter users not only provide information about themselves but also real-time updates of current events.[a]

Today Twitter is a source of information on such events, updated by millions of users[b] worldwide reacting to events as they unfold, often in real time. It was only a matter of time before the research community turned to it as a rich source of social, commercial, marketing, and political information.

My aim here is not a comprehensive survey on the topic but to focus on one of its most appealing applications: using its data to predict the outcome of current[c] and future events.

Such an application is natural in light of the excellent results obtained

» key insights

- Using social media to predict future events is a hot research topic involving multiple challenges, including bias in its many forms.
- Researchers' behavior can also be biased as they may not always report negative results while assuming conclusions from a few selected positive experiments.
- Ignoring negative results, researchers risk converting social media analysis into another *Literary Digest* poll (as in the 1936 U.S. presidential election), risking any future research into this kind of analysis.

[a] The 2008 Mumbai attacks and 2009 Iranian election protests are perhaps the best-known examples of Twitter playing such a role.

[b] As of mid-2009, Twitter reportedly had 41.74 million users.[7]

[c] Bill Tancer of Hitwise said predicting ongoing events should not be defined as "prediction" but rather as "data arbitrage."[13]

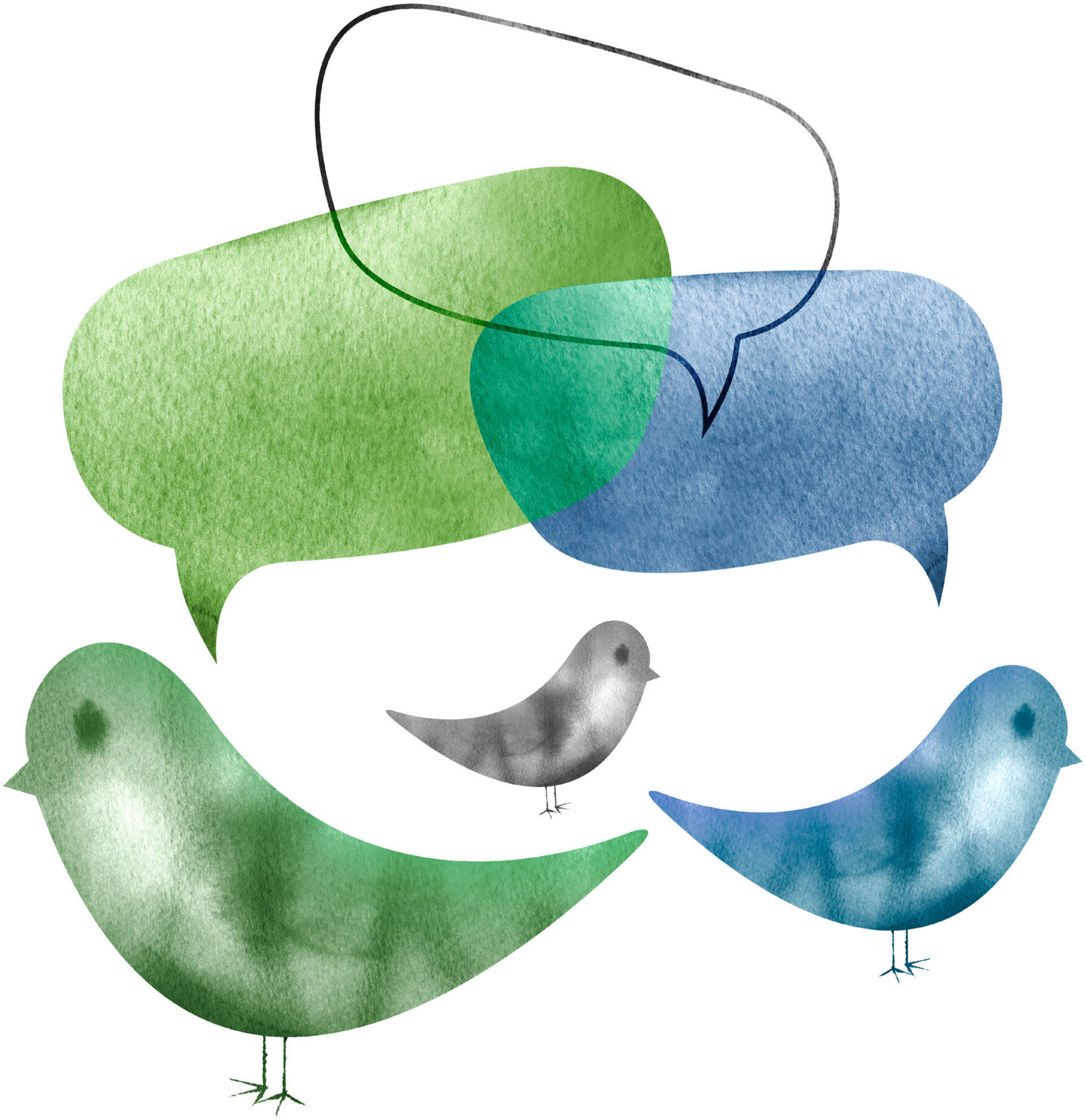


ILLUSTRATION BY CREATIVE IMPROV/SHUTTERSTOCK.COM

by mining query logs,[3,5] with a number of studies covering the topic. For example, in 2010, Asur and Huberman[1] used Twitter data to predict box-office revenue for movies; also in 2010, O'Connor et al.[10] correlated Twitter data with several public-opinion time series[d]; and Tumasjan et al.[14] claimed to have predicted the outcome of the 2009 German federal election by counting the number of mentions each candidate received on Twitter.

Such studies have been well received and generated a fair amount of attention, particularly on the theoretical possibility of predicting elections. Two informal experiments claimed the last electoral outcomes in the U.K. in 2010 and Belgium in 2010 were accurately predicted through Twitter data.[e]

Such reports seem to imply that predicting future events through Twitter data is fairly straightforward. Nevertheless, as I explore here, such direct correlation is simply not the case.

Can't (Always) Predict

As of December 2008, 11% of U.S. adults online were using Twitter and analogous services.[9] While that percentage is significant, the fact is the vast majority of Internet users, as well as people worldwide, simply do not use Twitter. Twitter users are just a sample, probably a biased one.

Another kind of bias permeates re-

[d] These authors found a correlation between Twitter data and consumer-confidence indices and presidential job-approval ratings, though no substantial correlation was found between Twitter data and data from polls during the 2008 U.S. presidential campaign.

[e] See 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

search—the tendency of researchers to report positive results while suppressing negative results. This so-called "file-drawer effect" can have a harmful influence if it is assumed that conclusions from a few selected positive experiments are directly applicable to any other conceivable scenario.

Twitter users are just a sample, probably a biased one.

It is 75 years since the famously ill-fated 1936 *Literary Digest* presidential poll predicting the outcome of the U.S. presidential election of 1936. Conducted among the magazine's own readers, people listed in the telephone directory nationwide, and a list of registered car owners, the poll concluded that the Republican candidate, Alf Landon, would win in a landslide. Roosevelt ultimately won with 61% of the popular vote.[f]

Ignoring negative results, current research risks turning social media analytics into the next *Literary Digest* poll. Here, I cover one such negative result—the experiment I conducted involving a large collection of tweets published during the 2008 U.S. presidential campaign predicting Obama would win every battleground state, as well as Republican stronghold Texas.

As with the *Literary Digest* poll, my experiment could be dismissed, attributing its failure to poor sampling methods or defects in the system that assigned voting intention to user tweets or even to bias and stereotypes regarding the political views of Twitter users.

Due to the nature of the experiment, the sampling was biased, but every prediction inferred from social media—even those with positive results—involve analogous bias. The sentiment analysis I performed was naïve, but even simpler systems have proved sound enough to achieve positive predictive results. Moreover, no matter how appealing ideological bias might be explaining this outcome, such a hypothesis must first be rigorously tested.

Beyond the two papers[10,14] mentioned earlier on predicting election outcomes, what could the present study contribute to the matter?

First, note that the findings of the earlier studies could seem to contradict one another. Tumasjan et al.[14] claimed the number of tweets mentioning a candidate was a reflection of vote share expected in the election under study and would have predictive power close to traditional polls. O'Connor et al.[10] found no substantial correlation between a much more complex sentiment analysis performed on Twitter data and several polls conducted during the 2008 U.S. presidential campaign.

Nevertheless, because both studies dealt with very different political scenarios (Tumasjan et al. on elections in Germany and O'Connor et al. on elections in the U.S.) and both used different kinds of ground-truth data (Tumasjan et al. comparing data with election results, the popular vote, and O'Connor et al. using pre-election polls, not actual election results), it is difficult to say if such predictions are possible from Twitter data alone. Moreover, even if they are possible, there are still serious questions as to what conditions would be required to make them.

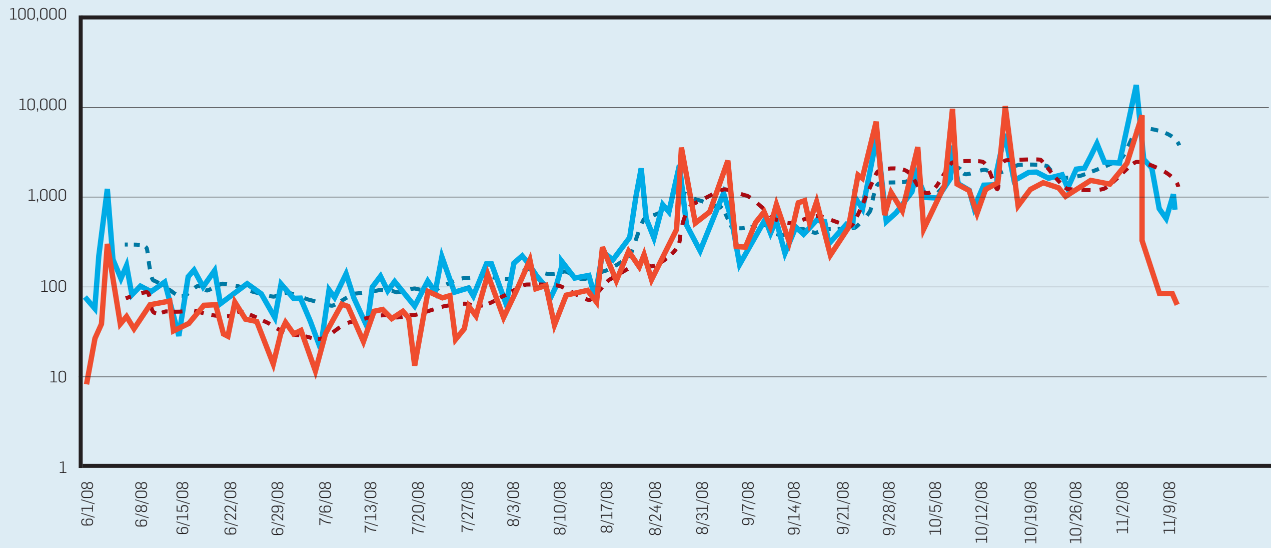
My aim here is to provide a balanced view of the real predictive possibilities of social media analytics. For that reason, my study involved a much more detailed analysis of electoral prediction from Twitter data than in the earlier studies.

Furthermore, my aim was not to compare Twitter data with pre-election polls or with the popular vote, as had been done previously, but to obtain predictions on a state-by-state basis. Additionally, unlike the other studies, my predictions were not to be derived from aggregating Twitter data but by detecting voting intention for every single user from their individual tweets. I applied four different sentiment-analysis methods described in the most recent literature and carefully evaluated their performance.

As I show here, the results for the 2008 U.S. presidential election could not have been predicted from Twitter data alone through commonly applied methods. While this conclusion is consistent with some of the results obtained by O'Connor et al., I went a step further by clarifying the nature of the failure (large overestimation of Obama's vote share), with a thorough analysis of its possible causes (such as urbanization

[f] The biased sample has been blamed as the source of the flawed result, though Squire[12] found that the biggest issue of the 1936 poll was not the biased sample but the nonresponse bias.

Time series depicting the volume of tweets regarding each of the two main parties' national tickets, Obama/Biden in light blue, McCain/Palin in red; dashed lines are seven-day moving averages.



and age demographics or, even possibly a "shy Republican" factor).

The lesson is clear, that researchers must be cautious about simplistic assumptions regarding forecasting based on so-called "big data" in general and on Twitter data in particular.

Election Twitter Data Set

For my study, I began collecting tweets shortly after the 2008 election to check the feasibility of using Twitter to predict future election outcomes. Using the Twitter search API, I included one query for each candidacy: `obama OR biden` for the Democratic candidates and `mccain OR palin` for the Republican candidates.

An API parameter to indicate a geographical area was used to consider only tweets published by U.S. residents; in addition, another API parameter was used to indicate a temporal interval for the query. Thus, by issuing queries limited both geographical and temporally, I obtained 100 tweets per candidate, per county, per day.

Doing this for every county in the U.S. would have involved submitting a large number of HTTP requests to Twitter's servers. The number of daily requests one IP address can submit is limited, and, more important, the Twitter index does not contain all published tweets but rather those within a sliding time frame. This meant it was critical for me to get the data as soon after it was tweeted as possible; I thus focused the collection on only a few selected states: one traditional stronghold state for each party (California for the Democrats, Texas for the Republicans), along with the six swing states (Florida, Indiana, Missouri, Montana, North Carolina, and Ohio).

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

The first thing that had to be checked was whether or not the data set could be considered a statistical representative sample.[g] I thus compared the number of tweets and unique users in each state to their populations. In ad-

[g] Note the data-collection method introduced a sample-selection bias; only a fraction of the Twitter firehose is provided as search results, and not every Twitter user provides a location, only about 50% of the profiles, according to data I collected in 2009.

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 sample (provided they were actually random).

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%

* Note: Data discarded due to large margin of error.

dition, sampling errors were computed on the assumption the collection was close to a random sample. The correlation between population and numbers of tweets and users was almost perfect (Pearson's *r* coefficients were 0.9604 and 0.9897, respectively), and all samples except Montana exhibited a fairly low sampling error (see Table 1). Hence, I discarded Montana.

After this preliminary analysis, I plotted a time series for each set of candidates, in addition to a seven-day moving average for each (see the accompanying figure). The plot exhibits peaks corresponding to relevant events, including 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 and October 7 and 15), the vice-presidential debate (October 2), and Election Day (November 4). The number of tweets, September to November, thus seemed consistent with a reasonable sampling; the amount of "conversation" through Twitter grew as the campaign progressed, showing bursts during important events and finally dropping off after Election Day.

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

This collection seemed to be a valid sample, following the trajectory of the national polls; moreover, there was a strong correlation between the volume of users and tweets from each state and its population. All this might seem to suggest an accurate sampling, and, given that the number of people involved in this data set was much larger than the samples in national polls, one might expect even greater accuracy, but that view would ultimately prove to be wrong.

Inferring Voter Intention

Despite the extensive literature on automatic sentiment analysis,[2,4] virtually all current research on microblogging analysis relies on rather simple methods. For my purpose, I applied four different methods: one based on mention counts, two based on polarity lexicons, and one based on the semantic-orientation method.[15]

The idea behind the first was 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 later vote for; this heuristic is pretty coarse but, interestingly, seemed to work predicting the outcome of elections in Germany, leading Tumasjan et al.[14] to write "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] consisting of a list of terms labeled either positive or negative; a tweet was labeled positive if it contained more positive than negative terms and vice-versa. Because each tweet in the collection concerned just one candidate it was possible to count, for each user, the number of positive and negative tweets for each set of candidates. I therefore

Table 2. Selected supportive and opposing phrases for both presidential candidates obtained through semantic orientation.

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

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%

[h] This was also the first time candidate John McCain was ahead in the national polls in the 2008 presidential campaign.

supposed a user would vote for the candidate with the highest score. Employing a similar method with mixed results, O'Connor et al.[10] wrote "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, called "Vote & Flip"[4] was also used, consisting basically of counting the number of positive, negative, neutral, and negation words in a sentence to later apply a set of rules to infer the sentence's polarity.

Finally, I also adapted semantic orientation.[15] The original approach consisted of finding phrases with either positive or negative polarity. Such a value was based on an estimation by means of a search engine of the Pointwise Mutual Information, a measure of semantic association based on word co-ocurrence between the phrase and the keywords "poor" and "excellent." However, the implemented version differed from the original in that it did not rely on either a search engine or on the pair "poor/excellent" but on a subset of tweets published by users who had clearly stated their voting intentions.[i]

Table 2 lists selected phrases the method found to be either in support or in opposition of each set of candidates; as expected, the patterns selected to build the subset appeared as top ranked while also revealing other useful patterns.

To evaluate the performance of each method, I also needed the actual votes of the users. An informal opinion-poll was conducted during the campaign by a Web site called TwitVote[j] asking users to declare their votes by publishing a tweet containing both their vote and the hashtag `#twitvote`. Thus, by collecting tweets published on November 4 and tagged with `#twitvote` I was able to identify the actual votes of a number of users.

Only 2,000 users (9% of my data set) used TwitVote, among whom 86.6% voted for Obama, with the rest for McCain.[k] These results, so different from the actual popular vote, did not bode well for the study because they apparently revealed a large bias in Twitter users toward the Obama/Biden ticket. Nonetheless, I used the data to evaluate the performance of each of the methods to infer user voting intention, ultimately proving they were inadequate (see Table 3).

Precision inferring Twitter-user support for Obama was rather high but poor with regard to McCain. Even more perplexing was that different methods achieved similar results. Indeed, all the methods seemed to drift toward a random classifier. This by itself was a bad sign, but in terms of relative performance, I found it reasonable to compare all the methods with a perfectly informed random classifier: one assigning voting intention with regard to the proportion of "votes" according to TwitVote.

Assuming the most frequently mentioned candidate would be the one ultimately chosen by users underperformed the random classifier (see Table 4). More intriguing was that the Vote & Flip method, which is more elaborate than the one that simply counts the number of polarized terms, underperformed it when it came to McCain. Finally, only two methods—Polarity Lexicon and Semantic Orientation—outperformed the random classifier with regard to precision. Because the former was better estimating McCain support and global accuracy, I used it to infer votes for all users in the data set.

I understood that no real-world application could rely on such poor classifiers but continued my study for other reasons. The first was that "sentiment analysis" is a difficult challenge requiring extreme caution when assuming a naïve classifier can do the work.

Presidential Election According to Twitter Data

Table 5 reflects the failure of simple sentiment analysis when trying to pre-

Table 4. Difference in performance between each method and a perfectly informed random classifier; that is, one assigning a vote to Obama with a 0.866 probability and to McCain with 0.134 probability. Such a method achieved 86.6% and 13.4% precision for each candidate and 76.8% accuracy.

Method	▲ Precision Obama	▲ Precision McCain	▲ Accuracy
Most frequent candidate	–4.8%	–41.8%	–34%
Polarity lexicon	2.5%	32.1%	–19.4%
Vote and Flip	7%	–20.1%	–34.1%
Semantic Orientation	6.6%	16.4%	–52.2%

Table 5. Prediction for the 2008 U.S. presidential election based on data collected in Twitter and the subset of users voting in TwitVote. The MAE is large, predicting victory for Obama, even in Texas. Nevertheless, the prediction using tweets was substantially better than the direct-poll conducted by TwitVote.

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%

[i] Tweets from users containing the phrases "I will vote for…," "I'm not voting…," and "I'd vote…"

[j] See http://twitvote.twitmarks.com/

[k] This result is no different from the results tallied by TwitVote: 85.9% Obama vs. 14.1% McCain.

dict the 2008 U.S. presidential election solely from Twitter data. The mean absolute error, or MAE, was 13.10% for the prediction based on Twitter data and 35.12% for TwitVote.[l]

Something was clearly wrong and so deserved a thorough analysis. The error could be attributed to the collected data but was probably not the case because the volume of tweets and users was highly correlated with the populations of the respective states or with the conversation exhibiting bursts of activity at key moments in the campaign. Indeed, given that my classifier largely overestimated McCain support (yet Obama came out on top), it seemed reasonable to assume self-selection bias had tainted the sample. Two hypotheses might explain such bias in Twitter:

- Urbanites and young adults were more likely to use Twitter and had a tendency toward liberal political opinions; and
- Republican voters used Twitter less than Democratic voters or were reluctant to express their political opinions publicly, reflecting the so-called "shy Republican" factor.

To test the first, I relied on the number of users per county, in addition to each county's population and population density. This way, I was able to look for correlations between percentage of users per county and population density. Using the actual results for the elections in each county I was also able to look for correlations between densely populated areas and a tendency toward Democratic voting.

All the states showed a positive correlation between population density and the Democratic vote in the 2008 U.S. presidential election (see Table 6). Moreover, all the states, except Missouri and Texas, reflected a positive correlation between population density and Twitter use. Hence, it seemed the collected sample overrepresented urban voters[m] who were more likely to vote for Obama.

Table 6. Correlation (Pearson's *r*) between percentage of users per county and population density per county and between population density and Democratic vote in the 2008 U.S. presidential election.

	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 7. Statistically correcting age bias for user age and participation of each age group in the 2004 and 2008 U.S. presidential elections.

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%

With regard to user age, Twitter neither solicits nor collects user birth date. However, using user name, along with county and location, I was able to identify the age of about 2,500 users in online public records; for example, 18–29-year-old users represented 23.7% of the total, and 30–44-year-old users represented 54.5% of the total. This contrasted with the age distribution in the overall voter population in the 2008 U.S. presidential election, where these groups were 18% and 29%, respectively. Thus, in 2008, younger people were clearly overrepresented in Twitter and explains part of the prediction error[n] in my study.

To test the idea of overrepresentation of younger voters in Twitter, I computed a second prediction based on users of known age, weighting their votes according to each group's participation in the 2004 and 2008 U.S. presidential elections. The MAE for the age-corrected predictions was 11.6% against 13.1% of the original one (see Table 7). Hence, though my collected Twitter data overrepresented the opinion of younger users, it is possible to correct such overestimation, provided the actual age distribution is known.

With regard to a hypothetical difference in behavior between Republican and Democratic voters (with Republican voters using Twitter less than Democratic voters or not discussing their political views publicly), little can be said with certainty based solely on the data at hand. Given the uneven support for Obama in TwitVote, as well as in the collected data set, it seems clear that

[l] Keeter et al.[6] wrote that eight out of 17 national phone polls during the 2008 campaign predicted the final margin for the election within one percentage point and most of the others within three percentage points. Thus, polling results achieved by analyzing Twitter data were still far less accurate than the predictive results achieved through traditional polling methods.

[m] This pattern was consistent with Lenhart and Fox[9] reporting that in 2009 35% of Twitter users lived in urban areas and only 9% in rural areas.

[n] This pattern was consistent with Lenhart and Fox[9] saying "Twitter users are overwhelmingly young," though "Twitter use is not dominated by the youngest of young adults." Smith and Rainie[11] said "Young voters tilt toward Obama specifically and towards the Democrats generally" and "stand out compared with their elders based on their creation of political commentary and writing," possibly justifying some of the bias.

Republicans, or at least McCain supporters, tweeted much less than Democratic voters during the 2008 election. This is consistent with Smith and Rainie,[11] writing that, due 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."

Lessons Learned

The outcome of the 2008 U.S. presidential election could not have been predicted from user content published through Twitter by applying the most common current sentiment analysis methods. This finding is consistent with O'Connor et al.[10] who did not find substantial correlation between a sentiment analysis of tweets and several pre-election polls during the campaign. In addition, the possible biases in the data were consistent with Lenhart and Fox[9] and Smith and Rainie.[11]

The problem with trying to predict the outcome of the 2008 U.S. presidential election was not data collection per se but how to minimize the importance of bias in social-media data and ignore how such data differs from the actual population, yielding several lessons:

Big-data fallacy. Social media is appealing because researchers can assemble large data collections to be mined. Nevertheless, just being large does not make such collections statistically representative samples of the overall population;

Beware demographic bias. Users of social media tend to be relatively young and, depending on the population of interest, can introduce important bias. To improve results researchers must know user age and try to correct for bias in the data;

Beware of naïve sentiment analysis. Some applications might achieve reasonable results by accounting for topic frequency or using simple approaches to sentiment detection. Nevertheless, as I've explored here, researchers should avoid noisy instruments and always check whether they are using a random classifier; dealing with political texts is especially difficult[17];

Silence speaks volumes. Nonresponses often play a more important role than collected data. If lack of information affects mostly only one group the results might differ considerably from reality. Estimating the degree and nature of a nonresponse is difficult if not impossible, so researchers must be wary of the related hazards; and

Past positive results do not guarantee generalization. Researchers should always be aware of the file-drawer effect and carefully evaluate positive reports before assuming the reported methods are straightforwardly applicable to any similar scenario with identical results.

Until social media is used regularly by a broad segment of the voting population, its users cannot be considered a representative sample, and forecasts from the data will be of questionable value at best and incorrect in many cases. Until then, researchers using such data should identify the various strata of users—based on, say, age, income, gender, and race—to properly weight their opinions according to the proportion of each of them in the population.

Acknowledgments

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

References

1. Asur, S. and Huberman, B.A. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Toronto, Aug. 31–Sept. 3). IEEE Computer Society, Los Alamitos, CA, 2010, 492–499.
2. Boiy, E., Hens, P., Deschacht, K., and Moens, M.F. Automatic sentiment analysis in online text. In *Proceedings of the 2007 Conference on Electronic Publishing* (Vienna, June 13–15). ÖKK Editions, Vienna, 2007, 349–360.
3. Choi, H. *Predicting the Present with Google Trends.* Tech. Rep. Google, Inc., Mountain View, CA, 2009; http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf
4. Choi, Y. and Cardie, C. 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* (Singapore, Aug. 6–7). Association for Computational Linguistics, Stroudsburg, PA, 2009, 590–598.
5. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature 457*, 7232 (Feb. 19, 2009), 1012–1014.
6. Keeter, S., Kiley, J., Christian, L., and Dimock, M. *Perils of Polling in Election '08.* Pew Internet and American Life Project, Washington, D.C., 2009; http://pewresearch.org/pubs/1266/polling-challenges-election- 08-success-in-dealing-with
7. Kwak, H., Lee, C., Park, H., and Moon, S. What is Twitter: A social network or a news media? In *Proceedings of the 19th International World Wide Web Conference* (Raleigh, NC, Apr. 26–30). ACM Press, New York, 2010, 591–600.
8. Hughes, A.L. and Palen, L. Twitter adoption and use in mass convergence and emergency events. In *Proceedings of the Sixth International Community on Information Systems for Crisis Response and Management Conference* (Gothenburg, Sweden, May 10–13, 2009).
9. Lenhart, A. and Fox, S. *Twitter and Status Updating.* Pew Internet and American Life Project, Washington, D.C. 2009; http://www.pewinternet.org/Reports/2009/Twitter-and-status-updating.aspx
10. O'Connor, B., Balasubramanyan, R., Routledge, B.R., and Smith, N.A. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media* (Washington, D.C, May 23-26). Association for the Advancement of Artificial Intelligence, Menlo Park, CA, 2010, 122–129.
11. Smith, A. and Rainie, L. *The Internet and the 2008 Election.* Pew Internet and American Life Project, Washington, D.C., 2008; http://www.pewinternet.org/Reports/2008/The-Internet-and-the-2008-Election.aspx
12. Squire, P. Why the 1936 *Literary Digest* poll failed. *Public Opinion Quarterly 52*, 1 (Spring 1988), 125–133.
13. Tancer, B. *Click: What Millions of People Are Doing Online and Why It Matters.* Hyperion New York, 2008.
14. Tumasjan, A., Sprenger, T.O., Sandner, P.G., and Welpe, I.M. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media* (Washington, D.C., May 23–26). Association for the Advancement of Artificial Intelligence, Menlo Park, CA, 2010, 178–185.
15. Turney, P.D. 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* (Philadelphia, July 6–12). Association for Computational Linguistics, Stroudsburg, PA, 2002, 417–424.
16. Wilson, T., Wiebe, J., and Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technology Conference, Conference on Empirical Methods in Natural Language Processing* (Vancouver, Canada, Oct. 6–8). Association for Computational Linguistics, Stroudsburg, PA, 2005, 347–354.
17. Yu, B., Kaufmann, S., and Diermeier, D. Exploring the characteristics of opinion expressions for political opinion classification. In *Proceedings of the 2008 International Conference on Digital Government Research* (Montréal, May 18–21). Digital Government Society of North America, Marina del Rey, CA, 2008, 82–91.

Daniel Gayo-Avello (dani@uniovi.es) is an associate professor in the Department of Computer Science of the University of Oviedo, Spain.