



No, You Cannot Predict Elections with Twitter

Daniel Gayo-Avello • University of Oviedo, Spain

In 2006, I attended a series of seminars in Madrid at which Isabel Aguilera, Google's managing director for Spain and Portugal at the time, was going to deliver an invited speech. According to one organizer, Aguilera told them in *petit comité* that at Google, they'd known beforehand the winner of the Real Madrid Club de Fútbol elections.

Unlike most professional football clubs in Spain, Real Madrid is member-owned and operated, so it periodically elects a club president. The elections were held only a few months before the seminars; Ramón Calderón was the winner.

My "informant" couldn't provide more details because Aguilera was rather cryptic in her comments. However, we both supposed that she was likely referring to the volume of queries for each candidate and assuming that the winner would be the one with the highest volume.

Once back home I tried to replicate those results using the new-at-the-time Google Trends service. The query `calderon` was indeed far more frequent around the electoral period than those for the other candidates. Unfortunately, the world is a strange place: the very same day Ramón Calderón was running for president of Real Madrid, another Calderón – Felipe – was running for president of Mexico (and also winning the election).

So, which Calderón were Google's users looking for? This little anecdote raised more questions than answers, and I thought it could become a hot research topic.

When Data Gives You Lemons

Fast forward to 2008. The US was holding its presidential election, and the Internet was supposed to play a major role. I was determined to test Aguilera's claim by obtaining data from

Google's Insights for Search on both Barack Obama and John McCain. Unfortunately, doing so in an automated fashion was far from trivial, and comparing search volumes across states was extremely difficult. On top of this, as the Yahoo Search Blog stated,

While searches aren't votes, they do provide insight into what's on the collective mind of the electorate. [...] In the past week, Senator Obama has drawn more than twice as many queries as Senator McCain. That's not necessarily all positive, as the nature of the queries indicate that people still have a lot of questions – lookups range from questions about his biography to his birth certificate.¹

In other words, searching for a candidate was hardly a form of support. So, I needed a different kind of user-generated content: something richer than queries and thus amenable to sentiment analysis. Luckily, Twitter was already a rather well-known service, and Obama's supporters were using it heavily.

I decided to collect Twitter data to check for its predictive power as regards elections. Nevertheless, collecting tweets was far from easy (no streaming API back then); moreover, I wanted geo-located tweets to obtain county-level details. Instead of trying to collect tweets during the elections, therefore, I collected them afterward, but only from those states that had played a major role in the outcome.

Everything looked promising in that collection: the size of the samples strongly correlated with states' populations, the volume of tweets seemed to closely follow the polling trends, and the dataset ultimately indicated that Obama was the predicted winner.

There was, however, one minor problem: My data had predicted a landslide victory in every

single state, including Texas! Needless to say, I considered such findings unpublishable and forgot about them until early 2010, when several events occurred in short succession that prompted me to get that report out of the drawer and submit it under a different perspective.

First, danah boyd wrote an incisive post about big data and social sciences;² I especially liked this part:

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 see traces of data doesn’t mean you always know the intention or cultural logic behind them. And just because you have a big N doesn’t mean that it’s representative or generalizable.

Amen to that!

Second, Daniele Fanelli published a paper³ discussing the bias toward positive results in scientific publications and mentioning the infamous “file-drawer effect” – that is, researchers’ tendency to not report negative results.

Third, the 2010 International AAAI Conference on Weblogs and Social Media (ICWSM 10) accepted two different papers for publication. The first, by Andranik Tumasjan and his colleagues, made the flamboyant claim that “the mere number of tweets reflects voter preferences and comes close to traditional election polls.”⁴ The second, by Brendan O’Connor and his colleagues linked Twitter sentiment to consumer confidence and presidential job approval but didn’t find any strong correlation between Twitter sentiment and polls conducted during the 2008 presidential campaign.⁵

So, there I was with a complete report on how to predict a landslide victory that never happened; this report was consistent with an independent study and, simultaneously,

reached the opposite conclusion of a third one.

Needless to say, the problem here is overgeneralization. I wasn’t proving that you can’t predict elections by mining Twitter but instead that I wasn’t able to accurately predict the US 2008 elections. In the same vein, the Tumasjan study had just demonstrated that the researchers predicted one particular election in one particular country. Nevertheless the mantra “Twitter can predict elections” had begun.

The Emperor Has No Clothes

Thus, I decided to write a paper that

- dealt with the need to publish negative results,
- provided a post-mortem of a my failed social media study, analyzing the sources of bias and ways to correct them, and
- offered some lessons and caveats for future research in the field.

This paper was ultimately published in late 2011.⁶

In the meantime, although I was vocal in my skepticism on this matter, uber optimistic claims similar to that in the Tumasjan paper appeared repeatedly. I certainly wasn’t the only skeptic on this issue. In fact, I joined forces with Takis Metaxas and Eni Mustafaraj to test the reproducibility of such aforementioned works, finding numerous worrisome flaws.⁷ A few other researchers stepped forward to debunk the naïve assumptions on which “Twitter election predictions” were based.⁸ This had little effect, however.

In paper after paper, at conference after conference, we were told the impossible: that applying crude sentiment analysis methods to noisy data produced by a biased and self-selected sample of the population is amazingly accurate when predicting elections – and not only elections. You would expect some

healthy disbelief to develop within the community. Yet, with regard to this topic it seems that the burden of proof lies on those of us trying to explain that the emperor has no clothes.

As a matter of fact, I have some anecdotal evidence for this – the reviews for one of the papers I’ve collaborated on. In them, the anonymous referees commented that “unless a negative results paper is methodologically impeccable, it’s hard for its conclusions to be believed” or “by concentrating on the failure of a specific set of techniques, it is not obvious how the reader can take this as evidence of failure of the idea in general.”

Does this mean the opposite is instead valid? That is, that papers with positive results need not be methodologically impeccable to be believed, or that showing how a specific method has been valid once is evidence of validity for the idea in general? This is wishful thinking at best or cargo cult science at worst.

If we’re really committed to advancing this electoral prediction based on social media data, we must recognize and avoid the common flaws plaguing current research:

- It isn’t prediction at all. I haven’t found a single paper predicting a future result. They all claim that a prediction could have been made, but the analysis is post hoc. And needless to say, negative results are rare.
- Chance isn’t a valid baseline because incumbency tends to play a major role in most elections.
- No commonly accepted way exists for “counting votes” in Twitter. Current research has used the raw volume of tweets, unique users, and many flavors of sentiment analysis.
- No commonly accepted way exists for interpreting reality. Some papers compare the predicted

results with polls, others with popular vote, yet others using the percentage of representatives each party achieves, and so on.

- Sentiment analysis is applied as a black box and with naïveté. Indeed, most of the time, sentiment-based classifiers are only slightly better than random classifiers.
- All the tweets are assumed to be trustworthy; however, just because something's been tweeted doesn't mean it's true. Twitter is plagued with rumors, propaganda, and misleading information that are processed as valid political opinion.
- Demographics are neglected. Social media isn't a representative and unbiased sample of the voting population. Not every age, gender, social, or ethnic group is equally represented.
- Self-selection bias is simply ignored. People tweet on a voluntary basis and, therefore, only the politically active produce data.
- Past positive results don't guarantee generalization, especially if we account for the file-drawer effect.

To avoid such flaws, I suggest several recommendations.

Some Recommendations

Elections are occurring virtually all the time. If you're claiming to have a prediction method, you should predict an election in the future. Although I'm aware that conference schedules make publishing a prediction in a paper rather difficult, what about putting up a blog post 24 or 48 hours before the election?

Check the degree of influence incumbency plays in the elections you're trying to predict. Your baseline shouldn't be chance but predicting that the incumbent will win. Apply that baseline to prior elections; if

your method's performance isn't substantially better than the baseline, then you have a convoluted Rube Goldberg version of the baseline.

Clearly define what constitutes a "vote." Provide sound and compelling arguments supporting your definition. For instance, why use all of the users if some have only a few tweets on the topic? Conversely, why drop users because they have only a few tweets on the topic? It might be tricky or unfair, but we need to know how you're counting votes.

Clearly define the golden truth you're using. Again, sound and compelling arguments are needed, but – in my opinion – you should use the "real thing" (that is, avoid polls).

Naïveté isn't bliss. Sentiment analysis is a core task and simplistic sentiment analysis methods should be avoided one and all. Political discourse is plagued with humor, double entendres, and sarcasm; this makes determining users' political preferences hard and inferring voting intention even harder. We shouldn't rely on simplistic assumptions and should instead devote more resources to the special case of sentiment analysis in politics before trying to predict elections.

Credibility should be a major concern. A substantial amount of data in these cases isn't trustworthy and should thus be discarded. An incipient body of work exists in this regard,^{9,10} so you should at least apply the available methods to justify that the data you're using have been checked for credibility, and that disinformation, puppets, and spammers have been removed.

Acknowledge demographic bias and correct predictions accordingly. You can do this based on different demographic groups' participation in a prior election, and on what proportion of your twitter users belong to each of these groups. The second point is by far the hardest, but you should try your best to obtain demographic data and political preference for the users in your dataset.

Not every twitterer is tweeting about politics. A minority of users is responsible for most of the political chatter and, thus, those people's opinions will drive what you can predict from social media. This self-selection bias is still an open problem and should be another central part of future research.

In short, if you're planning to conduct serious research in this topic, please consider all the suggested recommendations. Above all, however, don't cherry-pick references to support your point because – remember – you can't (consistently) predict elections from Twitter! ☐

References

1. "Yahoo Search Tracks the Road to the White House," blog, 30 Oct. 2008; www.ysearchblog.com/2008/10/30/yahoo-search-tracks-the-road-to-the-white-house/.
2. d. boyd, "Big Data: Opportunities for Computational and Social Sciences," blog, 17 Apr. 2010; www.zephoria.org/thoughts/archives/2010/04/17/big-data-opportunities-for-computational-and-social-sciences.html.
3. D. Fanelli, "Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data," *PLoS ONE*, vol. 5, no. 4, 2010; www.plosone.org/article/info:doi/10.1371/journal.pone.0010271.
4. A. Tumasjan et al., "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," *Proc. 4th Int'l AAAI Conf. Weblogs and Social Media*, AAAI Press, 2010, pp. 178–185.
5. B. O'Connor et al., "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," *Proc. Int'l AAAI Conf. Weblogs and Social Media*, AAAI Press, 2010, pp. 122–129.
6. D. Gayo-Avello, "Don't Turn Social Media into Another 'Literary Digest' Poll," *Comm. ACM*, vol. 54, no. 10, 2011, pp. 121–128.

7. P.T. Metaxas, E. Mustafaraj, and D. Gayo-Avello, "How (Not) to Predict Elections," *Proc. IEEE Int'l Conf. Privacy, Security, Risk, and Trust (PASSAT) and Int'l Conf. Social Computing (SocialCom)*, 2011, pp. 165–171.
8. A. Jungherr, P. Jürgens, and H. Schoen, "Why the Pirate Party Won the German Election of 2009, or the Trouble with Predictions: A Response to 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 Rev.*, April 2011; <http://ssc.sagepub.com/content/early/2011/04/05/0894439311404119>. abstract.
9. C. Castillo, M. Mendoza, and B. Poblete, "Information Credibility on Twitter," *Proc. 20th Int'l Conf. World Wide Web (WWW 11)*, ACM, 2011, pp. 675–684.
10. M.R. Morris et al., "Tweeting Is Believing? Understanding Microblog Credibility Perceptions," *Proc. 15th ACM Conf. Computer Supported Cooperative Work and Social Computing (CSCW 13)*, ACM, 2012, pp. 441–450.

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](https://twitter.com/PFCdgayo).

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.