Spend a few minutes reading the Rojas editorial and Linkins' reply. Be sure to consider Figure 1 and Table 1 carefully, and address the questions below.

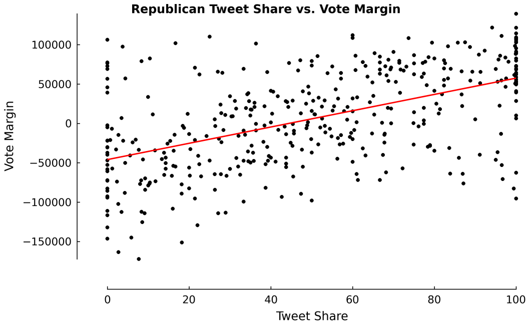

Figure 1: Bivariate relationship between the share of occurrences of Republican names in tweets and vote margin. We show a significant positive relationship at $P < .001$ with $R^2_{adj} = .283$.

Variable	Bivariate (SE)	Full Model (SE)
Republican Tweet Share	1035.0 (81.55) ***	154.7 (42.96) ***
Republican Incumbent		48932.53 (3014.15) ***
% McCain		2396.131 (131.38) ***
Median Age		-16.01 (406.56)
% White		439.82 (105.46) ***
% College Educated		-383.83 (207.91)
Median HH Income		79.77 (142.45)
% Female		-645.36 (1384.38)
CNN share		2.05 (36.77)
$Const$	-45832.6 (4853.35)	-116479.3 (69173.1)
N	406	406
R^2_{adj}	.28	.87

Table 1: Explaining Republican vote margin with the proportion of tweets that included a Republican candidate during the three months before the 2010 election. The share of Republican tweets that explain the relationship remains significant with $P < .001$ (***) after adding controls.

Statistics Hat

1. Write a sentence summarizing the findings of the paper.

2. Discuss Figure 1 with your neighbor. What is its purpose? What does it convey? Think critically about this data visualization. What would you do differently?

3. Interpret the coefficient of $RepublicanTweetShare$ in both models shown in Table 1. Be sure to include units.

4. Discuss with your neighbor the differences between the *Bivariate* model and the *Full Model*. Which one do you think does a better job of predicting the outcome of an election? Which one do you think best addresses the influence of tweets on an election?

5. Why do you suppose that the coefficient of *RepublicanTweetShare* is so much larger in the *Bivariate* model? How does this reflect on the influence of tweets in an election?

6. Do you think the study holds water? Why or why not? What are the shortcomings of this study?

Data Scientist Hat

Imagine that your boss, who does not have advanced technical skills or knowledge, asked you to reproduce the study you just read. Discuss the following with your neighbor.

1. What steps are necessary to reproduce this study? Be as specific as you can! Try to list the subtasks that you would have to perform.

2. What computational tools would you use for each task?

- working paper: `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2235423`

- published in *PLoS ONE*:`http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0079449` DiGrazia J, McKelvey K, Bollen J, Rojas F (2013) More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior. *PLoS ONE 8*(11): e79449.

- editorial in *The Washington Post* by Rojas: `http://www.washingtonpost.com/opinions/how-twitter-can-predict-an-election/2013/08/11/35ef885a-0108-11e3-96a8-d3b921c0924a_story.html`

- editorial in the Huffington Post by Linkins: `http://www.huffingtonpost.com/2013/08/14/twitter-predict-elections_n_3755326.html`

- editorial blog by Gelman: `http://andrewgelman.com/2013/04/24/the-tweets-votes-curve/`

Opinions

How Twitter can predict an election

Correction: *The op-ed originally gave the incorrect number for predicted elections. In the 2010 data, the analysis predicted the winner 92.8 percent of the time, or 404 out of 435 races when all are considered .The corrected version is below.*

By Fabio Rojas August 11, 2013

Fabio Rojas is an associate professor of sociology at Indiana University. He blogs at orgtheory.net.

Digital democracy is here. We no longer passively watch our leaders on television and register our opinions on Election Day. Modern politics happens when somebody comments on Twitter or links to a campaign through Facebook. In our hyper-networked world, anyone can say anything, and it can be read by millions.

This new world will undermine the polling industry. For nearly a century, conventional wisdom has argued that we can only truly know what the public thinks about an issue if we survey a random sample of adults. An entire industry is built on this view. Nearly every serious political campaign in the United States spends thousands, even millions, of dollars hiring campaign consultants who conduct these polls and interpret the results.

Digital democracy will put these campaign professionals out of work. New research in computer science, sociology and political science shows that data extracted from social media platforms yield accurate measurements of public opinion. It turns out that what people say on Twitter or Facebook is a very good indicator of how they will vote.

How good? In a paper to be presented Monday, co-authors Joseph DiGrazia, Karissa McKelvey, Johan Bollen and I show that Twitter discussions are an unusually good predictor of U.S. House elections. Using a massive archive of billions of randomly sampled tweets stored at Indiana University, we extracted 542,969 tweets that mention a Democratic or Republican candidate for Congress in 2010. For each congressional district, we computed the percentage of tweets that mentioned these candidates. We found a strong correlation between a candidate's "tweet share" and the final two-party vote share, especially when we account for a district's economic, racial and gender profile. In the 2010 data, our Twitter data predicted the winner in 404 out of 435 competitive races.

Why does this happen? We believe that Twitter and other social media reflect the underlying trend in a political race that goes beyond a district's fundamental geographic and demographic composition. If people must talk about you, even in negative ways, it is a signal that a candidate is on the verge of victory. The attention given to winners creates a situation in which all publicity is good publicity.

This finding is remarkable because it doesn't depend on exactly what people say or who says it. We measured only the total discussion and estimated each candidate's share. It is this relative level of discussion that matters for tracking public opinion in electoral contests. Furthermore, social media data mimic what polls measure. For example, in Ohio's 3rd Congressional District, we found that Republican Mike Turner got 65.4 percent of his district's tweet share. In the final election, he got 68.1 percent of the two-party vote. The tweet prediction was off by 2.7 percentage points — a figure that is within the margin of error of any poll.

This finding has profound implications for the democratic process. There are many nations that remain mired in poverty and do not have the infrastructure required for extensive polling. Furthermore, these nations often have governments that are suspicious of polling and try to suppress it. For these reasons, it is very hard to monitor elections. In contrast, as long as citizens have access to the Internet, they can talk about their views in a less-restricted manner. The "grassroots" buzz found in social media can be studied, and it will reveal how elections are conducted and if the state is respecting human rights. And as with U.S. elections, even if the people who use social media are not completely representative of the public, the amount of attention paid to an issue is an indicator of what is happening in society. Important events generate scrutiny that can be measured and studied.

Social media analysis is also important for elections in the United States. Polling favors the established candidates because it is relatively expensive. In contrast, social media analysis is cheap. Anyone with programming skills can write a program that will harvest tweets, sort them for content and analyze the results. This can be done with nothing more than a laptop computer.

Current polling practices also pay disproportionate attention to "big" races. Every four years, we have dozens of polls on the presidential election, but many other races for important offices will not be consistently polled. Some congressional races are never polled. Social media analysis can be used to systematically gather data on any race at any time. Thus, people in smaller states no longer need to rely on polling organizations for information. A single citizen can harvest social media data and learn about the election in his or her area.

Traditional polling will remain useful, especially for learning about voters' beliefs and backgrounds, but polls are no longer the only tool for forecasting elections. In the future, you will not need a polling organization to understand how your elected representative will fare at the ballot box. Instead, all you will need is an app on your phone.

Let's Calm Down About Twitter Being Able To Predict Elections, Guys

Posted: 08/14/2013 10:49 am EDT Updated: 11/11/2013 8:07 pm EST

Jason Linkins jason@huffingtonpost.com

Can Twitter help predict an election? Please, please, let the answer be "no." But Fabio Rojas, an associate professor of sociology at Indiana University, argues that it can in a recent Washington Post editorial. "Modern politics happens when somebody comments on Twitter or links to a campaign through Facebook," he writes, adding, "this new world will undermine the polling industry." Oh, well, it's been nice knowing you, polling industry!

The editorial reads more like, "Rah, Rah! [INSERT BUZZWORD HERE]" than anything resembling a piece of cogent political science. But Rojas and his coauthors lay out their case in a research paper, in which they describe how they painstakingly analyzed 542,969 tweets about Democratic or Republican candidates who ran in 2010. These were all sorted into specific races, and the percentage of tweets that mentioned each candidate was calculated. When this calculation, termed "tweet share," was matched up between opponents, the "tweet share" victor matched the winner in "404 out of 406 competitive races," Rojas writes. This was, he says, "a strong correlation."

Correlation does not imply ... what was it again?

In Rojas' mind, what he's stumbled upon is revolutionary because it's inexpensive, and polling is not. Furthermore, Rojas asserts that polling "favors the established candidates" and pays "disproportionate attention to 'big' races."

> Some congressional races are never polled. Social media analysis can be used to systematically gather data on any race at any time. Thus, people in smaller states no longer need to rely on polling organizations for information. A single citizen can harvest social media data and learn about the election in his or her area.

Terrific, I guess? I mean, as near as I can tell, a single citizen can access lots of polling data, too. Besides, one big reason that some congressional races are never polled is that some congressional races aren't much of a race.

Here's where I pass the mic to Stuart Rothenberg:

> Normally, when political scientists or journalists write about “competitive” races they are talking about contests where at least two candidates have at least some chance of victory. Obviously, there weren`t 406 “competitive” • House races in 2010 under that definition - at the Rothenberg Political Report, we rated just more than 100 House races as “not safe” • and a far fewer number in the truly competitive categories - so Rojas must be using the term to describe contested races.
>
> Most races aren`t real competitions, of course. Relatively few House challengers run robust campaigns, and voters generally are unfamiliar with challengers.
>
> Since House re-election rates have been over 90 percent in 19 of the past 23 elections, you don`t need polls or tweet counts to predict the overwhelming majority of race outcomes. In most cases, all you need to know is incumbency (or the district`s political bent) and the candidates` parties to predict who will win.

Rothenberg reckons that what "tweet share" can measure is name recognition, which is something that we tend to assert as fact without actually quantifying it in any way. (That said, I think that simple horse sense still usually wins out when evaluating name recognition.)

"But other than that," Rothenberg writes, "the idea that the content of tweets is irrelevant, and that it doesn`t matter if the tweets originate from inside a district or from people who cannot even vote in the race, seems to fly in the face of logic and everything that political scientists believe."

Oh, yeah, that's an important reminder: lots of people who write tweets about candidates are writing *negative* things about those candidates. Surely that makes raw "tweet share" completely useless as a measurement, right?

But Rojas says that it doesn't matter if the message is positive or negative.

> We believe that Twitter and other social media reflect the underlying trend in a political race that goes beyond a district`s fundamental geographic and demographic composition. If people must talk about you, even in negative ways, it is a signal that a candidate is on the verge of victory. The attention given to winners creates a situation in which all publicity is good publicity.

Well, then, congratulations to the next Mayor of New York City, Anthony Weiner!

[Would you like to follow me on Twitter? Because why not?]