More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior

Joseph DiGrazia,[1*] Karissa McKelvey,[2] Johan Bollen,[2] Fabio Rojas [1]

[1]Department of Sociology
[2]School of Informatics and Computing
Indiana University Bloomington, IN 47408, USA

[*]To whom correspondence should be addressed; E-mail: jdigrazi@indiana.edu.

Is social media a valid indicator of political behavior? We answer this question using a random sample of 537,231,508 tweets from August 1 to November 1, 2010 and data from 406 competitive U.S. congressional elections provided by the Federal Election Commission. Our results show that the percentage of Republican-candidate name mentions correlates with the Republican vote margin in the subsequent election. This finding persists even when controlling for incumbency, district partisanship, media coverage of the race, time, and demographic variables such as the district's racial and gender composition. With over 500 million active users in 2012, Twitter now represents a new frontier for the study of human behavior. This research provides a framework for incorporating this emerging medium into the computational social science toolkit.

An increasingly important question is whether social media activity can be used to assess offline political behavior. Online social networking environments present a tremendous scientific opportunity: they generate large scale data about the communication patterns and preferences of hundreds of millions of individuals (*1*), which can be analysed to form sophisticated models of individual and group behavior (*2, 3*). However, research also shows that social media content is largely focused on entertainment and emotional expressions (*4, 5*). Additionally, social media provides a self-selected sample of the electorate biased by population density, age, race, partisanship, income, and gender (*6–9*).

Despite these issues, there is a growing literature suggesting that online communication can still be a valid indicator of offline behavior. Searches for ethnic slurs correlate with lower vote tallies for minority politicians (*10*); film title mentions correlate with revenue (*11*); and online expressions of public mood correlate with fluctuations in stock market prices, sleep, work, and happiness (*12–14*). In addition, numerous studies have examined the relationship between online activity and election outcomes (*15, 16*). However, these studies do not account for confounding variables such as incumbency, partisanship, media coverage, and the sociodemographic makeup of the electorate (*17*).

Here we show a statistically significant relationship between tweets and electoral outcomes that persists after accounting for these potentially confounding variables. We compiled two large-scale datasets. First, we collected 2010 election outcomes and sociodemographic variables from all 435 U.S. House districts (*18*). Second, we retrieved a random sample of 537,231,508 tweets posted from August 1 and November 1, 2010. Then, we extracted 113,985 tweets that contained the name of the Republican or Democratic candidate for Congress.

For each candidate, we computed the number of tweets that contain their name (e.g., "Nancy Pelosi"). To account for the bias from a small number of extremely committed users or automated accounts generating tweets, we also computed the number of Twitter users who included

a candidate's name in at least one tweet. Each district i is assigned its share of Republican tweets, denoted $s_R(i)$, from the total of both Democratic and Republican frequencies, denoted $f_D(i)$ and $f_R(i)$ respectively. The $s_R(i)$ variable represents the amount of Twitter attention given to a particular candidate over their opposition in a particular race.

$$s_R(i) = \frac{f_R(i)}{f_D(i) + f_R(i)} \times 100 \tag{1}$$

Our dependent variable consists of the Republican vote margin for each district i, denoted $v_M(i)$ defined as the difference between the number of votes received by the Republican candidate, denoted $v_R(i)$, and the Democratic candidate, denoted $v_D(i)$, i.e. $v_R(i) - v_D(i)$. We did not use data from 29 districts where there was no opposition from a major party candidate. We normalize the dependent variable by district population (δ_i) to adjust for the small number of districts that deviate substantially from the average district size (μ_δ).

$$v_M(i) = (v_R(i) - v_D(i)) \times \frac{\mu_\delta}{\delta_i} \tag{2}$$

For example, Montana has an adult population of 772,072 represented by a single Congressional district, whereas μ_δ = 542,886.

We estimate the effect of Twitter share on vote margin by performing an Ordinary Least Squares regression (OLS). We include variables commonly used in other studies of congressional elections, such as incumbency and district partisanship (*19, 20*). Incumbency is coded as 1 if the Republican candidate is an incumbent and 0 otherwise. District partisanship is measured by the percentage of the 2008 Presidential vote share captured by John McCain.

In addition, we include controls capturing relevant aspects of the sociodemographics of each district such as median age, percent white, percent college educated, median household income and percent female (*21–23*). To control for the extent to which a candidate is covered in the traditional media, we have included a measure of how frequently the candidates were

mentioned in a transcript of a broadcast on the cable news network, CNN, during the same time period.

We estimate the effect of the tweet and user share variables with two models: a bivariate model and a full model including all controls. Across all models in Tables 1 and 2, the coefficients for both the tweet and user share show statistically significant effects ($P < .001$). Each additional percentage point of tweet share is associated with an increase in vote margin by 1,035 votes. The bivariate relationship between tweet share and vote margin is shown in Figure 1.

Although the effect size is reduced to 154.7 in the full model, the effect remains highly significant. Both models fit the data well; the R^2_{adj} for the bivariate model is .283 and increases to .871 in the full model. The effect for user share is 1,071 in the bivariate model and 173.7 in the full model, indicating that, net of all other factors, each additional percentage point increase in user share is associated with 173.7 extra normalized Republican votes. The effect of user share is also significant, indicating that this relationship is not driven by a small number of users.

To give a better sense of the magnitude of the effects, one standard deviation increase in the tweet share in the full model is associated with an increase in the vote margin equal to 4,978.9 votes. One standard deviation increase in user share is associated with an increase 5,622.67 votes. While these effects are much smaller than the effects for the Twitter measures in the bivariate model, modest increases in the tweet share measures still produce substantively meaningful and highly significant predicted changes in the vote share.

There are also a number of significant effects for some of the other control variables that are worth noting. Consistent with previous research, Republican incumbency and baseline district partisanship, as measured by McCain vote share, have highly significant effects in both models (*19, 20*). Interestingly, the percentage of whites also has a highly significant positive effect on vote margin in both models. This may indicate that voting in this election was particularly

racialized even compared to the 2008 contest.

We can assess the limitations of this model by looking at outliers. We examine those Congressional districts were the residual was at least two standard deviations above or below the predicted value. We find that districts where the model under-performs tend to be relatively uncompetitive. If there is little doubt about who the winner will be, there may be little reason to talk about the election. In the baseline model, for example, we obtain outliers such as California district 5 and West Virginia district 2. These areas lean heavily Democratic. California district 5 has voted Democrat since 1949. Since 2000, every Democrat polled at least 70%, with the exception of a 2005 special election, where the winning Democrat won with 67% of the vote. Similarly, West Virginia district 2 shows a strong partisan orientation. A single Republican has held the seat since 2001. However, a lack of competition does not explain every outlier. Some districts have idiosyncratic features that merit more research. For example, Oklahoma district 2 is a rural area that has voted for a Democratic Congressman while voting strongly for McCain and Bush.

Finally, we test the robustness of the results by examining the model across different time periods before the election, and across 2 different election cycles. First, because the link between tweeting and voter preferences may vary during the period before the election, we estimate the same models using only monthly shares of Twitter data from August, September, and October. As shown in Fig. 2, the effect of Twitter share is similar in terms of size and statistical significance across all three months as indicated by their overlapping confidence intervals. Second, a preliminary analysis of the 2012 U.S. House elections yields similar results. Data from 389 districts with competitive races yields a bi-variate OLS regression coefficient of 1,624 ($P < .001$).

The robust effect of tweet content on electoral outcomes is consistent with prior psycholinguistic research. A common finding is that people are more likely to say a word when it has a

positive connotation within the mind of the speaker (*24–26*). The over-representation of a word within a corpus of text may indicate that it signifies something that is viewed in a relatively positive manner. Directly testing this hypothesis is beyond the scope of this paper, but it does suggest a more general explanation for the consistently observed link between online discourse and behavior.

These findings have a number of important implications for the quantitative analysis of social media. First, the data do not include any information about the meaning or context of a name mention (e.g., "I love Nancy Pelosi" vs. "Nancy Pelosi should be impeached"). The relative share of attention compared to the opponent is all that is needed. This is evidence for the conventional wisdom that "all publicity is good publicity." Second, the models show that social media matters even when controlling for traditional television media, such as CNN, which many scholars have argued is important because it shapes political reality via agenda setting (*27, 28*), but does not seem to have a significant effect in our models. Finally, this study adds to the mounting evidence that online social networks are not ephemeral, spam-infested sources of information. Rather, social media may very well provide a valid indicator of the American electorate.

Acknowledgements We would like to thank Emily Winters and Matt Stephens for data collection as well as Clem Brooks, Elizabeth Pisares, and the Politics, Economy, and Culture Workshop at Indiana University for helpful discussions and contributions. We gratefully acknowledge support from the National Science Foundation (grants SBE 0914939, CCF 1101743), the Andrew W. Mellon Foundation, and the McDonnell Foundation.

References and Notes

1. W. Bainbridge, *Science* **317**, 4726 (2007).

2. D. Lazer, *et al.*, *Science* **323**, 7213 (2009).

3. A. Vespignani, *Science* **325**, 4258 (2009).

4. M. Naaman, J. Boase, C.-H. Lai, *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10 (ACM, New York, NY, USA, 2010), pp. 189–192.

5. A. Java, X. Song, T. Finin, B. Tseng, *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (ACM, 2007), pp. 56–65.

6. A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, J. N. Rosenquist, *ICWSM '11: 5th International AAAI Conference on Weblogs and Social Media* (Barcelona, Spain, 2011), pp. 554–557.

7. M. D. Conover, B. Gonc, A. Flammini, F. Menczer, *EPJ Data Science* **1**, 1 (2012).

8. E. Hargittai, *Journal of Computer-Mediated Communication* **13**, 276 (2007).

9. T. Correa, A. W. Hinsley, H. G. d. Ziga, *Computers in Human Behavior* **26** (2010).

10. S. Stephens-Davidowitz, *Quarterly Journal of Economics* (2011).

11. S. Asur, B. A. Huberman, *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10 (IEEE Computer Society, Washington, DC, USA, 2010), pp. 492–499.

12. J. Bollen, H. Mao, X. Zeng, *Journal of Computational Science* **2**, 1 (2011).

13. S. Golder, M. Macy, *Science* **333**, 187881 (2011).

14. P. Dodds, K. Harris, I. Kloumann, C. Bliss, C. Danforth, *PloS one* **6**, e26752 (2010).

15. A. Tumasjan, T. O. Sprenger, P. G. Sandner, I. M. Welpe, *Word Journal Of The International Linguistic Association* **280**, 178 (2010).

16. B. OConnor, R. Balasubramanyan, B. R. Routledge, N. A. Smith, *Proceedings of the International AAAI Conference on Weblogs and Social Media* (AAAI Press, 2010), vol. 5, p. 122129.

17. D. Gayo-avello, *Arxiv preprint arXiv12046441* pp. 1–13 (2012).

18. U. S. FEC, *Federal Elections 2010: Election Results for the U.S. Senate and the U.S. House of Representatives* (2010), pp. 39–150.

19. C. Klarner, *PS: Political Science & Politics* **41**, 723728 (2008).

20. A. I. Abramowitz, *The Western Political Quarterly* **28** (1975).

21. H. Brady, S. Verba, K. Schlozman, *American Political Science Review* pp. 271–294 (1995).

22. K. Schlozman, N. Burns, S. Verba, *Journal of Politics* **56**, 963 (1994).

23. S. Verba, K. Schlozman, H. Brady, N. Nie, *British Journal of Political Science* **23**, 453 (1993).

24. J. Boucher, C. E. Osgood, *Journal of Verbal Learning and Verbal Behavior* **8** (1969).

25. D. Garcia, A. Garas, F. Schweitzer, *EPJ Data Science* **1**, 1 (2012).

26. P. Rozin, L. Berman, E. Royzman, *Cognition & Emotion* **24** (2010).

27. M. E. McCombs, D. L. Shaw, *Public Opinion Quarterly* **36**, 176 (1972).

28. M. S. Roberts, *Journalism and Mass Communication Quarterly* **69**, 878 (1992).

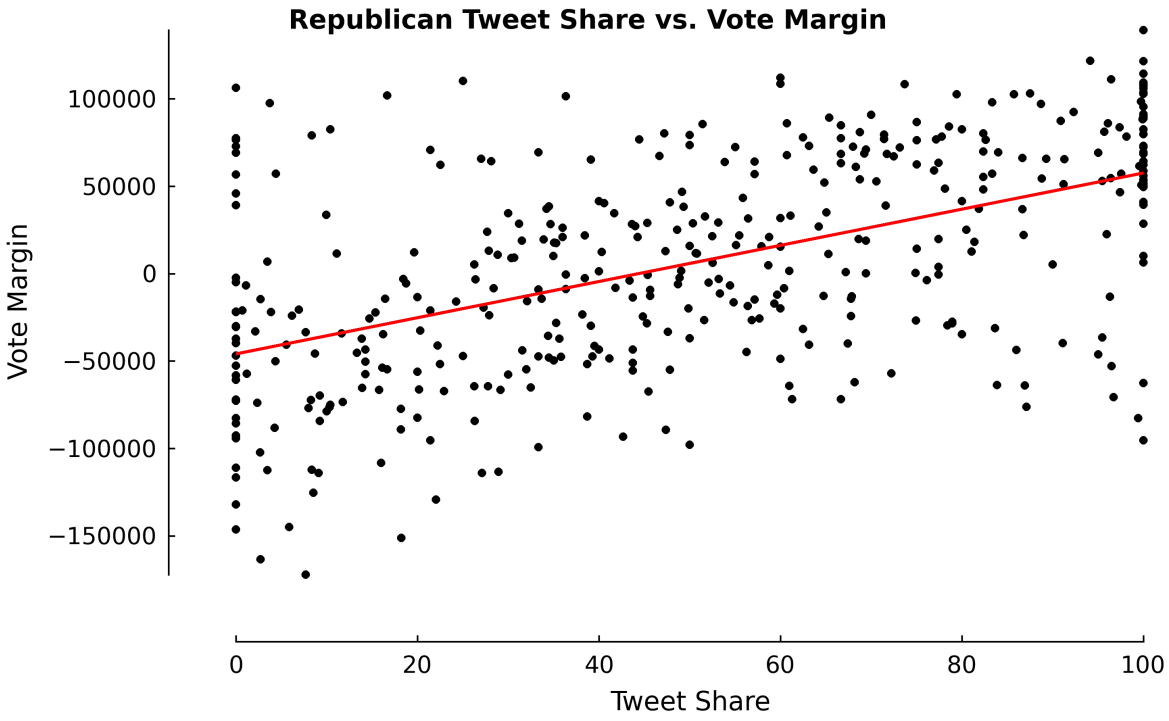

Figure 1: Bivariate relationship between the share of occurrences of Republican names in tweets and vote margin. We show a significant positive relationship at $P < .001$ with $R^2_{adj} = .283$.

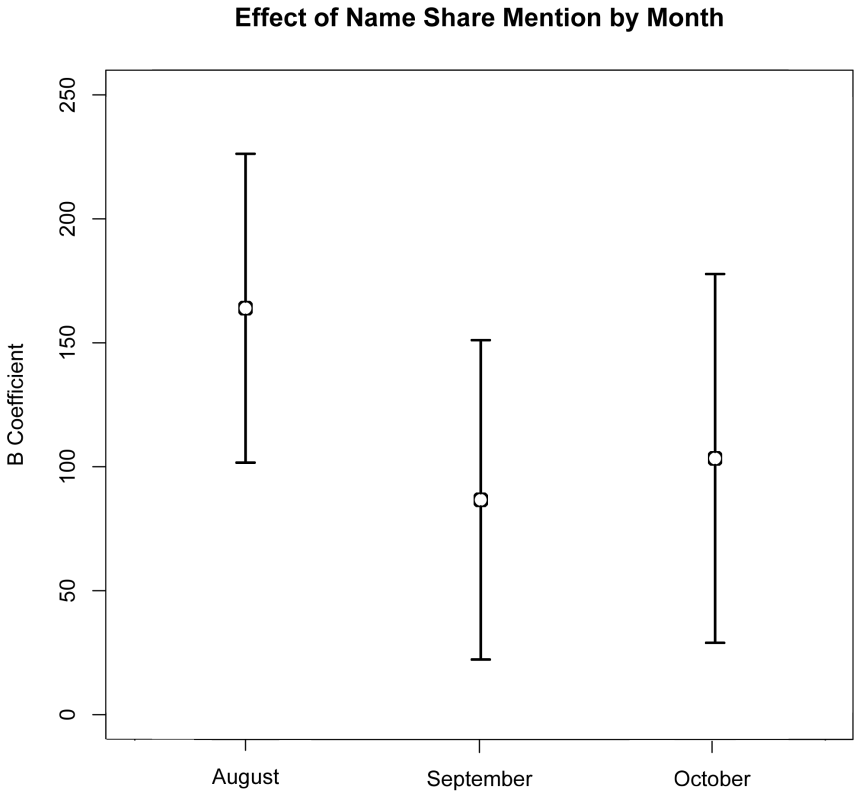

Figure 2: Effects of Republican tweet share during the months of August, September, and October with a 95% confidence interval.

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.

Variable	Bivariate (SE)	Full Model (SE)
Republican User Share	1071.0 (79.72) ***	173.65 (43.07) ***
Republican Incumbent		48563.34 (3001.07) ***
% McCain		2373.81 (131.39) ***
Median Age		-39.43 (404.72)
% White		447.06 (104.94) ***
% College Educated		-394.87 (206.98)
Median HH Income		83.31 (141.82)
% Female		-500.89 (181.072)
CNN share		-2.44 (36.63)
$Const$	-45832.6 (4769.36)	-123170.1 (68999.13)
N	406	406
R^2_{adj}	.307	.87

Table 2: Explaining Republican vote margin with the proportion of users who included a Republican candidate in at least one tweet. The relationship remains significant with $P < .001$ (***) after adding controls.