Who's Tweeting About the President? What Big Survey Data Can Tell Us About Digital Traces?

Pasek, Josh; McClain, Colleen A.; Newport, Frank; Marken, Stephanie

Article

Social Science Computer Review
2020, Vol. 38(5) 633-650

DOI: 10.1177/0894439318822007
journals.sagepub.com/home/ssc
SAGE

Who's Tweeting About the President? What Big Survey Data Can Tell Us About Digital Traces?

Josh Pasek[1], Colleen A. McClain[1], Frank Newport[2], and Stephanie Marken[2]

Abstract

Researchers hoping to make inferences about social phenomena using social media data need to answer two critical questions: What is it that a given social media metric tells us? And who does it tell us about? Drawing from prior work on these questions, we examine whether Twitter sentiment about Barack Obama tells us about Americans' attitudes toward the president, the attitudes of particular subsets of individuals, or something else entirely. Specifically, using large-scale survey data, this study assesses how patterns of approval among population subgroups compare to tweets about the president. The findings paint a complex picture of the utility of digital traces. Although attention to subgroups improves the extent to which survey and Twitter data can yield similar conclusions, the results also indicate that sentiment surrounding tweets about the president is no proxy for presidential approval. Instead, after adjusting for demographics, these two metrics tell similar macroscale, long-term stories about presidential approval but very different stories at a more granular level and over shorter time periods.

Keywords

Twitter sentiment, presidential approval, demographics, trends over time

This article is part of the SSCR special issue on "*Integrating Survey Data and Digital Trace Data*", guest edited by Sebastian Stier, Johannes Breuer, Pascal Siegers (GESIS – Leibniz Institute for the Social Sciences) & Kjerstin Thorson (Michigan State University).

Researchers are increasingly turning to digital trace data to generate novel insights about society, both *as a supplement to* and *a potential replacement for* surveys and experiments. These digital expressions are the culmination of a series of psychological and social processes that researchers cannot fully observe. Hence, scholars must make assumptions about what each social media metric

[1] University of Michigan, Ann Arbor, MI, USA
[2] Gallup, Inc., Washington, DC, USA

Corresponding Author:
Josh Pasek, University of Michigan, 105 S. State Street, 5413 North Quad, Ann Arbor, MI 48103, USA.
Email: jpasek@umich.edu

can tell us about society. Yet our assumptions about *what* tweets tell us and *who* they tell us about are difficult to straightforwardly test.

In work on this topic so far, scholars have expressed concerns about the self-selected nature of platform users. Individuals posting on social media sites represent a subset of a group that is unrepresentative of the public, even before considering the content and frequency of their posts, or whether tweets or accounts represent individuals at all. Consequently, researchers—especially those accustomed to probability-sample survey data—have worried that trace data describe nonprobability samples of the public (Couper, 2013; Murphy et al., 2014); some assume that if we could account for self-selection processes, then the correspondence between social media metrics and other social indicators would improve.

Yet, is an individual-level correspondence between survey and social media data even necessary for digital traces to predict survey trends, *if* we assume that "what" they tell us about is similar to survey data? Not all social media accounts lend themselves to this sort of analysis. Some do not represent humans (e.g., bots and corporate accounts); there is a healthy debate on our ability to detect these (Almaatouq et al., 2016; Chu, Gianvecchio, Wang, & Jajodia, n.d.; Guo & Chen, 2014; Vosoughi, Roy, & Aral, 2018). Even detectable, nonpersonal accounts may post information from news sources in ways that mirror attention among the larger public (Schober, Pasek, Guggenheim, Lampe, & Conrad, 2016). More broadly, Twitter posts are designed for an audience rather than solicited in response to a formulated question (Marwick & Boyd, 2011; Schober et al., 2016). Although different from the data generating process underlying survey data, they could potentially tap areas of general societal attention (Jungherr, Schoen, Posegga, & Jürgens, 2016; Pasek, Yan, Conrad, Newport, & Marken, 2018). Thus, individually or collectively, they may capture some form of public opinion.

In this study, we take as our starting point a strong assumed link in the literature that sentiment about the president on Twitter should track presidential approval (O'Connor, Balasubramanyan, Routledge, & Simon, 2010). Prior studies typically rely on the face validity correspondence between the metrics and the directional nature of both measures (Bae & Lee, 2012; Barberá, 2016; Cody, Reagan, Dodds, & Danforth, 2016; Marchetti-Bowick & Chambers, 2012); to the extent that this assumed link has been tested, it is done by attempting to match these variables to one another rather than to some external "ground truth."[1] We examine viability of such a link by assessing whose attitudes appear to be described by Twitter sentiment. To accomplish this, we first discuss a series of implications of *tweets-as-individuals'-opinions* and *tweets-as-societal-window* models for correspondence. We then articulate a new strategy leveraging rich survey data to determine whether tweets about Obama track approval levels among Americans generally, some subgroup of the public, or some combination of subgroups. Finally, we discuss what our results imply for producing survey-like inferences from Twitter data.

Tweets-As-Individuals'-Opinions

Thinking about Twitter posts as a form of public opinion makes intuitive sense. Individuals who use Twitter respond to the prompt "what's happening?", sometimes expressing their views on figures and events in the news.[2] Although different from the process that generates survey responses—the dominant contemporary form of public opinion—the notion of public expressions as opinions resonates with many early definitions of the concept (Anstead & O'Loughlin, 2015; Splichal, 2012).

If Twitter users represent self-selected subsets of individuals posting public opinions, aggregations of their posts constitute nonprobability samples of the public. Drawing on literature on nonprobability surveys (see Baker et al., 2013), some have attempted to correct for demographic imbalances in Twitter use and posting behavior to produce inferences about the public (see, e.g., Barberá, 2016; Diaz, Gamon, Hofman, Kıcıman, & Rothschild, 2016; Gayo-Avello, Metaxas, &

Mustafaraj, 2011; Sang & Bos, 2012). But these adjustments are fraught; users provide minimal biographic information and only some produce data that enable demographic inferences (see Cesare, Grant, & Nsoesie, 2017; Freelon, 2019; Zheng, Han, & Sun, 2017). Further, predicting and validating these variables can introduce error (Hsieh & Murphy, 2017; McClain, Mneimneh, Raghunathan, & Singh, 2018), and conclusions vary depending on whose tweets are analyzed (e.g., An & Weber, 2015; Cohen & Ruths, 2013).

Tweets-As-Societal-Window

Alternately, aggregations of tweets may tell us something about social trends but little about what it means for any given individual to post about a particular topic. Instead of collections of users' thoughts, they may serve as indicators of underlying social phenomena and thereby yield aggregate insights (Burnap, Gibson, Sloan, Southern, & Williams, 2016). For example, tweets could track news in ways that reveal dynamics of public attention (Jungherr et al., 2016). Given that Twitter facilitates news dissemination and information diffusion, users might be tweeting about topics relevant to the entire public sphere (cf. Pasek et al., 2018). Similarly, the broadcasting nature of tweets could lead users to imagine a general audience for posts and consequently cater the content they produce (Marwick & Boyd, 2011; Schober et al., 2016). In this model, the composition of users is not particularly relevant; aggregated tweets could provide a portrait of the public even if the user base was highly skewed.

Toward a Middle Ground

Most likely, elements of both models have some explanatory power. That is, tweets likely tell us something both about the particular individuals who are tweeting and about the larger information environment within which they are embedded. To understand these processes, we must distinguish the extent to which patterns of tweets reflect the attitudes and attention of Twitter users as opposed to larger scale social patterns. Yet making individual-level links between social media data and other forms of data is difficult, both because only a moderate proportion of respondents consent to data linkages (e.g., Barthel & Shearer, 2015; Vaccari et al., 2013) and because the probability that any given user will have posted about any given topic is small (Hobbs et al., 2017). If we can assess how groups of people vary in the extent to which their Twitter sentiment corresponds with expressed approval, however, we might unpack how these two models apply.

Leveraging Big Survey Data

In the current study, we employ a novel approach to determine what trends in the sentiment of tweets about Barack Obama tell us about presidential approval in the public at large and among notable population subgroups. This method flips the analytic problem on its head as compared to previous work—beginning from the survey data, with "known" demographics, and assessing the subgroups of survey respondents for whom approval ratings track a complementary Twitter metric. Rather than making strong assumptions about the demographics of Twitter users, this aggregate approach leverages information we already know, striking a middle ground between one-to-one linkages and aggregate trend analyses. We treat the sentiment of tweets mentioning "Obama" as a collection of individual assessments of presidential approval from an unknown demographic subset of individuals—making a key assumption of equivalence of sentiment to approval for the purposes of our investigation. Some individuals represented by survey responses (and those like them) are unlikely to tweet about the president; others will show up in the data to varying degrees and levels of correspondence. This model allows for *daily* and *overtime* correspondences between Twitter sentiment and public opinion that may be a product of different individual-level and societal-level forces, ranging from self-presentational concerns (microlevel) to event shocks in the data (macrolevel). Both perspectives are important to consider when assessing correspondence between sentiment and approval.

If groups are relevant, and if we have sufficient data to group users, we can identify when correspondence between surveys and social media differs by demographic. Notably, these differences could occur either because certain groups are unrepresented in the data or because the data generating process for survey expression versus social media expression differs by subgroup.

We are interested in four distinct questions. First, as in previous literature, how does Twitter sentiment track presidential approval? Second, does attention to differential behaviors across demographic groups improve the over time correspondence between Twitter sentiment and public opinion? Third, could the demographic groups whose approval levels best correspond with sentiment reasonably be the people whose tweets are responsive to changing attitudes about Obama during his presidency? And finally, what does the sensitivity of our inferences to robustness checks say about the processes underlying the relations we observe?

Data and Methods

Data

To assess whether Twitter sentiment reflected attitudes of identifiable subgroups of the population, we juxtaposed the positivity of tweets mentioning "Obama" with Gallup presidential approval survey data from January 2009 through September 2014. One hundred and twenty million tweets containing "Obama" were collected using Topsy, a firehose data collection vendor that was purchased by Apple in 2013 and is no longer available. Data were collected in batches of up to 40,000 posts matching the keyword per hour (few hours had more tweets). Data from the Gallup Daily survey, which surveyed a cross section of 500 respondents per day, were used to identify levels of presidential approval across various demographic subgroups. Both data streams were analyzed over 1,960 days shared across the data sets (see Online Appendix A for details on data collection and missing days).

Twitter sentiment. Twitter sentiment was calculated using the Lexicoder Sentiment Dictionary (Young & Soroka, 2012). For each day, all tweets posted were treated as a single text document. The corpus was run through the dictionary to identify words in four categories: positive, negative, negated positive, and negated negative. Daily sentiment was calculated: ((positive − negated positive) − (negative − negated negative))/total words. The corpus consisted of 2.9 billion words. Among these, ~70 million words were positive, ~89 million negative, with ~787,000 negated positives and ~473,000 negated negatives. Resulting daily scores ranged from −.034 to .083. Because it is unclear whether sentiment has an objective neutral point (i.e., positive and negative words may not be used in equal frequency) and our goal involved comparing Twitter sentiment with survey metrics, we generated a standardized measure of daily sentiment for comparisons.

Presidential approval. Respondents to the Gallup Daily survey were asked, "Do you approve or disapprove of the way Barack Obama is handling his job as president?" 286,630 respondents answered the approval question during the target time frame (excluding ~22,000 respondents who answered "don't know" or refused). ~154,000 approved (53.8% of valid responses) and ~133,000 disapproved overall (46.2% of valid responses). Daily approval levels were calculated as the proportion of individuals who approved among those who approved or disapproved; these ranged from

.35 to .85.[3] To mirror the estimation strategy used for sentiment and for consistency in modeling, daily approval means were standardized for comparisons.

Population subgroups. Eight variables from Gallup were used for one- and two-way demographic groupings. These included sex (two categories), age (seven), partisanship (five), education (six), race (five), census region (four), employment (five), and income (nine). The categories used can be found in Online Appendix A. In total, 44 primary demographic subgroups and 828 two-way combinations were used in the analyses.[4]

Analytical Strategy

To compare both overall approval and subgroup approval of Barack Obama with the sentiment in tweets about "Obama," we conduct a five-pronged analysis. First, we examine overall time trends observed from 2009 to 2014 in the overall approval and sentiment measures. Second, we generate both raw and smoothed estimates for each measure to compare approval and sentiment. Third, we calculate approval for primary and two-way combinations of demographic variables and compare these to sentiment to determine whether trends in approval among select groups of individuals more closely mirror sentiment. Fourth, we run a set of constrained regressions to determine whether combinations of primary or two-way demographic categories better account for correspondence between Twitter sentiment and approval. And finally, we conduct a series of analyses to determine whether the correspondences between data streams and the categories that drive them may exist by chance.

There are many reasons to think that social media metrics may be sensitive to short-term variations different from the factors that influence survey measures (Pasek et al., 2018). In particular, posts may vary in magnitude and valence as a function of events that produce large public reactions (Guggenheim & Pasek, 2013; Jang & Pasek, 2015; Kwak, Lee, Park, & Moon, 2010; Zhao et al., 2011). Daily means in survey responses, while also potentially impacted by events (Iyengar & Simon, 1993; Mutz & Soss, 1997), tend to be highly autocorrelated and are influenced by noise from variations due to sampling error (Kish, 1965). To this end, it is important to assess correspondence over different time windows and analytic strategies. For the current study, we conduct comparisons between daily means across data streams as well as data smoothed using generalized additive models (Hastie, 2017). For smoothed models, splines on the date variable were used to generate predicted values for sentiment and approval for each day. Each spline was given 44 knots, the square root of the total number of days examined, such that the number of knots matched the average number of days between knots. This smoothing closely mirrored overall trends without capturing daily instability. Additional details on this process are provided in Online Appendix B.

Understanding the results. We further assess the impact of potential methodological (e.g., overfitting and spurious correlations) and theoretical (e.g., daily variations differing from longer term trends) concerns. First, we examine whether our conclusions about correlations across data streams hold over shorter periods of time by rerunning the correlations over 1-year rolling windows instead of the full period. Subsequently, weaker correspondences would imply either that the results were spurious or that they were driven by large and rare shocks that influenced both data streams. Second, we examine how conclusions about the demographic groups for which approval reflects sentiment and the combination of predictive groups would vary depending on the specific dates chosen by bootstrapping the days used for the comparison. This assesses whether the demographic categories identified as driving trends are indeed stable. Third, we use detrended moving-average cross-correlation analysis (DMCA; Kristoufek, 2014), which extracts long-term trends from the data when analyzing correlations between data streams, to examine whether nonstationary processes might be inflating the correspondence between streams. We calculate moving averages over a period of 132 days.[5]

Results

Trends in Approval and Sentiment

Across all survey respondents, we replicated a long-standing finding that presidential approval typically wanes over time (Stimson, 1976). Figure 1A shows unweighted daily estimates of approval (dots) as well as a smoothed trend line. Barack Obama began his first term with the approval of more than 75% of Gallup respondents. By the end of his first year, his approval rates were equivocal and continuing to slide. Approval levels rebounded temporarily in early 2011 and improved during his reelection campaign throughout 2012. They fell again throughout 2013 and 2014.

Although more difficult to discern due to a few outlier days, sentiment in tweets about Obama also trended down over time (Figure 1B). Whereas average sentiment scores just after Obama was inaugurated hovered around .01 (corresponding to a distribution of *all words* that were 1 percentage point more positive than negative), by the end of the field period, sentiment was typically less than −.01. Notably, sentiment on select days differed considerably from the overarching trend. Sentiment spiked to .08 when Obama was awarded the Nobel Peace Prize (December 10, 2009) and fell to −.03 when hackers hijacked the Associated Press Twitter feed and reported that Obama had been injured in a series of explosions (April 23, 2013).

Correspondence Between Approval and Sentiment

At first glance, presidential approval and Twitter sentiment about Obama corresponded moderately over time, replicating prior work (O'Connor et al., 2010). When both metrics were smoothed, the correlation was .67; when daily estimates were used, the correlation was .44 (shown below). In subsequent analyses, we focus on the smoothed data and treat replications with daily data as a robustness check.

Correspondence between the sentiment of tweets and approval trends in particular demographic groups was typically similar to those of the aggregate trends, though there were notable variations. Pearson's correlations ranged from .50 to .81 when using smoothed data for primary demographic subgroups and from .03 to .86 for interacted subgroups. The strongest corresponding primary subgroup was those with a postgraduate education ($r = .81$), and the strongest interacted subgroup was those with a postgraduate education making US$36,000–US$47,000 per year ($r = .86$; see Online Appendix C).

Modeling Sentiment From Approval

It does not necessarily follow that the combination of the demographic categories that most strongly corresponded with Twitter sentiment would produce the best overall prediction of that sentiment. If multiple distinct groups of individuals were expressing feelings about President Obama on Twitter, a combination of these demographic groups could predict more effectively. Further, the demographic categories used to generate trends for bivariate comparisons were not mutually exclusive, meaning that some subgroups may have corresponded because individuals who posted heavily about their attitudes happened to fall into multiple groups.[6] By modeling sentiment as a combination of the attitudes of different groups of individuals, we project what kinds of people may have posted the messages observed or more particularly whose posting behavior may have varied over time in ways that corresponded with approval.

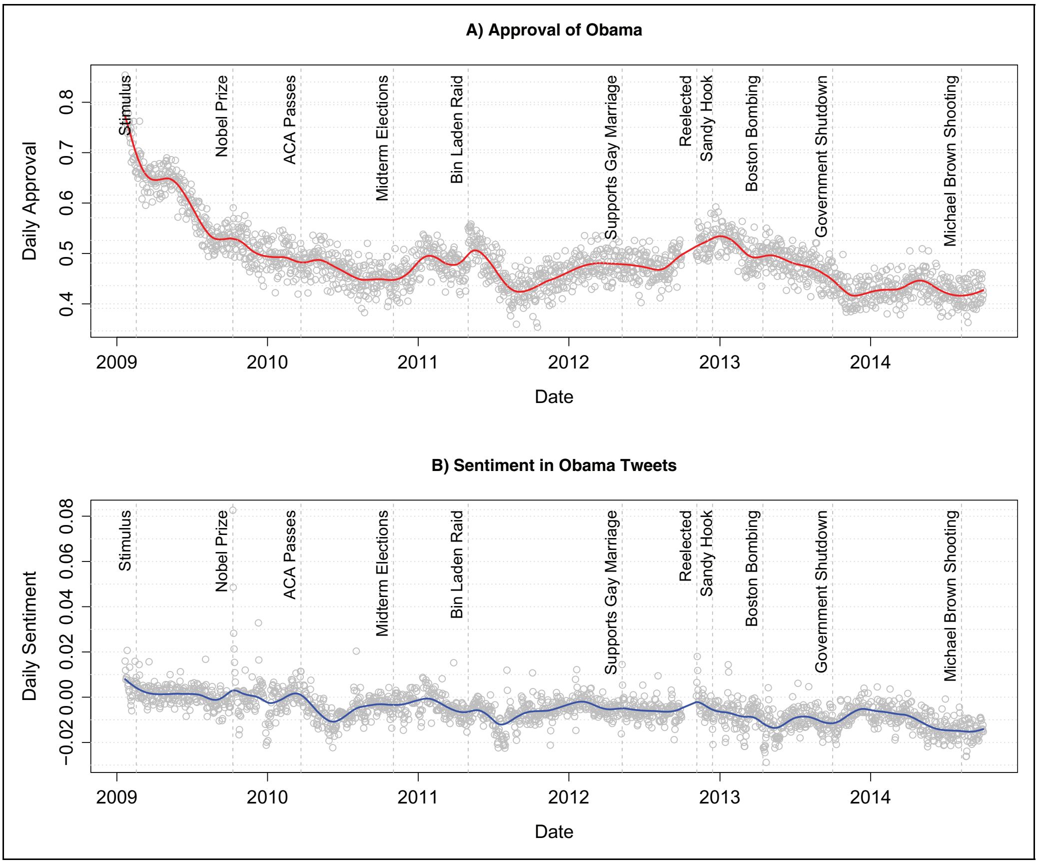

Figure 1. Presidential approval and sentiment in tweets about Obama over time.

To generate estimates of these combinations of demographic groups without overfitting the data, we decided a priori on a strategy of introducing increasingly restrictive constraints into the model until we derived a set of predictors that were both parsimonious and substantively interpretable.[7] As expected, putting all demographic predictors into a single ordinary least squares regression yielded results that were neither parsimonious (i.e., too many predictors entered the model and the fit was nearly perfect) nor substantively interpretable (i.e., a large proportion of the predictors were negative, yet it is impossible for groups to be negatively represented on Twitter). The simplest, most obvious, and most successful constraint was a requirement that the coefficients in the model be uniformly positive. To accomplish this, we employed Lawson–Hanson nonnegative least squares, as implemented in the *nnls* package for R (Mullen & van Stokkum, 2007) as our first tested constraint. This proved sufficient to yield parsimonious models, prevent overfitting, and eliminate spurious negative coefficients. Hence, no further constraints were tested.[8]

Using individual categories, 5 of the 44 subgroups entered the model as unique predictors of Twitter sentiment; the rest were constrained at 0. As shown in Table 1, these were trends for those with a postgraduate education ($b = .42$), Democratic leaners ($b = .16$), Blacks ($b = .15$), those with less than a high school education ($b = .13$), and other race individuals ($b = .03$). To assess the significance of these estimates and to see if alternative models might produce equally well-fitting results, we bootstrapped these regressions by randomly resampling dates with replacement. Across 10,000 resampled models, only one additional predictor emerged, Republican identification, and it

Table 1. Nonnegative Least Squares Predictions of Twitter Sentiment Trends Using Survey Trends With Primary Demographic Subgroups.

	Proportion of Resamples (%)	Average Coefficient	*N* of Respondents in Category	Proportion of all Respondents (%)
Postgraduate work or degree	100.0	.42	201,266	21.4
Lean Democrat	100.0	.16	125,015	13.3
Black	100.0	.15	79,009	8.4
Less than high school diploma	99.8	.13	62,185	6.6
Other	83.8	.03	25,753	2.7
Republican	2.4	.0003	285,140	30.3

did so in only 2% of resamples. Aside from other race, which appeared in 84% of resamples, the aforementioned demographic categories each appeared in at least 99% of resamples. Collectively, a model based on these demographic categories produced an estimate that correlated much more highly with smoothed Twitter sentiment than the overall approval measure (.81 vs. .67 for overall approval). This corresponds to about 65% of the variance explained (compared to 45% for overall approval).

Our estimates of sentiment improved further when approval trends from two-way demographic combinations were used as the predictors. Across 828 unique combinations of demographic categories, trends from between 9 and 17 groups uniquely predicted in any given bootstrapped model. These are shown in Table 2 (sorted by average coefficient). This model reveals that trends in Twitter sentiment best track a combination of the approval rates of Republicans lacking a high school diploma ($b = .23$), advanced degree holders making US$36,000–US$48,000 per year ($b = .21$), Black non-Hispanics earning US$24,000–US$36,000 per year ($b = .15$), and postgraduates earning US$$6,000–US$12,000 annually ($b = .12$). A trend estimate based on this combination correlated at .90 with sentiment, implying that these variables collectively explained 81% of the variance in smoothed sentiment.

Could these demographic groups plausibly be the ones producing variations in Obama-related Twitter content? The answer to this question is unclear. It does make some sense that the attitudes of highly educated individuals might contribute more to variations in sentiment than less educated individuals or that partisan leaners might be more sensitive to real-world cues than stronger partisans as the primary demographic model indicates. The results from crossed demographic categories are less clear-cut, though it is hard to contend that any group that shows up in these models is obviously wrong. Indeed, every demographic category associated with Twitter sentiment appears to use Twitter to a moderate degree in Pew studies (Gottfried & Shearer, 2016). This is an area that requires further investigation.

Robustness of Correspondences

While these overall results seem promising, one of the central concerns about correlations of smoothed data over long time periods is that correspondence may be driven either by overarching trends in the data or by seminal events that shock otherwise-divergent data streams in similar ways (Bode et al., 2019). A visual inspection of trends in Figure 1 provides some evidence for both processes. Sentiment and approval each experience downward shifts over the course of the Obama presidency. There are also notable improvements in sentiment and approval when Obama won the Nobel Prize, shortly after Osama Bin Laden was killed, and around his reelection.

Table 2. Nonnegative Least Squares Predictions of Twitter Sentiment Trends Using Survey Trends With Interacted Demographic Subgroups.

	Proportion of Resamples (%)	Average Coefficient	*N* of Respondents in Category	Proportion of all Respondents (%)
Republican × Less Than High School Diploma	100.0	.23	14,227	1.5
Postgraduate Work or Degree × US$36,000–US$47,999	100.0	.21	20,066	2.1
Black × US$24,000–US$35,999	100.0	.15	11,236	1.2
Postgraduate Work or Degree × US$6,000–US$11,999	100.0	.12	3,394	0.4
Female × Other Race	99.5	.09	11,798	1.3
Unemployed × $120,000 and Over	100.0	.08	6,577	0.7
Other Race × Employed Part Time, Do Not Want Full Time	99.7	.05	3,720	0.4
College Graduate × US$720–US$5,999	100.0	.04	2,378	0.3
Lean Democrat × Postgraduate Work or Degree	80.9	.03	29,091	3.1
Black × Under $720	86.0	.02	3,002	0.3
Age 35–44 × Black	83.7	.02	12,329	1.3
Lean Democrat × Employed Full Time, Self	38.4	.01	13,839	1.5
Age 65–74 × Other Race	62.6	.010	3,066	0.3
Black × US$$48,000–US$59,999	58.9	.009	8,461	0.9
Less Than High School Diploma × US$60,000–US$89,999	37.3	.005	3,845	0.4
Democrat × Other Race	23.4	.003	8,985	1.0
Other Race × US$$36,000–US$47,999	13.2	.001	3,148	0.3
Other Race × Northeast	9.2	.0006	3,776	0.4
Democrat × Postgraduate Work or Degree	4.5	.0005	72,697	7.7
Some College × US$720–US$5,999	0.9	.00005	4,334	0.5
Lean Democrat × Under US$720	0.2	.00001	2,972	0.3
Asian × Under US$720	0.3	.00001	500	0.1
Hispanic × Employed Part Time, Do Not Want Full Time	0.04	.000003	6,696	0.7
Independent, No Lean × Other Race	0.1	.000002	3,635	0.4
Postgraduate Work or Degree × Unemployed	0.01	.000001	14,337	1.5

Determining what to make of these correspondences is challenging on both theoretical and empirical levels. Overall trends and event-based responses are meaningful signals with respect to public opinions about the president. This phenomenon is different from that operating in time-series analyses in economics, for instance, where macrolevel trends often obfuscate microlevel relations and where macrolevel trends are typically not of theoretical relevance. But it is simultaneously important to know whether social media data provide substantive information about periods that are not heavily shaped by large-scale trends or responses to key events. Many analytical goals rely on an ability to use social media data to interpolate trends between surveys, for instance. To these ends, it is important that any given social media metric produces reliable correspondence over shorter time spans.

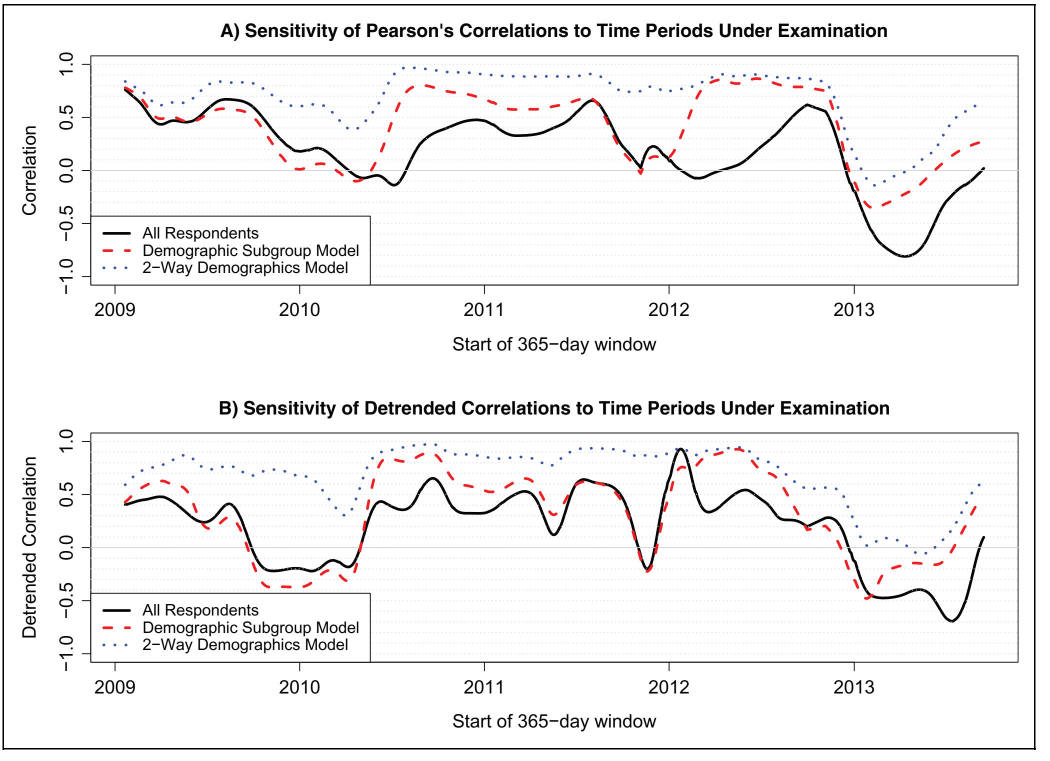

Figure 2. Sensitivity of correlations to time periods under examination.

Daily data. Daily correlations provide a window into the mechanisms by which social media data and survey data may correspond. To the extent that tweets reflect expressions of presidential approval or at least expressions among certain subgroups, we should observe correspondences between daily values that mirror those of the overall trends. To some extent this holds, but when we run the bootstrapping procedure on the daily data instead of the smoothed data, we find it much harder to identify a consistent set of groups that account for variations in sentiment (see Online Appendix D). Further, the overall correlation between the best fitting model and Twitter sentiment ($r = .59$) is far lower than for the smoothed version (.67). This implies that processes generating daily-level data are somewhat distinct from those underlying aggregate time trends.

Yearlong correspondence. We further examine whether correlations between data streams are consistent across shorter time spans. In a prior study, we found that the correspondence between economic confidence and tweets about "jobs" was strong over an 8-year period but highly variable when the window was only a single year (Pasek et al., 2018). Replicating this analysis, we compared sentiment and approval over all possible yearlong windows in the current study. As had been the case for economic confidence, we found lower average trends and considerable overtime variability in correspondence between approval and Obama sentiment (solid line in Figure 2A). Across all possible yearlong periods, the average correlation was .18, far lower than the .67 overall correlation between data streams. Further, 28% of the 1,595 spans yielded negative correlations instead of positive ones. Hence, while correspondence between overall approval and sentiment appeared highly sensitive to major shocks and trends over long time periods, this did not seem to apply to smaller windows.

Table 3. Pearson's (Above Diagonal) and DMCA (Below Diagonal) Correlations Between Smoothed Metrics.

	Smoothed Correlations				Daily Correlations			
	Twitter sentiment	Global approval	Subgroup approval model	Crossed subgroup approval model	Twitter sentiment	Global approval	Subgroup approval model	Crossed subgroup approval model
Twitter sentiment	—	.67	.82	.90	—	.44	.53	.59
Global approval	*.18*	—	.89	.87	*.15*	—	.89	.87
Subgroup model approval	*.32*	*.75*	—	.98	*.26*	*.83*	—	.95
Crossed subgroup model approval	*.63*	*.70*	*.87*	—	*.43*	*.82*	*.96*	—

Note. Pearson's correlations above the diagonal and DMCAs in italics below the diagonal. DMCA = detrended moving-average cross-correlation analysis.

Estimating the correspondence between approval among select demographic subgroups and sentiment across yearlong periods yielded stronger and more frequently positive correspondences, suggesting some promise for our approach but raising caution about interpreting why sentiment and approval might correspond only for very small subgroups. When using primary demographic groups, the yearlong correspondence averaged .40 and was positive for 85% of spans (dashed line in Figure 2A). Using interacted demographic groups, the average yearlong correspondence was .69, and 94% of yearlong periods were positively correlated with sentiment (dotted line in Figure 2A). It therefore seems that sentiment on Twitter not only tracks approval for these same subgroups better than among general population over the entire field period but also over shorter time periods as well. And this prediction might improve if we allowed the selection of subgroups to change dynamically over time. Although modeling the changing composition of Twitter posting over time is of theoretical and empirical importance for future work, it was beyond the scope of the current study.

Detrended analyses. When we replicated the analyses using detrending methods, some correlations disappeared whereas others remained strong. The overall correlations between all survey responses and tweets over the full time period essentially evaporated when detrending was used; these dropped from .67 to .18 (compare row 1 column 2 with row 2 column 1 in Table 3). This implies, similarly to our other robustness checks, that large-scale trends in both data streams—most notably, the overall decline in approval over time—were inducing much of the correspondence between streams. That said, once we restricted survey trends to the subset of respondents modeled to best reflect Twitter sentiment from the overall smoothed data, we found that correlations remained robust, though somewhat lower, across data streams. After detrending, the correlations dropped from .90 (row 1 column 4) to .63 (row 4 column 1). This indicates that the approval of subgroups identified in the nonnegative least squared models does capture trends that are similar to those in the Twitter data. The drops in correlations when detrending for daily estimates were far smaller for the demographic models than for overall approval and the resulting correlations were higher (correspondences for demographic subgroups are shown in Online Appendix C).

Whereas detrended overall correlations vacillated widely when comparisons were conducted over 1-year rolling spans, with an average DMCA of .19 and negative correlations for 30% of days (solid line in Figure 2B), the DMCA relations linking the two-way demographic prediction model with sentiment were positive for 97% of yearlong spans with an average correlation of .69 (dotted line in Figure 2B). These improved predictions were highly stable, even though the nonnegative least squares model used to generate them was not run using detrended data.

Discussion

Our results suggest that the initially strong correspondence between sentiment toward Obama and presidential approval from 2009 to 2014 was at least in part illusory. Although both data streams replicated the well-established finding that presidential popularity diminishes over time (Stimson, 1976), correlations between data streams became increasingly tenuous as we shifted the focus from macrolevel trends to assessing correspondence over smaller time periods or accounting for long-term processes. Overall, tweets about the president provided little to no information about how presidential approval was changing over the course of a year.

Nonetheless, there did appear to be particular subsets of respondents whose reported approval mapped onto trends in sentiment in far more nuanced ways. The fact that sentiment about Obama on Twitter mirrored the attitudes of a combination of less educated Republicans, lower income individuals with advanced degrees, lower income Blacks, women who were not White, Black, or Hispanic, and high-income unemployed individuals is substantively interesting, if not easily explicable. It would be valuable to assess how this compares to other methods of determining the demographic profiles of those tweeting about the president. It does fit, broadly, with the work of Hobbs and colleagues (2017), who report that a very small number of individuals are election poll "sensors" on Twitter. Further work should seek to understand how and why certain users or groups of people may serve as sensors, what the implications are of using certain subgroups for analysis, and what this means for understanding Twitter versus survey expression.

Why Doesn't Sentiment Consistently Track Approval?

Across the current results, we find that Twitter sentiment does not consistently track overall presidential approval aside from mirroring a steady decline over time. There are three basic reasons why this divergence might occur. First, it is possible that Twitter sentiment tells us about some notable subgroup(s) of the population, for which sometimes the trends diverge from the population at large. Second, Twitter sentiment might be a mixture of large-scale trends that reflect presidential approval and also smaller scale responses to other occurrences, such as events in the news. Finally, it could be the case that Twitter sentiment and survey-based approval are measures of distinct phenomena and *any* correspondence is spurious. Our data provide evidence for each of these possibilities.

Some demographic groups. The idea that only some people drive correspondence fits well with the finding that sentiment better tracks trends in approval among subsets of respondents than among the population as a whole. The groups we identify as most closely linked with Twitter sentiment make some conceptual sense. Twitter, at the time, was disproportionately used by individuals who were younger, lower income, non-White, or at the far ends of the education distribution (Smith & Brenner, 2012). It is of course possible that additional groups tweet about Obama but in ways that do not correspond with variations over time in the survey data. For example, if a group of strong Republicans posted negative content about Obama on Twitter in a way that never varied over time, they would not uniquely predict change over time in approval (and would therefore get dropped from our models).

Sensitivity. The daily estimates for Twitter sentiment provide some support for the idea that the measure is more sensitive to momentary events than the approval metric. Daily sentiment reacts to important speeches, terror plots, and the president's birthday (Online Appendix E). This is exactly what would be expected if the correspondence is largely overwhelmed by noisy processes that yield more microcosmic estimates. These sorts of processes have been proposed in other circumstances for why it may sometimes be appropriate to use nonprobability data streams to track trends over time (see Page & Shapiro, 1992; Pasek, 2016). An event-based view of social media data may help to reveal the extent to which these momentary spikes reflect long-term changes as opposed to irrelevant cues.

Sentiment and approval. Yet, beyond the sensitivity distinction, our work suggests good reasons to doubt the starting assumption in much of the literature of equivalence between sentiment and approval. Scholars have long known that the face validity of a correspondence is no guarantee of its accuracy (Cronbach & Meehl, 1955). And the notion that sentiment and approval measure the same construct relies on tenuous presumptions about the nature of both. For one, like all comparisons between Twitter-based metrics and survey results, sentiment differs from approval in that one is organically expressed and the other is solicited (Schober et al., 2016). Similarly, the wording of the approval question focuses on the president's job performance, whereas sentiment could capture all sorts of other constructs. Perhaps there is some other construct, instead of approval, that better captures what people are posting about on Twitter (cf. Jungherr et al., 2016).

The long-standing attempt to produce electoral predictions from Twitter highlights the challenge of understanding why prediction does or doesn't work. For one, it is unclear whether the volume of tweets about candidates and parties (e.g., Tumasjan, Sprenger, Sandner, & Welpe, 2011), the sentiment of those tweets (e.g., O'Connor et al., 2010), or some other metric is most appropriate for making predictions. The debate raises concerns about the extent to which substantive findings from social media studies may be spurious (Burnap et al., 2016), attributable to exploratory analyses and the so-called researcher degrees of freedom (see Beauchamp, 2016; Gayo-Avello, 2013; Simmons, Nelson, & Simonsohn, 2011). In our data, the correspondence in overall smoothed trends between Twitter sentiment and survey approval is a tenuous one. But why, exactly, it falls apart when we examine shorter time spans and consider daily data remains unclear.

Limitations and Next Steps

The evidence presented highlights several important factors that may underlie conditions under which social media data can yield insights similar to survey metrics. The specific methods we use, however, are limited. Although the sheer volume of both the Gallup and Twitter datasets are considerable, the models linking the two sources of data are based on numerous assumptions: That the correspondence between sentiment and approval is relevant, that the demographic categories used are sufficient for identifying who might be posting information, and that variations in posting behavior over time are similar across groups tweeting about the president. It is possible that most tweets about Obama are produced by extreme partisans on each side of the aisle. If these individuals produced positive or negative tweets at a consistent rate over time, their contributions to the Twitter stream would be orthogonal for our over time analyses; that is, they wouldn't vary and would not be picked up in a correlation coefficient of any sort. Thus, when we find that approval among a limited population tracks the changes in Twitter sentiment, we *cannot* and *should not* conclude that these are the only people tweeting about the president, even if we could be sure that those individuals were producing content. Testing these assumptions is beyond the scope of the current study. Future scholarship should therefore interrogate the content of Twitter posts among individuals that are also surveyed. A true test of the tweets-as-individuals'-opinions model requires this sort of one-to-one linked correspondence.

Further, although we conduct additional analyses to determine whether the attributes that account for correspondence do so in consistent ways over time, there is no true out-of-sample validation in this study (see Beauchamp, 2016). We therefore cannot be certain that our models are not overfitting the data. It would be valuable to replicate the results for other time periods and with other Twitter metric/survey response combinations to assess whether correspondence is consistently higher for the groups identified or whether these links were idiosyncratic. Similarly, it may well be the case that the conclusions we would reach from a similar comparison on a different social media site would differ.

Overall, the current study indicates that elements of Twitter sentiment track how presidential approval changes for certain subgroups of individuals but that there are a number of reasons that it is inappropriate to regard sentiment and approval as interchangeable. Our results suggest that Twitter sentiment blends a combination of individual-level approval assessments with other microlevel trends that are difficult to identify. If we hope to interpolate survey data with Twitter metrics, we need to continue working to discern the underlying data generating processes. The current analyses suggest that those processes may be multifaceted and more complex than they first appear.

Data Availability

The replication data and original analysis code for this article are available via ICPSR—http://doi.org/10.3886/E107205V1

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Software Information

All analyses were conducted using R 3.5.0 using the gam, nnls, and weights packages. Analysis code is available via ICPSR

Supplemental Material

Supplemental material for this article is available online.

Notes

1. It is not really clear how one could meaningfully ground truth a correlation of this sort, though see Barberá (2016).
2. This has been the prompt since November 2009 (see https://blog.twitter.com/official/en_us/a/2009/whats-happening.html).
3. Daily valid *N*s ranged from 364 to 962 ($m = 480$). Although the Gallup Daily survey is probability, results presented in the article do not use weights because there is little theoretical reason to believe that Twitter users will more closely mirror the national population than the population of Gallup respondents. These weights also would not have been appropriate for subgroup analyses.
4. The smallest primary subgroup category was 19,418 Asian, non-Hispanics; the median primary subgroup had 125,015 respondents. For interacted subgroups, the smallest category was 404 Asian respondents making US$720–US$6,000 per year. The median interacted subgroup had 18,462 respondents.

5. This was chosen because, at 3 times the average span between knots used for smoothing, it would ensure that the moving average would not capture the same variations as the smoothing but would still be sensitive to moderate-term trends.
6. The use of nonnegative least squares prevents this from being a multicollinearity issue in the models.
7. By interpretable, we mean that the set of predictors would not include closely related variables that were operating in opposing predictive directions. Although we were substantively interested in whether the predictors derived matched plausible theoretical expectations about the types of people who might use Twitter to post about Obama, we planned to generate models with consideration only on the number and overlap of predictors.
8. We had planned to use increasingly restrictive *glmnet* models as a next option.

References

Almaatouq, A., Shmueli, E., Nouh, M., Alabdulkareem, A., Singh, V. K., Alsaleh, M., ... Pentland, A. S. (2016). If it looks like a spammer and behaves like a spammer, it must be a spammer: Analysis and detection of microblogging spam accounts. *International Journal of Information Security*, *15*, 475–491.

An, J., & Weber, I. (2015). Whom should we sense in "social sensing"—Analyzing which users work best for social media now-casting. *European Physical Journal Data Science*, *4*, 22.

Anstead, N., & O'Loughlin, B. (2015). Social media analysis and public opinion: The 2010 UK General Election. *Journal of Computer-Mediated Communication*, *20*, 204–220. doi:10.1111/jcc4.12102

Bae, Y., & Lee, H. (2012). Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology*, *63*, 2521–2535.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., ... Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*, 90–143. Retrieved from http://www.aapor.org/AM/Template.cfm?Section=Reports1&Template=/CM/ContentDisplay.cfm&ContentID=5963

Barberá, P. (2016). Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data. doi:10.1101/095778

Barthel, M., & Shearer, E. (2015). *How do Americans use Twitter for news*. Washington, DC: Pew Research Center.

Beauchamp, N. (2016). Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science*, *61*, 490–503.

Bode, L., Budak, C., Ladd, J. M., Newport, F., Pasek, J., Singh, L. O., ... Traugott, M. W. (2019). *Words that matter: How the news and social media shaped the 2016 presidential election*. Washington, DC: Brookings Institution Press.

Burnap, P., Gibson, R., Sloan, L., Southern, R., & Williams, M. (2016). 140 Characters to victory? Using Twitter to predict the UK 2015 General Election. *Electoral Studies*, *41*, 230–233. doi:10.1016/j.electstud.2015.11.017

Cesare, N., Grant, C., & Nsoesie, E. O. (2017, February 6). Detection of user demographics on social media: A review of methods and recommendations for best practices. Retrieved from http://adsabs.harvard.edu/abs/2017arXiv170201807C

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (n.d.). Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*, 811–824. doi:10.1109/TDSC.2012.75

Cody, E. M., Reagan, A. J., Dodds, P. S., & Danforth, C. M. (2016, August 5). Public opinion polling with Twitter. Retrieved from https://arxiv.org/abs/1608.02024

Cohen, R., & Ruths, D. (2013). Classifying political orientation on Twitter: It's not easy!. *Proceedings of the Seventh Annual AAAI Conference on Weblogs and Social Media*, *7*, 91–99.

Couper, M. P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, *7*, 145–156.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. doi:10.1037/h0040957

Diaz, F., Gamon, M., Hofman, J. M., Kıcıman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PLoS One*, *11*, e0145406. doi:10.1371/journal.pone.0145406

Freelon, D. (2019). Inferring individual-level characteristics from digital trace data: Issues and recommendations. In N. J. Stroud & S. McGregor (Eds.), *Digital discussions: How big data informs political communication*. New York, NY: Routledge.

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review*, *31*, 649–679. doi:10.1177/0894439313493979

Gayo-Avello, D., Metaxas, P. T., & Mustafaraj, E. (2011). Limits of electoral predictions using Twitter. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 1–4.

Gottfried, J., & Shearer, E. (2016). *News use across social media platforms 2016*. Washington, DC: Pew Research Center.

Guggenheim, L., & Pasek, J. (2013). Binders full of tweets: Stimulus-response curves in Twitter reactions to news events. *11th Annual American Political Science Association Preconference on Political Communication*, Philadelphia, PA.

Guo, D., & Chen, C. (2014). Detecting non-personal and spam users on geo-tagged Twitter network. *Transactions in GIS*, *18*, 370–384. doi:10.1111/tgis.12101

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S* (pp. 249–307). New York, NY: Routledge. doi:10.1201/9780203738535-7

Hobbs, W., Friedland, L., Joseph, K., Tsur, O., Wojcik, S., & Lazer, D. (2017). "Voters of the year": 19 voters who were unintentional election poll sensors on Twitter. *Proceedings of the 11th International Conference on Web and Social Media (ICWSM)*, 544–547.

Hsieh, Y. P., & Murphy, J. (2017). Total Twitter error. In *Total survey error in practice* (Vol. *74*, pp. 23–46). Hoboken, NJ: John Wiley. doi:10.1002/9781119041702.ch2

Iyengar, S., & Simon, A. (1993). News coverage of the Gulf crisis and public opinion: A study of agenda-setting, priming, and framing. *Communication Research*, *20*, 365–383. doi:10.1177/009365093020003002

Jang, S. M., & Pasek, J. (2015). Assessing the carrying capacity of Twitter and online news. *Mass Communication and Society*, *18*, 577–598. doi:10.1080/15205436.2015.1035397

Jungherr, A., Schoen, H., Posegga, O., & Jürgens, P. (2016). Digital trace data in the study of public opinion. *Social Science Computer Review*, *35*, 336–356. doi:10.1177/0894439316631043

Kish, L. (1965). *Survey sampling*. New York, NY: John Wiley.

Kristoufek, L. (2014). Measuring correlations between non-stationary series with DCCA coefficient. *Physica A: Statistical Mechanics and Its Applications*, *402*, 291–298. doi:10.1016/j.physa.2014.01.058

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web*, 591–600.

Marchetti-Bowick, M., & Chambers, N. (2012). Learning for microblogs with distant supervision: Political forecasting with Twitter. *EACL '12 Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 603–612.

Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, *13*, 114–133. doi:10.1177/1461444810365313

McClain, C. A., Mneimneh, Z., Raghunathan, T. E., & Singh, L. O. (2018). *Ground truthing in social media research: Assessing methods used for demographic inference from Twitter*. Presented at the 73rd Annual Conference of the American Association for Public Opinion Research, Denver, CO.

Mullen, K. M., & van Stokkum, I. (2015). The nnls Package: The Lawson-Hanson algorithm for Non-negative Least Squares. Version 1.4. Retrieved from ftp://ftp.gr.xemacs.org/mirrors/CRAN/web/packages/nnls/nnls.pdf

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., … Harwood, P. (2014). Social media in public opinion research executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, *78*, 788–794. doi:10.1093/poq/nfu053

Mutz, D. C., & Soss, J. (1997). Reading public opinion: The influence of news coverage on perceptions of public sentiment. *Public Opinion Quarterly*, *61*, 431–451. doi:10.2307/2749580

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Simon, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *International AAAI Conference on Weblogs and Social Media*, *11*, 122–129. Retrieved from https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=84890614558&origin=inward

Page, B. I., & Shapiro, R. Y. (1992). *The rational public*. Chicago, IL: University of Chicago Press.

Pasek, J. (2016). When will nonprobability surveys mirror probability surveys? Considering types of inference and weighting strategies as criteria for correspondence. *International Journal of Public Opinion Research*, *28*, 269–291. doi:10.1093/ijpor/edv016

Pasek, J., Yan, H. Y., Conrad, F. G., Newport, F., & Marken, S. (2018). The stability of economic correlations over time identifying conditions under which survey tracking polls and Twitter sentiment yield similar conclusions. *Public Opinion Quarterly*. doi:10.1093/poq/nfy030

Sang, E. T. K., & Bos, J. (2012). Predicting the 2011 Dutch Senate election results with Twitter. *Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks*, Avignon, France, 53–60.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social media analyses for social measurement. *Public Opinion Quarterly*, *80*, 180–211. doi:10.1093/poq/nfv048

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Smith, A., & Brenner, J. (2012). *Twitter use 2012*. Washington, DC: Pew Research Center.

Splichal, S. (2012). Public opinion and opinion polling: Contradictions and controversies. In C. Holtz-Bacha & J. Strömbäck (Eds.), *Opinion polls and the media* (Vol. *1*, pp. 25–46). London, England: Palgrave Macmillan. doi:10.1057/9780230374959_2

Stimson, J. A. (1976). Public support for American presidents: A cyclical model. *Public Opinion Quarterly*, *40*, 1–21. doi:10.1086/268264

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, *29*, 402–418. doi:10.1177/0894439310386557

Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J. T., Nagler, J., & Tucker, J. (2013). Social media and political communication: A survey of Twitter users during the 2013 Italian general election. *Rivista Italiana Di Scienza Politica*, *XLIII*, 381–410. doi:10.1426/75245

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*, 1146–1151.

Young, L., & Soroka, S. N. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*, 205–231. doi:10.1080/10584609.2012.671234

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E. P., Yan, H., & Li, X. (2011). Comparing Twitter and traditional media using topic models. In *Advances in Information Retrieval* (Vol. *6611*, pp. 338–349). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-20161-5_34

Zheng, X., Han, J., & Sun, A. (2017). A survey of location prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, *30*, 1652–1671.

Author Biographies

Josh Pasek is associate professor of communication studies, faculty associate at the Center for Political Studies, Institute for Social Research, and core faculty in the Michigan Institute for Data Science at the University of Michigan, Ann Arbor, MI, USA.

Colleen A. McClain is a PhD candidate in the Program in Survey Methodology at the University of Michigan, Ann Arbor, MI, USA.

Frank Newport is a Gallup editor in chief, Gallup, Inc., Washington, DC, USA.

Stephanie Marken is a chief methodologist of the Gallup Poll at Gallup, Inc., Washington, DC, USA.