Oxford Handbooks Online

Measuring Public Opinion with Social Media Data

Marko Klašnja, Pablo Barberá, Nick Beauchamp, Jonathan Nagler, and Joshua Tucker

The Oxford Handbook of Polling and Polling Methods
Edited by Lonna Rae Atkeson and R. Michael Alvarez

Subject: Political Science, Political Methodology, Political Behavior
Online Publication Date: Oct 2017 DOI: 10.1093/oxfordhb/9780190213299.013.3

Abstract and Keywords

This chapter examines the use of social networking sites such as Twitter in measuring public opinion. It first considers the opportunities and challenges that are involved in conducting public opinion surveys using social media data. Three challenges are discussed: identifying political opinion, representativeness of social media users, and aggregating from individual responses to public opinion. The chapter outlines some of the strategies for overcoming these challenges and proceeds by highlighting some of the novel uses for social media that have fewer direct analogs in traditional survey work. Finally, it suggests new directions for a research agenda in using social media for public opinion work.

Keywords: social networking sites, Twitter, public opinion, public opinion surveys, social media, political opinion, representativeness

Social Media and Public Opinion: Opportunities and Challenges

Social media sites such as Facebook and Twitter are playing an increasingly central role in politics. As Kreiss (2014) shows, the 2012 Barack Obama and Mitt Romney presidential election campaigns relied heavily on social media to appeal to their supporters and influence the agendas and frames of citizens and journalists. In 2016 the role of social media accelerated, with Twitter, for example, becoming a central pillar of the Trump campaign. Social media sites have also been essential for disseminating information and organizing during many recent episodes of mass protest, from the pro-democracy revolutions during the Arab Spring to Euromaidan to the recent wave of pro–civil rights demonstrations in the United States (see, e.g., Tufekci and Wilson 2012; Tucker et al. 2016). The influence of social media has also become pervasive in traditional news outlets. Twitter is commonly used as a source of information about breaking news events, journalists and traditional media often solicit feedback from their viewers through social media, and political actors can rely on social media rather than press releases to reach the public. Most fundamentally, for numerous political organizations and millions of users, social media have become the primary means of acquiring, sharing, and discussing political information (Kwak et al., 2010; Neuman et al., 2014).

This chapter examines to what extent one can aggregate political messages published on social networking sites to obtain a measure of public opinion that is comparable or better than those obtained through surveys. It is well known that public opinion surveys are facing growing difficulties in reaching and persuading reluctant respondents (De Leeuw and De Heer 2002). According to the Pew Research Center, the typical contact rates dropped from 90% to 62% between 1997 and 2012, with response rates dropping from about 40% to 9% (Pew Research Center 2012).[1] One important reason for these trends is the falling rate of landline phone use, coupled with the fact that federal regulations prohibit the use of automated dialers for all unsolicited calls to cell phones (but not landline phones). According to one estimate, the share of cell-phone-only households in the United States has grown by 70% in four years, reaching 44% of all households in 2014.[2] While the relationship between nonresponse rates and nonresponse bias—which arises when those who answer are different from those who do not—is complex (Groves 2006; Groves and Peytcheva 2008), survey responders tend to be more likely to vote, contact a public official, or volunteer than are survey nonresponders (e.g., Pew Research Center 2012). The responders' answers tend to exhibit less measurement error and lower social desirability bias (Abraham, Helms, and Presser 2009; Tourangeau, Groves, and Redline 2010). The cell-phone-only respondents can differ in political preferences than those with landline phones; for example, they were significantly more likely to support Obama in 2008, especially older voters (Mokrzycki, Keeter, and Kennedy, 2009). These trends have raised questions about the reliability and precision of representative surveys and have increased the costs of fielding high-quality polls, at the same time that funding available for a number of established large-scale surveys has been threatened.[3]

These factors are increasing the incentives for using social media to measure public opinion. First and foremost, social media provide an opportunity to examine the opinions of the public without any prompting or framing effects from analysts. Rather than measure what someone thinks about politics in the artificial environments of a front porch, dinnertime phone call, or survey web page, we can observe how people spontaneously speak about politics in the course of their daily lives. And instead of depending on the analyst's view of which topics are important at any given time, we can observe the topics that the public chooses to raise without our prompting.

The second major appeal of social media data is their reach: over time, across individuals, cross-nationally, and within small geographical regions. Due to the fine-grained nature of Twitter and Facebook data, for example, it should be possible to measure changes in opinion on a daily or even hourly basis. Similarly, because hundreds of millions of people use Twitter and Facebook regularly, the scope of opinion that can be measured goes far beyond anything we could previously have attempted. And since social media can be found throughout the world, they provide a convenient platform for sampling opinion in many countries where it would otherwise be difficult or impossible for survey researchers to work. In fact, it is likely that the Twitter archive is already the largest cross-national time-series data set of individual public opinion available to the mass public.[4]

The third appeal of using social media to measure public opinion is the cost and practicality. With a little programming and a decent-sized hard drive, anyone can capture, for example, every comment made about a presidential debate, in real time and for free. To the extent that we care about public opinion because we think it helps to hold rulers more accountable and to make policy more responsive to the mass citizenry, the potential to dramatically reduce the cost of studying public opinion may be perhaps the most exciting opportunity afforded by social media.[5]

Of course while social media have desirable properties that traditional public opinion surveys cannot match, truly developing tools to effectively harness their potential involves enormous challenges, discussed in the next section. Each of the strengths discussed above also constitutes a challenge—both theoretical and technical—for measuring opinion in the ways we are used to using traditional surveys. First, identifying emergent topics and sentiments is hugely challenging, not just computationally but theoretically, as we strive to understand machine- or human-generated summaries and reconcile them with previous survey measures and research agendas. Second, the breadth and scale of social media use is counterbalanced by the opacity of its user population, and the steps needed to reweigh this entirely unrepresentative “survey” in order to measure any population of interest remain difficult and uncertain. Third, the technical challenges of collecting and aggregating the data are nontrivial, particularly given the diffident and often opaque cooperation of private social media providers like Twitter and Facebook.

We believe that many of these challenges involved in using social media data to study public opinion can be overcome, and that the potential payoff certainly justifies the effort. But we also believe it is crucial to be upfront about these challenges moving forward, and therefore one of the goals of this chapter is to lay these out explicitly. In the second section we discuss in greater detail these three main challenges and how they have arisen in past social media research. In the third section we discuss some of the strategies for overcoming many of these challenges, both drawing upon past work that suggests various successful strategies and suggesting new ones. And in the fourth section we discuss in greater detail some of the novel uses for social media, ones that have fewer direct analogs in traditional survey work. We conclude in the fifth section with a series of recommendations for a research agenda that uses social media for public opinion work, as well as providing a list describing *how* social media data were collected that we suggest scholars and practitioners use when reporting any results based on social media data, but especially when reporting results claiming to be representative of public opinion.

We focus here on Twitter data because they are widely used, mainly public, and relatively easy to collect; for these reasons, these data have been the focus of the majority of recent social media research. But of course all of these concepts apply more generally, and the difficulties and solutions we propose here will likely continue well into a future in which social media platforms that do not exist yet may dominate the landscape.

Challenges in the Measurement of Public Opinion with Social Media Data

In the study of public opinion, a *survey* is commonly defined as a systematic method for gathering information from a sample of individuals for the purposes of constructing quantitative descriptors of the attributes of the larger population of which the individuals are members (see, e.g., Groves et al. 2011). This information is commonly gathered by asking people questions. The three core components of a survey are thus a standardized questionnaire, a population frame from which individuals are sampled using a probability sampling method, and a method to aggregate individual responses to estimate a quantity of interest.

Without any adjustment, treating social media data as a survey fails to meet any of these three criteria: the opinions expressed by individuals are unprompted and unstructured, the probability that an individual is included in the sample varies in systematic but opaque ways, and the collection and aggregation of data into quantities of interest are problematic due to uncertainties in the data-generating and -collection processes. In this section we describe these difficulties and why they are critical for the measurement of public opinion with social media data.

When we speak of trying to measure "public opinion" we are primarily concerned with the traditional notion of who "the public" is: adults in a particular polity (or set of polities). However, one of the benefits of social media is that there is no such constraint on whose opinion is uttered on social media. We are potentially able to measure subpopulations of interest within a polity, such as ethnic groups, ideological groups, or speakers of particular languages (Metzger et al. 2016). On the other hand, this also extends to populations such as children, political activists, persecuted minorities, and other subpopulations that want, expect, or deserve privacy in their online activities. Fully tackling the myriad ethical issues entailed in using social media to measure public opinion would require an entire chapter, but we should be aware that such issues permeate every stage discussed below. Some of these issues are common to any sort of collection of publicly available data, including the issue of consent regarding data that have been made public but may be used in ways not anticipated by the participant, and the collection of data from minors and others not able to give consent themselves. Other issues are common to data collection and storage more generally, including data protection and anonymization, and specific to sharing data, particularly for replication purposes. Other questions are more specific to social media data, including how to deal with posts that were deleted by users after the data were collected, the potential privacy violations inherent in using sophisticated machine-learning methods to infer demographic and other characteristics that had not been publicly revealed, and the question of whether the results of these analyses could put exposed users at risk of political or other forms of retaliation (Flicker Haans and Skinner 2004; Tuunainen, Pitkänen, and Hovi 2009; Zimmer 2010; Solberg 2010; Bruns et al. 2014). The scope of ethical considerations in social media studies is rapidly growing, but for our purposes we focus here on the technical challenges of measuring public opinion.

Identifying Political Opinion

If we seek to fit social media into the framework of existing public opinion measurement, we may consider social media posts as something like unstructured and entirely voluntary responses to external stimuli analogous to public opinion questions. In this sense, like a survey question, the stimuli set (or affect) the topics, and our job is to identify these topics and turn the unstructured responses into something like sentiment, approval levels, feeling thermometers, or the like. In both traditional surveys and unstructured social media, we have something like a subject (the question, or a post’s topic) and a predicate (the numeric response, or a post’s sentiment), and we seek to turn the raw data of the unstructured social media text and metadata into something more like the structured survey responses we are familiar with. This analogy is often latent in research using social media to measure public opinion, but making it more explicit clarifies a number of issues in putting social media to such a use. This distinction has also been referred to as the distinction between “designed data” and “organic data” (Groves 2011). Whereas traditional collections of public opinion data, or data on economic behavior based on survey responses, are curated and created by the designer with intent in mind, many data sets now available are based on data that exist simply because much human behavior occurs online—and is recorded.

First, the questions are not directly asked of people; instead, people give their opinions in response to events and discussions. How do we define what topics to examine and which tweets are relevant for a given topic? For example, if we want to measure users’ sentiment toward the candidates in the 2016 presidential election, how do we identify a corpus of relevant tweets? The vast majority of studies focus on tweets mentioning candidate names, without discussing the possibility of systematic selection bias in determining the search criteria in this way (but see King, Lam, and Roberts 2014). For example, focusing only on tweets that mention Hillary Clinton or Donald Trump may miss a number of social media messages that also relate to the 2016 election but do not mention candidate names (He and Rothschild 2014). If tweets that refer to either candidate *without* using the candidate’s name tend to be either more positive or negative than tweets that do explicitly mention the candidate’s name, then obviously selecting tweets based on the use of the name will generate a large amount of selection bias. And that’s just bias relative to the full corpus of tweets on the candidates, even apart from bias relative to the population of interest; perhaps only persons with some particular characteristic use particular terms. If that is the case, and we omit terms used by that group, we will fail to measure group opinion accurately. Even without generating bias by collecting based on candidate names, collections based on names may include substantial noise or miss substantial numbers of tweets. Tweets containing “Hillary” in 2016 may be predominantly about Hillary Clinton, but tweets containing “Trump” or “Cruz” may not be about either Donald Trump or Ted Cruz, thus adding noise to the corpus. Filtering on tweets containing “Donald Trump” or “Ted Cruz” may miss many tweets actually focused on the candidates. In general, the choice of the relevant corpus of tweets is almost invariably ad hoc, in part because the analyst cannot be omniscient about what constitutes the set of tweets related to a given topic.

In addition to defining the topics and content that shape the collected data, measuring the topics in individual tweets, particularly when topics may be rapidly changing over time or responding to major events, remains both a technical and theoretical challenge. Are people responding to changing questions on the same topic, are the topics themselves changing, or do we need a complex hierarchical and temporal structure of all our content before we can begin to quantify public opinion in a systematic way? For example, during the presidential debates there was considerably more commentary and discussion among users than at other times, when information-sharing tweets (with high frequency of URLs within tweets) were more common (Diaz et al. 2014). Similarly, during politically charged events, such as the Wisconsin labor strikes of 2011, many social media users seem to have been particularly focused on tweeting non-mainstream news and alternative narratives of the protest, unlike during less contentious events (Veenstra et al. 2014). The same occurs during mass protest events, since regime elites can respond strategically to protest and try to shift the focus of the discussion (Munger 2015; King, Pan, and Roberts 2016). Topics themselves can change, because the comments on social media may represent a change in public opinion: either the development of a new issue that was previously not part of political discourse or the disappearance of an issue from public concern. The set of topics dominating political discussion in 2000 would be very different than the set of topics dominating political discussion in 2016. And just as refusal to answer surveys may not be random, but may vary systematically with the likely response, discussion on any issue on social media may vary with context. During a period of “good news” for a candidate, we may see more tweets by the candidate’s supporters, and vice versa. Thus the population, topics, and sentiments may all be continually shifting in ways that are very challenging to measure.

Even assuming we are able to resolve the subject—the topics—what of the predicate: the sentiment, approval, enthusiasm, and so forth? What exactly is the quantity of interest? Simple counts of mentions of political parties or issues have in some cases produced meaningful results. For example, Tumasjan et al. (2010) and Skoric et al. (2012) showed that mentions of parties on Twitter were correlated with election results. However, that is often not the case (Metaxas, Mustafaraj, and Gayo-Avello 2011; Bermingham and Smeaton 2011). In fact, Gayo-Avello (2011) showed that tweet-counting methods perform worse than a random classifier assigning vote intentions based on the proportion of votes from a subset of users who directly revealed their election-day intentions to the researcher. Similarly, Jungherr, Jürgens, and Schoen (2012) criticize the tweet-counting method used by Tumasjan et al. (2010) to predict German elections for focusing only on tweets mentioning the largest parties. They show that if tweets mentioning a new party—the Pirate Party—were counted as well, the results differed considerably and mispredicted the election outcome, as the Pirate Party was a clearly predicted election winner, whereas in fact it won only 2% of the vote (see also Jungherr et al. 2016).

One common alternative to counting methods is the use of sentiment analysis, which aims at measuring not the volume of tweets on a particular topic, but the valence of their content. This method often relies on existing dictionaries of positive and negative words, in which the ratio of positive to negative words that co-occur with a topic on, for example, a given day, is taken as a measure of the overall public sentiment about that topic on that day. For example, O'Connor et al. (2010) show that Twitter sentiment over time in economics-related tweets is correlated with consumer confidence measures in the United States. The downside of this approach is that its performance can vary in unpredictable ways. The approach depends on potentially ad-hoc dictionaries and often exhibits low out-of-sample accuracy (González-Bailón and Paltoglou 2015) and even significant differences in its performance across different applications within a similar context. For example, Gayo-Avello (2011) finds that the performance of a lexicon-based classifier was considerably more reliable for tweets about Barack Obama than about John McCain during the 2008 election campaign.

Finally, even if we have a good method for measuring topics and, for example, sentiments, it is not at all clear that what we are measuring is necessarily an honest expression of opinion. It remains unknown to what degree the (semi-)public nature of social media could induce stronger social desirability bias than in the context of traditional survey responses. On the one hand, given potential social stigma, users may be even less likely to reveal attitudes on sensitive topics than they are in standard surveys (Newman et al. 2011; Pavalanathan and De Choudhury 2015), and individuals can control their content after it is posted, with changes and edits potentially inducing selection bias in the type of content that remains (Marwick and Boyd 2011). On the other hand, Twitter in particular does allow users a certain degree of anonymity (though perhaps less than they think) and thus may allow individuals to express their true preferences and attitudes more honestly than in many traditional surveys (Joinson 1999; Richman et al. 1999). However, to our knowledge this potential issue has not been examined systematically in the context of measuring public opinion on political (particularly sensitive) topics.

Representativeness of Social Media Users

One crucial advantage we lose with social media relative to traditional surveys is the opportunity to control our sampling frame. Traditional surveys attempt to guarantee a known probability of any individual in the population being asked a survey question. Where those surveys fail is in both high and non-random nonresponse, and non-random item nonresponse. With social media, since control of the sampling frame is lost, we can neither know the likelihood that someone has been asked a "question" nor know the likelihood of a response. In a traditional survey, to generalize to a target population we have to assume that nonresponses are missing at random, or that they are missing at random conditioning on measured covariates. In the best of worlds, this is a strong assumption; it may be that people who choose not to reveal their preferences on something are systematically different than those who do reveal their preferences. On social media, where we do not ask the question but depend on the participants to reveal their opinions, we might have more trouble. The set of people offering *unprompted* opinions on a topic may be more passionate or different in myriad other ways from the set of people who offer opinions on that topic when explicitly asked. This presumably makes our missing data problems far worse than those caused by traditional survey nonresponse.

It would of course be extremely unwise to generalize directly from Twitter behavior to any of the standard populations of interest in most surveys. A number of studies have demonstrated that Twitter users are not representative of national populations (Duggan and Brenner 2015; Mislove et al. 2011; Malik et al. 2015). For example, in the United States most populous counties are overrepresented, and the user population is nonrepresentative in terms of race (Mislove et al. 2011). Comparing geotagged tweets and census data, Malik et al. (2015) also demonstrate significant biases toward younger users and users of higher income. Differences in usage rates of social media platforms across countries are also an obstacle for the comparative study of public opinion (Mocanu et al. 2013). These differences are also present, although perhaps to a lesser extent, in the analysis of other social media platforms like Facebook (Duggan and Brenner 2015).

For the purposes of the study of public opinion, however, it is more important whether and how representative the politically active Twitter users are relative to the general population. But here, too, the evidence is consistent with Twitter users being highly nonrepresentative. For example, women are the majority of Twitter users, but a much smaller minority among politically active Twitter users (Hampton et al. 2011); politically active Twitter users are more polarized than the general population (Barberá and Rivero 2014); and they are typically younger, better educated, more interested in politics, and ideologically more left wing than the population as a whole (Vaccari et al. 2013). Crucially, nonrepresentativeness may even vary by topic analyzed, as different issues attract different users to debate them (Diaz et al. 2014).[6]

Evaluating the representativeness of Twitter users is not straightforward, given that unlike standard surveys, Twitter does not record precise demographic information. Instead, most studies try to infer these characteristics. While some approaches have been quite successful (see, e.g., Al Zamal, Liu, and Ruths 2012a; Barberá and Rivero 2014), these are still approximations. These difficulties can be compounded by the possibility of bots and spammers acting like humans (Nexgate 2013), especially in the context of autocratic regimes (Sanovich 2015). It becomes much harder to infer how representative tweets are of any given population if some tweets come from automated computer programs, not people. And even determining how many people are paying attention to discussions is problematic, as fake accounts can be used to inflate common metrics of popularity. For example, one study found at least twenty sellers of followers on eBay, at an average price of $18 per thousand followers, demonstrating how fake accounts can rack up followers very easily (Barracuda Labs 2012).[7] In addition, there may be important deviations from one-to-one correspondences between individual users and individual accounts, given the existence of duplicate and parody accounts and accounts that represent institutions, companies, or products, such as the White House, Walmart, or Coca-Cola.

Moreover, the demographic composition of users can change over time, particularly in response to important events, such as presidential debates or primaries. These changes may be quite unpredictable. For example, during the presidential debates in the 2012 presidential election in the United States, the male overrepresentation among political tweeters dropped significantly, whereas the geographic distribution of tweets (by region) became considerably less representative (Diaz et al. 2014). In the Spanish context, Barberá and Rivero (2014) find that important events during the 2011 legislative election, such as party conferences and the televised debates, increased the inequality on Twitter by increasing the rate of participation of the most active and most polarized users. It is important to keep these shifts in mind, since raw aggregates of public opinion may be due to these shifts in demographic composition rather than any shifts in actual opinion (see, e.g., Wang et al. 2015).

Aggregating from Individual Responses to Public Opinion

A number of other platform-specific issues also affect researchers’ ability to aggregate individual social media messages. At present, access to 100% of tweets is only available through third-party companies like Gnip (recently bought by Twitter) at prices often beyond what most researchers can afford. Instead, researchers rely on Twitter’s streaming application programming interface (API), which only provides content in real-time, not historical, data. That means most researchers have to anticipate in advance the period of study they will focus on. Results can change significantly when using different time windows (Jungherr 2014), which can lead to ad hoc choices of period of coverage and a non-negligible likelihood of missing key events.

Most important, Morstatter et al. (2013) and González-Bailón et al. (2014) found significant differences between the full population of tweets (the so-called Twitter “firehose”) and the samples obtained through Twitter’s streaming API, the most popular source of data used by researchers. In particular, it appears the rate of coverage (the share of relevant content provided by the streaming API relative to all content) varies considerably over time; the topics extracted through text analysis from Streaming API can significantly differ from those extracted from the Firehose data; that users who participate less frequently are more likely to be excluded from the sample; and that top hashtags from the streaming API data can deviate significantly from the full data when focusing on a small number of hashtags.

In those cases in which researchers are interested in aggregating data from social media to specific geographic units such as a state or congressional district, they face the problem that only a small proportion of tweets are annotated with exact coordinates

(Leetaru et al. 2013). Geolocated tweets are highly precise but are not a representative subset of all tweets (Malik et al. 2015). An alternative is to parse the text in the “location” field of users’ profiles. While this increases the degree of coverage, it is not a perfect solution either, as Hecht et al. (2011) found that up to a third of Twitter users do not provide any sort of valid geographic information in this field.

Finally, one important issue often overlooked in social media studies is that, given Twitter’s opt-in nature, tweets often cannot be treated as independent because many individuals tweet multiple times. It is often the case that a minority of unique individuals dominates the discussion in terms of tweet and retweet volume, making oversampling of most active users very likely (Barberá and Rivero 2014; Gruzd and Haythornthwaite 2013; Mustafaraj et al. 2011). For example, in the run-up to the 2012 presidential election, 70% of tweets came from the top 10% of users, with 40% of the tweets coming from the top 1% of users (Barberá and Rivero 2014). This problem is exacerbated by practices such as astroturfing—coordinated messaging from multiple centrally controlled accounts—disguised as spontaneous behavior (Castillo, Mendoza, and Poblete 2011; Morris et al. 2012). Importantly, politically motivated actors use astroturf-like strategies to influence the opinions of their candidates during electoral campaigns (Kreiss 2014; Mustafaraj and Metaxas 2010). The more influential are the attempts to characterize the behavior of online users, the greater may be the incentive to manipulate such behavior (Lazer et al. 2014).

How Should It Be Done? Potential Solutions and Areas for Future Research

We argue that the three concerns about using social media data to measure public opinion outlined in the previous section—measuring opinion, assessing representativeness, and overcoming technical challenges in aggregation—are the main challenges to overcome in this field. Each of these stages has its analog in traditional survey methodology, but each presents unique challenges when using social media. In this section we describe current efforts by previous studies to address these issues and potential solutions that could be implemented in future research.

Better Methods for Identifying Political Opinion

In choosing the corpus of tweets that will be included in the analysis, previous studies often defined a set of ad hoc search criteria, such as a list of hashtags related to an event or the names of political actors. This is partially driven by the limitations imposed by Twitter’s streaming API and researchers’ inability to collect historic data freely. We claim that it is necessary to establish more systematic criteria to select what set of tweets will be included in the sample.

One approach that has yielded promising results is the development of automated selection of keywords. He and Rothschild (2014) apply such a method in their study of the 2012 U.S. Senate elections. They started with a corpus drawn based on candidate names, then iteratively expanded it by identifying the most likely entities related to each candidate. Their final corpus is 3.2 times larger, which gives an indication of the magnitude of the potential biases associated with simple keyword selection methods. For example, they find that the aggregate sentiment of tweets mentioning only candidate names is different from that of the extended corpus after applying their selection method. King, Lam, and Roberts (2014) also propose a similar method that adds human supervision in the selection of new keywords to resolve linguistic ambiguities and reduce the proportion of false positives.

An alternative solution is to abandon keyword filtering altogether and instead sample at the user level. As Lin et al. (2013) demonstrate, tracking opinion shifts within a carefully selected group of Twitter users can overcome some of the limitations mentioned above by learning from users' prior behavior to detect their biases and controlling for it in any analysis.[8] These "computational focus groups" can be further improved if they are combined with surveys of Twitter users that contain questions about sociodemographic and political variables (Vaccari et al. 2013).

In addition to topics, the other half of assessing opinion is the predicate side, such as the estimation of sentiment about those topics. One of the most successful examples of sentiment analysis applied to election prediction, the Voices from the Blogs project (Ceron et al. 2014; Ceron, Curini, and Iacus 2015), combines supervised learning methods with human supervision in the creation of data sets of labeled tweets that are *specific to each example*. González-Bailón and Paltoglou (2015) conducted a systematic comparison of dictionary and machine-learning methods, finding similar results: classifiers trained with a random sample of the data set to be used for prediction purposes outperformed dictionary methods, which are in many cases no better than random. One possible refinement of application-specific methods is the combination of topic models and sentiment analysis (Fang et al. 2015), which could leverage differences in words' usage across different topics to improve the performance of these techniques.

Increasing Representativeness

The majority of studies using Twitter data, particularly those estimating voting preferences and predicting election outcomes, do not attempt to address the nonrepresentativeness of (politically active) Twitter users (the exceptions include Gayo-Avello, 2011; Choy et al., 2011; 2012). In fact, many of these studies do not clearly specify the target population, which in the case of electoral predictions should be the voting population. The implicit assumption is that the size of the data, the diversity of Twitter users, and the decentralized nature of social media may compensate for any potential bias in the sample. Of course as we know in cases where it has been studied, the set of Twitter users is *not* representative of typical target populations such as voters or eligible voters (see, e.g., Duggan and Brenner 2015).

Significantly more work is needed to examine the plausibility of these assumptions. On the one hand, for predictive purposes, the skew in the sample may not be problematic if politically active users on Twitter act as opinion leaders who can influence the behavior of media outlets (Ampofo, Anstead, and O'Loughlin 2011; Farrell and Drezner 2008; Kreiss 2014) or a wider audience (Vaccari et al. 2013). On the other hand, as discussed in the previous section, the nonrepresentativeness of these users relative to the general population may be quite severe, suggesting that the biases may not balance out unless addressed by reweighting.

One potentially promising method is multilevel regression and post-stratification (MRP), particularly because it relies on post-stratification adjustments to correct for known differences between the sample and the target population (Little 1993; other potential weighting approaches can be found in AAPOR 2010). Somewhat like traditional weighting in telephone or online polls, this approach partitions the target population into cells based on combinations of certain demographic characteristics, estimates via multilevel modeling the variable of interest in the sample within each cell (e.g., average presidential approval for white females, ages 18–29), and then aggregates the cell-level estimates up to the population level by weighting each cell by the proportion in the target population (Park, Gelman, and Bafumi 2004; Lax and Phillips 2009). This approach has been fruitfully used to generate quite accurate election predictions from highly nonrepresentative samples, such as XBox users (Wang et al. 2015).

The main challenge with this approach is of course to obtain the detailed sample demographics needed for post-stratification. Twitter does not collect or provide data on demographics. And unlike some other platforms such as Facebook, Twitter metadata and profile feeds contain limited information to directly classify users. There are two ways to address this concern: first, consider demographic variables as latent traits to be estimated, and second, augment Twitter data with other types of data, such as voter registration records or surveys.

Pennacchiotti and Popescu (2011) and Rao et al. (2010) provide proofs of concept that demonstrate that coarse categories of age, political orientation, ethnicity, and location can be estimated by applying a variety of supervised machine-learning algorithms to user profiles, tweets, and social networks. Al Zamal, Liu, and Ruths (2012b) demonstrate that users' networks (i.e., whom they follow and their followers) can be particularly informative about their age and gender. However, these studies often rely on small convenience samples of labeled users, and it is still an open question whether these methods can scale up to the large samples researchers often work with.

One of the key variables in MRP applications has been party identification (Park, Gelman, and Bafumi 2004; Lax and Phillips 2009). Thus it is extremely useful to be able to infer ideological orientation and partisanship, in addition to gender, ethnicity, age, and geographic location. There are several promising approaches in this direction. Barberá (2015) shows that Twitter users' ideology can be accurately estimated by observing what political actors they decide to follow. Other studies estimate political ideology or partisan identification using different sources of information, such as the structure of retweet interactions, follower networks, or similarity in word use with respect to political elites (Boutet, Kim, and Yoneki 2013; Cohen and Ruths 2013; Conover et al. 2011; Golbeck and Hansen 2011; Wong et al. 2013). One limitation of these approaches is that ideology, as well as the other demographic variables, often cannot be estimated for the entire sample of users, or at least with the same degree of accuracy, especially if they rely on usage of specific hashtags, which can vary significantly across users.

An alternative solution to this problem is to augment Twitter data with demographic information from other sources. For example, Bode and Dalrymple (2014) and Vaccari et al. (2013) conducted surveys of Twitter users by sampling and directly contacting respondents through this platform, achieving relatively high response and completion rates. By asking respondents to provide their Twitter user names, they were able to learn key characteristics of a set of Twitter users directly from survey responses provided by those users. Matching Twitter profiles with voting registration files, publicly available in the United States, can also provide researchers with additional covariates, such as party affiliation, gender, and age (see, e.g., Barberá, Jost, et al. 2015). The subset of users for which this information is available could then be used as a training data set for a supervised learning classifier that infers these sociodemographic characteristics for all Twitter users.[9] These matching approaches could also be conducted at the zipcode or county level with census data to control for aggregate-level income or education levels (see, e.g., Eichstaedt et al. 2015).

Improving Aggregation

It is perhaps the last step—aggregating from tweets to a measure of public opinion—on which most attention has been placed in previous studies. We now have a good understanding of the biases induced by how Twitter samples the data that will be made available through the API (Morstatter et al. 2013; González-Bailón et al. 2014), the power-law distribution of users' Twitter activity (Barberá and Rivero 2014; Wu et al. 2011), and the fact that very few tweets contain enough information to locate their geographic origin (Leetaru et al. 2013; Compton, Jurgens, and Allen 2014). Researchers need to be aware of these limitations and address them in their analyses. For example, if the purpose of a study is to measure public opinion about a topic, then the analysis should add weights at the user level to control for different levels of participation in the conversation. When such a solution is not possible, the study should include a discussion of the direction and magnitude of the potential biases introduced by these limitations.

Finally, regardless of the approaches to aggregation, weighting, or opinion measurement that we choose, an important step in any analysis should be the removal of spam messages and accounts (or *bots*), which in some cases can represent a large share of the data set (King, Pan, and Roberts 2016). One option is to apply simple filters to remove users who are not active or exhibit suspicious behavior patterns. For example, in their study of political communication on Twitter, Barberá, Jost, Nagler, Tucker, and Bonneau (2015) only considered users who sent tweets related to at least two different topics, which should filter spam bots that “hijack” a specific trending topic or hashtag (Thomas, Grier, and Paxson 2012). Ratkiewicz et al. (2011) and Castillo, Mendoza, and Poblete (2011) implemented more sophisticated methods that rely on supervised learning to find accounts that are intentionally spreading misinformation. Their study shows that spam users often leave a distinct footprint, such as a low number of connections to other users, high retweet count among a limited set of strongly connected (and likely fake) users, and a string of very similar URLs (e.g., differing only in mechanically created suffixes). Therefore, it appears possible and therefore potentially warranted to invest more effort in preprocessing the data by removing the suspect content, or at least in inspecting the sensitivity of the results to the presence of bot accounts.

Validation

Once we have specified our data collection and aggregation strategies, our population of interest and weighting strategies, and our opinion measurement methods, it is essential to validate these purported measures against trusted ground truths, or at least against previously established measures. The success of these approaches must be examined relative to clear benchmarks, such as previous election results, existing surveys, public records, and manually labeled data (Metaxas, Mustafaraj, and Gayo-Avello 2011; Beauchamp 2016). This validation should be conducted with out-of-sample data, ideally forward in time, and should be measured statistically, by computing the predicted accuracy. Depending on the application, other forms of validity should be considered, such as convergent construct validity (the extent to which the measure matches other measures of the same variable) or, in the case of topic-specific measures, semantic validity (the extent to which each topic has a coherent meaning).[10]

Conversely, rather than engaging in demographics-based weighting and topic/sentiment estimation to predict public opinion, it may also be possible to reverse the validation process and instead train machine-learning models to sift through thousands of raw features (such as word counts) to find those that directly correlate with variations in the quantities of interest (such as past polling measures of vote intention) (Beauchamp 2016). In this way, one could potentially go directly from word counts and other metadata (such as retweets, URLs, or network data) to opinion tracking with no worry about demographics, topics, or sentiments—although potentially at the cost of interpretability and generalizability to other regions, times, and political circumstances.

New Directions for Measuring Public Opinion: Going beyond Survey Replication

As we have said, each of the challenges to using social media data to measure public opinion also reveals how social media can be taken well beyond existing survey methods.

Weighting and demographics aside, the sheer size of social media data make it theoretically possible to study subpopulations that would not be possible with traditional survey data, including those defined by demographic, geographic, or even temporal characteristics (Aragón et al. 2016; Barberá, Wang, et al. 2015). Social media also enable us to measure opinion across national borders. While Twitter penetration varies in different countries (Poblete et al. 2011), as long as we know something about the characteristics of who in a country is on Twitter, we can try to generalize from tweets to a measure of mass opinion in a country.

Because of their organic nature, social media data are generated continuously, and thus we can track changes over time at very fine-grained temporal units (e.g., Golder and Macy 2011 track changes in mood across the world over a course of one day). This means we can aggregate the data by any temporal unit we choose, and simulate designed data for tracking opinion change over time and for using in traditional time-series analysis. Moreover, because social media data come with individual identifiers, they also constitute panel data. We (often) have repeated observations from the same informant. This high frequency means that social media can reveal public opinion changes over time about issues that are not polled very frequently by traditional surveys. Reasonably dense time-series survey data exist for some issues, such as presidential approval or consumer sentiment, but social media data offer the opportunity to put together dense time series of public opinion on a host of specific issues that are rarely or infrequently polled.[11] And by taking advantage of information identifying characteristics of informants, those time series could be evaluated for distinct subgroups of populations.

The temporal nature of social media also lets us observe the emergence of public opinion. This is perhaps social media’s greatest strength; it does not depend on the analyst to ask a preconceived question. So while at some point almost no survey firm would think to invest in asking respondents whether or not they thought gay marriage or marijuana should be legal, by collecting sufficiently large collections of social media posts in real time, it should be possible to observe when new issues emerge, and indeed to identify these newly emerging issues *before* we even know we should be looking for them. While the free-form nature of opinion revelation on social media can be a barrier to measuring what *everyone* is thinking about an issue, it may give us a way to measure not just sentiment, but intensity of sentiment via content and retweeting, as well as richer measures of sentiment along as many potential dimensions as there are topics.[12]

Social media also allow us to measure not just mass opinion, but especially that of political activists and other elites. Legislators, political parties, interest groups, world leaders, and many other political elites tweet. And while these are public revelations and not necessarily truthful, we do see what these actors choose to reveal. We are able to see this contemporaneously with mass opinion revelation. And again, by taking advantage of the fine-grained, temporal nature of social media data, we can observe how elite opinion responds to mass opinion and vice versa (Barberá et al. 2014; Franco, Grimmer, and Lee 2016) and how both groups respond to exogenous events. While of course we need to be sensitive to the fact that elites know that what they are posting on Twitter or Facebook is intended for public consumption and thus may not reflect genuine "opinion" in the sense that we are trying to measure mass opinion, the social media record of elite expression may nevertheless prove extremely valuable for studying both elite communication strategy generally as well as *changes* in the issues that elites are emphasizing. Thus, while we may not know whether a particular politician genuinely believes gun control laws need to be changed, social media can easily help us measure whether that politician is emphasizing gun control more at time t than at time t-1.

Social media also come with a natural set of contextual data about revealed opinion: the social network of the individual informant (Larson et al. 2016). Measuring this with survey questions is notoriously difficult because people have not proven capable of stating the size of their networks, much less providing information necessary for contacting network members. Yet social media provide us not only with the size of the social networks of informants, but also a means to measure the opinions of network members. Traditional surveys have tried to measure the connection of network ties primarily by depending on self-response, which has proven to be unreliable.

Social media data also provide the potential to link online information directly to opinion data. Many social media users reveal enough information about themselves to make it possible to link them to public records such as voter files. Social media data can also be directly supplemented with survey data obtained by contacting social media users directly through Twitter, via "replies" (as in Vaccari et al. 2013) or promoted tweets targeted to a specific list of users. However, it remains challenging to construct matched samples using these more direct methods that will be sufficiently large and representative or that can be reweighted to ensure representativeness.

Finally, social media significantly democratize the study of public opinion. Researchers can potentially address novel research questions without having to field their own surveys, which are often much more costly. Access to large volumes of social media is free and immediate, unlike many existing surveys that may be embargoed or restricted. Moreover, this accessibility extends well beyond scholars of public opinion: Anyone—from campaigns selling political candidates or consumer goods, to regimes trying to understand public wants—can access social media data and see what people are saying about any given issue with minimal expense. Even in a world in which the numbers of surveys and surveyed questions are proliferating, social media potentially offer a spectrum of topics and a temporal and geographic density that cannot be matched by existing survey methods.

A Research Agenda for Public Opinion and Social Media

Making better use of social media for measuring public opinion requires making progress on multiple fronts. Perhaps the issue that remains the most theoretically challenging is the measurement of topic and sentiment: the “question” and “response.” The unstructured text is what is most unusual about social media—it is *not* an answer to a question specified by the researcher but rather free-form writing—and attempts to transform it into the traditional lingua franca of opinion research—answers to specific questions—remains an open problem. We may even eventually discover that this approach is obsolete, as we move entirely beyond the constraints imposed by traditional surveys into the naturally high-dimensional world of free-form text. The tremendous opportunity presented by social media data makes the payoff for solving such problems worth the investment. Social media data can give us measures of public opinion on a scale both geographically, temporally, and in breadth of subject that is vastly beyond anything we can measure with other means.

There are of course major mistakes that can be made when analyzing social media data. One advantage we have pointed out about social media is that they democratize measuring public opinion: anyone can do it. That means anyone can do it badly. And the endemic deficiency of ground-truths can make it difficult to know when a measure is a bad measure. Reputable scholars or organizations reporting measures based on traditional polls have adopted standardized practices to increase the transparency of their reporting (items such as sample size and response rates). We thus conclude with some obvious standards in reporting that social medi–based measures of opinion should adhere to in order to at least guarantee a minimum amount of transparency and allow readers, or users of the measures created, to better evaluate the measures. We describe and list standards with respect to analyzing data from Twitter, but these can be easily applied to other sources of social media with appropriate modifications. The points apply generally across organic public opinion data.

First, researchers need to be clear about the technical means of gathering data. Data gathered in real time through a rate-limited means could be incomplete, or differ in unpredictable ways from data purchased after the fact from an archive or collected thru Firehose. Second, researchers need to very clearly report the limitations placed on the sampling frame. Data on social media can be gathered based on user identification or on content of text, and can be further filtered based on other metadata of users or individual tweets (such as language or time of day).

Second, researchers need to explain whether data were gathered based on the context of the text, the sender of the text, or some contextual information (such as time or place of tweet).

Third, researchers need to very precisely describe the criteria for inclusion in their sample, and how those criteria were arrived at. If a sample is based on keywords, researchers need to describe how the keywords were selected. If someone claims to be measuring opinion about gun control, they could state that: “we collected all tweets about gun control.” But this claim could not be evaluated unless the full set of keywords used to gather tweets is provided. One could collect all tweets containing the expression “gun control,” but that would omit many tweets using assorted relevant hashtags and phrases. Or, if an analyst were to try to measure public opinion about Barack Obama with all tweets containing “Barack Omaba,” the analyst would miss any tweets that are about Barack Obama, but do not use his name. If one were to omit all tweets that contain “Barack Hussein Obama” rather than “Barack Obama,” then this could obviously cause significant measurement error. Thus precise description of what can be included in the corpus of text is important.

If the corpus is collected with a set of keywords, the analyst should explain how the set was generated. Did the analyst assume omniscience and create the list of keywords, or was it generated using some algorithim proceeding from a set of core terms and finding co-occuring terms? And no matter how the keywords were chosen, or how the topic was collected, the analyst should describe any testing done to confirm that the corpus was in fact about the chosen topic.

Researchers also need to note whether filters or constraints were imposed on the collection that would exclude some tweets based on language or geography. If we are *only* measuring opinions expressed in English, that is important. Similarly, if we filter out all tweets that cannot be identified as being from a particular geographic region, that is important. And researchers need to clearly explain how any such constraints were implemented. If a language filter was used, precisely what was the filter? If a geographic constraint was imposed, what was it? Were only geocoded tweets considered? Or were metadata about the source of the tweet considered, and if so, how?

If tweets are aggregated by topic, the analyst must explain the aggregation method used. Or the analyst must explain how tweets were assigned to topics, whether by topic modeling or by human assignment. If by topic modeling, the analyst should provide information on how robust the results are to variations in the number of topics selected. Information about the criteria for tweets being assigned to topics (such as top terms from linear discriminant analysis) is essential. And the analyst should indicate whether the topics were validated against human judgments.

If a collection of tweets claiming to measure public opinion excludes tweets *by some individuals*, that information is crucial and needs to be provided. Such exclusions could be based on the individual’s frequency of social media use, on whether he or she is part of a set of individuals following particular political (or nonpolitical) actors. Such exclusion could also be based on the individual's characteristics, such as use of language, demographic characteristics, or political characteristics. And as such characteristics are often estimated or inferred from metadata, the analyst must be precise and transparent about how characteristics for individuals are inferred. Precise data on *who* is eligible to be included in the data set are essential for any attempt to draw a population inference.

If a measure of sentiment is given, the analyst must carefully explain how sentiment was calculated. If a dictionary was used, the analyst should explain how robust the measure was to variations in the dictionary—or across dictionaries. And the analyst should explain whether sentiment measures were validated in any way against human judgments.

Following is a set of guidelines:

1. Describe the source of data for the corpus.
  (a) Were data taken from the Twitter Firehose, or from one of the rate-limited sources?
  (b) Were data retrieved using the rest API or streaming API?

2. Describe whether the criteria for inclusion in the corpus were based on the text, the sender of the text, or contextual information (such as the place or time of the tweet).
3. For corpora collected based on keywords or regular expressions in the text, describe the criteria for inclusion.
  (a) What were the criteria by which the keywords or regular expressions were selected?
  (b) Were keywords or regular expressions chosen based on

- the analyst's expertise or prior beliefs?
- an extant document?
- an algorithm used to generate keywords based on an initial seeding and further processing of the text?

  (c) Describe any testing done to confirm that tweets gathered were relevant for the intended topic for which the keywords, or regular expressions, were used.

4. For corpora collected using *topics* as the criterion for inclusion:
  (a) Describe how individual documents were determined to be relevant to the chosen topic (i.e., what were the technical requirements for inclusion in the corpus?).
  (b) Describe any testing done to estimate, or determine exactly, the number of documents in the corpora that were germane to their assigned topic(s).

5. If the content of tweets was *aggregated* by topic, after some selection criteria into the corpus, how were topics generated?
  (a) Were topics hand-coded?

(b) Was some form of automated topic generation method used?

(c) How robust are the topics to variations in the sample or the number of topics selected?

(d) Was inclusion of text into topics validated against human judgments, and if so, how?

6. Describe any limitations placed on the sampling frame that could limit *whose* opinions could be in the data. State any limitations of the sampling frame based on metadata provided by informants, either *directly* provided in metadata or *inferred* from metadata. If characteristics were inferred, explain the procedure for inference. This would include

(a) exclusions based on language *of the informant*,

(b) exclusions based on geography of the informant, and

(c) exclusions based on gender, age, or other demographic characteristics of the informant.

7. For corpora in which selection was based on the sender, how were the senders chosen?

8. Describe any constraints (or filters) imposed on the collection that would exclude some tweets from being included based on characteristics of the tweet, such as constraints on geography or language.

(a) Describe whether any geographic constraints were based on geocoding or on an algorithim used to infer geography from metadata.

(b) Describe how language was determined if language constraints were imposed.

9. If a sentiment measure, or a related measure, was applied, describe how the measure was calculated.

(a) If a dictionary was used, describe how robust the measure was to variations in the dictionary or across dictionaries.

(b) Were the sentiment measures validated against human judgments?

(c) What information, other than the text of the tweet, such as linked content or images, characteristics of the sender, or context, was used to determine sentiment?

10. Describe the aggregation method used to generate the quantity of interest.

(a) Describe precisely the temporal units used, including reference to time zone.

(b) Describe how retweets are treated.

For anyone interested in studying public opinion, it would be foolish to ignore the information about public opinion revealed by social media data. However, it would also be foolish to treat measurement of social media data in the same manner one treats a well-designed survey yielding something approximating a random sample of a population of interest. We have listed many of the reasons that this is not a viable strategy. Either one accepts that one has a nonrepresentative opt-in sample, which may or may not be a useful sample for some goal *other than* measuring mass public opinion, or one attempts to weight the sample. We think continued work on studying public opinion via social media is a fruitful endeavor. And we urge scholars and practioners to both work on improving our ability to measure mass public opinion via social media and to follow solid guidelines for reporting results obtained via social media.

References

AAPOR. 2010. “AAPOR Report on Online Panels.” **http://poq.oxfordjournals.org/content/early/2010/10/19/poq.nfq048.full.pdf?ijkey=0w3WetMtGItMuXs&keytype=ref**.

Abraham, K. G., S. Helms, and S. Presser. 2009. “How Social Processes Distort Measurement: The Impact of Survey Nonresponse on Estimates of Volunteer Work in the United States.” *American Journal of Sociology* 114 (4): 1129–1165.

Al Zamal, F., W. Liu, and D. Ruths. 2012a. “Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors.” In Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, 387–390, AAAI Press, Palo Alto, California.

Ampofo, L., N. Anstead, and B. O’Loughlin. 2011. “Trust, Confidence, and Credibility: Citizen Responses on Twitter to Opinion Polls During the 2010 UK general Election.” *Information, Communication & Society* 14 (6): 850–871.

Aragón, P., Y. Volkovich, D. Laniado, and A. Kaltenbrunner. 2016. “When a Movement Becomes a Party: Computational Assessment of New Forms of Political Organization in Social Media.” Proceedings of the Tenth International AAAI Conference on Weblogs and Social Media, 12–21, AAAI Press, Palo Alto, California.

Barberá, P. 2015. “Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data.” *Political Analysis* 23 (1): 76–91.

Barberá, P., R. Bonneau, P. Egan, J. T. Jost, J. Nagler, and J. Tucker. 2014. “Leaders or Followers? Measuring Political Responsiveness in the US Congress using Social Media Data.” Paper presented at the 110th American Political Science Association Annual Meeting.

Barberá, P., J. T. Jost, J. Nagler, J. Tucker, and R. Bonneau. 2015. “Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber?” *Psychological Science* 26 (10): 1531–1542.

Barberá, P., and G. Rivero. 2014. “Understanding the Political Representativeness of Twitter Users.” *Social Science Computer Review* 33 (6): 712–729.

Barberá, P., N. Wang, R. Bonneau, J. T. Jost, J. Nagler, J. T., and S. González-Bailón. 2015. “The Critical Periphery in the Growth of Social Protests.” *PloS one* 10 (11): e0143611.

Barracuda Labs. 2012. “The Twitter Underground Economy: A Blooming Business.” Internet security blog. **https://www.barracuda.com/blogs/labsblog?bid=2989**.

Beauchamp, Nicholas. 2016. “Predicting and Interpolating State-level Polls using Twitter Textual Data.” *American Journal of Political Science* 61 (2): 490–503.

Bermingham, A., and A. F. Smeaton. 2011. “On Using Twitter to Monitor Political Sentiment and Predict Election Results.” In *Sentiment Analysis: Where AI Meets Psychology (SAAIP)* Workshop at the International Joint Conference for Natural Language Processing, **http://doras.dcu.ie/16670/**.

Bode, L., and K. E. Dalrymple. 2014. “Politics in 140 Characters or Less: Campaign Communication, Network Interaction, and Political Participation on Twitter.” *Journal of Political Marketing* 15(4): 311–332.

Boutet, A., H. Kim, and E. Yoneki. 2013. “What’s in Twitter: I Know What Parties Are Popular and Who You Are Supporting Now!” *Social Network Analysis and Mining* 3 (4): 1379–1391.

Brick, J. M., and D. Williams. 2013. “Explaining Rising Nonresponse Rates in Cross-sectional Surveys.” *Annals of the American Academy of Political and Social Science* 645 (1): 36–59.

Bruns, A., K. Weller, M. Zimmer, and N. J. Proferes. 2014. “A Topology of Twitter Research: Disciplines, Methods, and Ethics.” *Aslib Journal of Information Management* 66 (3): 250–261.

Castillo, C., M. Mendoza, and B. Poblete. 2011. “Information Credibility on Twitter.” In *Proceedings of the 20th International Conference on World Wide Web*, Association for Computing Machinery, New York, NY, 675–684.

Ceron, A., L. Curini, and S. M. Iacus. 2015. “Using Sentiment Analysis to Monitor Electoral Campaigns Method Matters—Evidence from the United States and Italy.” *Social Science Computer Review* 33 (1): 3–20.

Ceron, A., L. Curini, S. M. Iacus, and G. Porro. 2014. “Every Tweet Counts? How Sentiment Analysis of Social Media Can Improve Our Knowledge of Citizens’ Political Preferences with an Application to Italy and France.” *New Media & Society* 16 (2): 340–358.

Choy, M., M. Cheong, M. N. Laik, and K. P. Shung. 2012. “US Presidential Election 2012 Prediction using Census Corrected Twitter Model.” **https://arxiv.org/abs/1211.0938**.

Choy, M., M. L. F. Cheong, M. N. Laik, and K. P. Shung. 2011. “A Sentiment Analysis of Singapore Presidential Election 2011 sing Twitter Data with Census Correction.” **https://arxiv.org/abs/1108.5520**.

Cohen, R., and D. Ruths. 2013. “Classifying Political Orientation on Twitter: It’s Not Easy!” Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 91–99, AAAI Press, Palo Alto, California.

Compton, R., D. Jurgens, and D. Allen. 2014. “Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization.” In *2014 IEEE International Conference on Big Data (Big Data)*, 393–401, **http://ieeexplore.ieee.org/abstract/document/7004256/?reload=true**.

Conover, M. D., B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. 2011. “Predicting the Political Alignment of Twitter Users.” In *Privacy, Security, Risk and Trust (PASSAT), and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, 192–199, **http://ieeexplore.ieee.org/document/6113114/**.

De Leeuw, E., and W. De Heer. 2002. “Trends in Household Survey Nonresponse: A Longitudinal and International Comparison.” In *Survey Nonresponse*, edited by R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, 41–54. New York: John Wiley & Sons.

Diaz, F., M. Gamon, J. Hofman, E. Kiciman, and D. Rothschild. 2014. “Online and Social Media Data as a Flawed Continuous Panel Survey.” Working Paper, Microsoft Research.

Duggan, M., and J. Brenner. 2015. *The Demographics of Social Media Users, 2014*. Pew Research Center’s Internet & American Life Project, vol. 14. Washington, DC: Pew Research Center.

Eichstaedt, J. C., H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, … M. Sap. 2015. “Psychological Language on Twitter Predicts County-level Heart Disease Mortality.” *Psychological Science* 26 (2): 159–169.

Fang, A., I. Ounis, P. Habel, and C. Macdonald. 2015. “Topic-centric Classification of Twitter User’s Political Orientation.” In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, 791–794.

Farrell, H., and D. W. Drezner. 2008. “The Power and Politics of Blogs.” *Public Choice* 134 (1–2): 15–30.

Flicker, S., D. Haans, and H. Skinner. 2004. “Ethical Dilemmas in Research on Internet Communities.” *Qualitative Health Research* 14 (1): 124–134.

Franco, A., J. Grimmer, and M. Lee. 2016. “Changing the Subject to Build an Audience: How Elected Officials Affect Constituent Communication.” Unpublished Manuscript.

Gayo-Avello, D. 2011. “Don’t Turn Social Media into Another ‘Literary Digest’ Poll.” *Communications of the ACM* 54 (10): 121–128.

Golbeck, J., and D. Hansen. 2011. "Computing Political Preference among Twitter Followers." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, 1105–1108.

Golder, S. A., and M. W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures." *Science* 333 (6051): 1878–1881.

González-Bailón, S., and G. Paltoglou. 2015. "Signals of Public Opinion in Online Communication A Comparison of Methods and Data Sources." *Annals of the American Academy of Political and Social Science* 659 (1): 95–107.

González-Bailón, S., N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno. 2014. "Assessing the Bias in Samples of Large Online Networks." *Social Networks* 38: 16–27.

Groves, R. M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70 (5): 646–675.

Groves, R. 2011. "'Designed Data' and 'Organic Data'." **http://directorsblog.blogs.census.gov/2011/05/31/designed-data-and-organic-data/**.

Groves, R. M., and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: a Meta-analysis." *Public Opinion Quarterly* 72 (2): 167–189.

Groves, R. M., F. J. Fowler, Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2011. *Survey Methodology*. New York: John Wiley & Sons.

Gruzd, A., and C. Haythornthwaite. 2013. "Enabling Community Through Social Media." *Journal of Medical Internet Research* 15 (10), **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842435/**.

Hampton, K., L. Sessions Goulet, L. Rainie, and K. Purcell. 2011. "Social Networking Sites and Our Lives." Pew Internet & American Life Project Report, **http://www.pewinternet.org/2011/06/16/social-networking-sites-and-our-lives/**.

He, R., and D. Rothschild. 2014. "Who Are People Talking about on Twitter?" Working Paper, Microsoft Research.

Hecht, B., L. Hong, B. Suh, and E. H. Chi. 2011. "Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, 237–246.

Hillygus, D. S. 2011. "The Practice of Survey Research: Changes and Challenges." In *New Directions in Public Opinion*, edited by A. Berinsky. Routledge Press, New York, NY.

Joinson, A. 1999. "Social Desirability, Anonymity, and Internet-based Questionnaires." *Behavior Research Methods, Instruments, & Computers* 31 (3): 433–438.

Jungherr, A. 2014. “Twitter in Politics: A Comprehensive Literature Review.” **https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2402443**.

Jungherr, A., P. Jürgens, and H. Schoen. 2012. “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 Review* 30 (2): 229–234.

Jungherr, A., H. Schoen, O. Posegga, and P. Jürgens. 2016. “Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support.” *Social Science Computer Review* 35 (3): 336–356.

Keeter, S. 2012. “Presidential Address: Survey Research, Its New Frontiers, and Democracy.” *Public Opinion Quarterly* 76 (3): 600–608.

King, G., J. Pan, and M. E. Roberts. 2016. “How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument.” Unpublished Manuscript.

King, G., P. Lam, and M. Roberts. 2014. “Computer-Assisted Keyword and Document Set Discovery from Unstructured Text.” **https://gking.harvard.edu/publications/computer-assisted-keyword-and-document-set-discovery-fromunstructured-text**.

Kreiss, D. 2014. “Seizing the Moment: The Presidential Campaigns’ Use of Twitter During the 2012 Electoral Cycle.” *New Media & Society* 18 (8): 1473–1490.

Kwak, H., C. Lee, H. Park, and S. Moon. 2010. “What Is Twitter, a Social Network or a News Media?” In *Proceedings of the 19th International Conference on World Wide Web*, Association for Computing Machinery, New York, NY, 591–600.

Larson, J., J. Nagler, J. Ronen, and J. A Tucker. 2016. “Social Networks and Protest Participation: Evidence from 93 Million Twitter Users.” *SSRN,***https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2796391**.

Lax, J. R., and J. H. Phillips. 2009. “How Should We Estimate Public Opinion in the States?” *American Journal of Political Science* 53 (1): 107–121.

Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014. “The Parable of Google Flu: Traps in Big Data Analysis.” *Science* 343 (March 14): 1203–1205.

Leetaru, K., S. Wang, G. Cao, A. Padmanabhan, and E. Shook. 2013. “Mapping the Global Twitter Heartbeat: The Geography of Twitter.” *First Monday* 18 (5), **http://firstmonday.org/article/view/4366/3654?__hstc=225085317.e835c34ab7bf88e972fdd7a7debc8575.1436140800094.1436140800095.**

Lin, Y.-R., D. Margolin, B. Keegan, and D. Lazer. 2013. “Voices of Victory: A Computational Focus Group Framework for Tracking Opinion Shift in Real Time.” In *Proceedings of the 22nd International Conference on World Wide Web*, Association for Computing Machinery, New York, NY, 737–748.

Little, R. J. A. 1993. “Post-Stratification: A Modeler’s Perspective.” *Journal of the American Statistical Association* 88 (423): 1001–1012.

Malik, M. M., H. Lamba, C. Nakos, and J. Pfeffer. 2015. “Population Bias in Geotagged Tweets.” In *Ninth International AAAI Conference on Weblogs and Social Media*, 18–27, AAAI Press, Palo Alto, California.

Marwick, A. E., and D. Boyd. 2011. “I Tweet Honestly, I Tweet Passionately: Twitter Users, Context collapse, and the Imagined Audience.” *New Media & Society* 13 (1): 114–133.

Metaxas, P. T., E. Mustafaraj, and D. Gayo-Avello. 2011. “How (Not) to Predict Elections.” In *Privacy, Security, Risk and Trust (PASSAT), and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, 165–171.

Metzger, M., R. Bonneau, J. Nagler, and J. A. Tucker. 2016. “Tweeting Identity? Ukranian, Russian, and #Euromaidan.” *Journal of Comparative Economics* 44 (1): 16–50.

Mislove, A., S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. 2011. “Understanding the Demographics of Twitter Users.” *ICWSM* 11 (5).

Mocanu, D., A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. 2013. “The Twitter of Babel: Mapping World Languages Through Microblogging Platforms.” *PloS One* 8 (4): e61981.

Mokrzycki, M., S. Keeter, and C. Kennedy. 2009. “Cell-phone-only Voters in the 2008 Exit Poll and Implications for Future Noncoverage Bias.” *Public Opinion Quarterly* 73 (5): 845–865.

Morris, M. R., S. Counts, A. Roseway, A. Hoff, and J. Schwarz. 2012. “Tweeting Is Believing? Understanding Microblog Credibility Perceptions.” In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, Association for Computing Machinery, New York, NY, 441–450.

Morstatter, F., J. Pfeffer, H. Liu, and K. M. Carley. 2013. “Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose.” In *ICWSM*.

Munger, K. 2015. “Elites Tweet to Get Feet Off the Streets: Measuring Elite Reaction to Protest Using Social Media.” Working paper, New York University.

Mustafaraj, E., and P. Metaxas. 2010. "From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search." Paper presented at WebSci10: Extending the Frontiers of Society On-Line, April 26–27, Raleigh, NC.

Mustafaraj, E., S. Finn, C. Whitlock, and P. T. Metaxas. 2011. "Vocal Minority Versus Silent Majority: Discovering the Opionions of the Long Tail." In *Privacy, Security, Risk and Trust (PASSAT), and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, 103–110.

Neuman, W. R., L. Guggenheim, S. M. Jang, and S. Y. Bae. 2014. "The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data." *Journal of Communication* 64 (2): 193–214.

Newman, M. W., D. Lauterbach, S. A. Munson, P. Resnick, and M. E. Morris. 2011. "It's Not That I Don't Have Problems, I'm Just Not Putting Them on Facebook: Challenges and Opportunities in Using Online Social Networks for Health." In *Proceedings of the ACM 2011 Conference on Computer-Supported Cooperative Work*, Association for Computing Machinery, New York, NY, 341–350.

Nexgate. 2013. "2013 State of Social Media Spam." Nexgate Report. **http://go.nexgate.com/nexgate-social-media-spam-research-report**.

O'Connor, B., R. Balasubramanyan, B. R. Routledge, and N. A. Smith. 2010. "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." *ICWSM* 11: 122–129.

Park, D. K., A. Gelman, and J. Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12 (4): 375–385.

Pavalanathan, U., and M. De Choudhury. 2015. Identity Management and Mental Health Discourse in Social Media." In *Proceedings of the 24th International Conference on World Wide Web Companion*, Association for Computing Machinery, New York, NY, 315–321.

Pennacchiotti, M., and A.-M. Popescu. 2011. "A Machine Learning Approach to Twitter User Classification." *ICWSM* 11: 281–288.

Pew Research Center. 2012. "Assessing the Representativeness of Public Opinion Surveys." **http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/**.

Poblete, B., R. Garcia, M. Mendoza, and A. Jaimes. 2011. "Do All Birds Tweet the Same? Characterizing Twitter Around the World." In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, New York, NY, 1025–1030.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. 2010. “How to Analyze Political Attention with Minimal Assumptions and Costs.” *American Journal of Political Science* 54 (1): 209–228.

Rao, D., D. Yarowsky, A. Shreevats, and M. Gupta. 2010. “Classifying Latent User Attributes in Twitter.” In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, Association for Computing Machinery, New York, NY, 37–44.

Ratkiewicz, J., M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. 2011. “Detecting and Tracking Political Abuse in Social Media.” In *ICWSM*. 297–304.

Richman, W. L., S. Kiesler, S. Weisband, and F. Drasgow. 1999. “A Meta-analytic Study of Social Desirability Distortion in Computer-administered Questionnaires, Traditional Questionnaires, and Interviews.” *Journal of Applied Psychology* 84 (5): 754.

Sanovich, S. 2015. “Government Response Online: New Classification with Application to Russia.” Unpublished Manuscript, New York University.

Skoric, M., N. Poor, P. Achananuparp, E.-P. Lim, and J. Jiang. 2012. “Tweets and Votes: A Study of the 2011 Singapore General Election.” In *System Science (HICSS), 2012 45th Hawaii International Conference on Systems Science*. (HICSS-45 2012). Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2583–2591.

Solberg, L. B. 2010. “Data Mining on Facebook: A Free Space for Researchers or an IRB Nightmare?” *Journal of Law, Technology and Policy* 2: 311–343.

Thomas, K., C. Grier, and V. Paxson. 2012. “Adapting Social Spam Infrastructure for Political Censorship. In *Proceedings of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats*. USENIX Association, Berkeley, CA, 13–13.

Tourangeau, R., R. M. Groves, and C. D. Redline. 2010. “Sensitive Topics and Reluctant Respondents: Demonstrating a Link Between Nonresponse Bias and Measurement Error.” *Public Opinion Quarterly* 74 (3): 413–432.

Tucker, J. A., J. Nagler, M. M. Metzger, P. Barberá, D. Penfold-Brown, and R. Bonneau. 2016. “Big Data, Social Media, and Protest: Foundations for a Research Agenda.” In *Computational Social Science: Discovery and Prediction*, edited by R. M. Alvarez. Cambridge University Press, New York, NY, 199–224.

Tufekci, Z., and C. Wilson. 2012. “Social Media and the Decision to Participate in Political Protest: Observations from Tahrir Square.” *Journal of Communication* 62 (2): 363–379.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M Welpe. 2010. “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.” *ICWSM* 10: 178–185.

Tuunainen, V. K., O. Pitkänen, and M. Hovi. 2009. "Users' Awareness of Privacy on Online Social Networking Sites—Case Facebook." In *Proceedings of the 22nd Bled eConference, eEnablement: Facilitating an Open, Effective and Representative eSociety*, Association for Information Systems, Atlanta, GA, 42–58.

Vaccari, C., A. Valeriani, P. Barberá, R. Bonneau, J. T. Jost, J. Nagler, and J. Tucker. 2013. "Social Media and Political Communication: A Survey of Twitter Users during the 2013 Italian General Election." *Rivista Italiana di Scienza Politica* 43 (3): 381–410.

Veenstra, A., N. Iyer, N. Bansal, M. Hossain, and J. Park, 2014. "#Forward! Twitter as Citizen Journalism in the Wisconsin Labor Protests." Paper presented at the Annual Meeting of the Association for Education in Journalism and Mass Communication, St. Louis, MO.

Wang, W., D. Rothschild, S. Goel, and A. Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting*, 31 (3): 980–991.

Wong, F., M. Fai, C. W. Tan, S. Sen, and M. Chiang. 2013. "Quantifying Political Leaning from Tweets and Retweets." In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, AAAI Press, Palo Alto, California, 640–649.

Wu, S., J. M. Hofman, W. A. Mason, and D. J. Watts. 2011. "Who Says What to Whom on Twitter." In *Proceedings of the 20th international Conference on World Wide Web*. Association for Computing Machinery, New York, NY, 705–714.

Zimmer, M. 2010. "'But the Data Is Already Public': On the Ethics of Research in Facebook." *Ethics and Information Technology* 12 (4): 313–325.

Notes:

([1]) While the response rates for the "gold standard" surveys such as the General Social Survey, the American National Election Study, and the National Household Education Survey are higher, they too have been falling off markedly (Brick and Williams 2013; Hillygus 2011).

([2]) See, for example, http://www.businesswire.com/news/home/20150402005790/en#.VR2B1JOPoyS.

([3]) The newspaper industry, a major source of public opinion polls, shrank 43% from 2000 to 2012 (see http://www.stateofthemedia.org/2012/overview-4/). The declining public support to higher education due to the financial crisis of 2008–2009 led to the closing of some university-based survey research centers (Keeter 2012), and there has been increasing political pressure to defund such initiatives as the American Community Survey and the Economic Census. Overall interest in polls, however, has only grown, with the total number of active pollsters (with at least ten polls per campaign) having risen since 2000: in presidential years, this has increased from appoximately ten to twenty polls per presidential campaign over the last two decades and from approximately five to ten polls for midterm elections (based on our analysis of data from http://projects.fivethirtyeight.com/pollster-ratings/).

([4]) The Twitter archive is of course dwarfed by the Facebook archive, but this is not yet available to the public. And to be clear, by "available" we mean available for purchase; collecting relatively large amounts of Twitter data is free in real time, but it is not free to retrieve tweets with a broad backward-looking search.

([5]) It also raises all sorts of new questions for social scientists, who will find themselves in the future wanting to work with huge private companies, such as Facebook or Twitter, much in the way that natural scientists have had to learn how to work with big pharma. Although this discussion is beyond the scope of this article, this too will likely pose all sorts of new challenges for researchers, the likes of which we have previously rarely encountered.

([6]) Note that all of the studies cited here are country specific; we cannot really make these claims about the global set of Twitter users.

([7]) Such concerns could be particularly pernicious if politicians are buying bots precisely for the *purpose* of manipulating measures of public opinion. Although we do not yet have evidence of this occurring, it does not seem to be a large leap to imagine politicians moving from simply buying followers to buying accounts that will deliver positive sentiment about themselves (or negative sentiment about opponents) in an attempt to manipulate reports in the media about online popularity.

([8]) See below for a discussion of working with a randomly chosen set of users.

([9]) As an additional challenge, social media users and their demographic distributions are presumably constantly evolving, so these models will have to be frequently updated to keep up with this rapidly shifting landscape.

([10]) See Quinn et al. (2010) for a more extensive discussion of different types of validity.

([11]) For example, such topics might include intermittently polled issues in the United States, like gun control or immigration; government approval measures in less well-polled nations; public opinion about specific foreign or domestic policies (e.g., Syria or the Affordable Care Act) or factual questions (e.g., climate change or genetically modified organisms); and more local issues, such as opinion on the policies or services in specific cities.

([12]) In addition to issues with representativeness, the public nature of social media means that these sentiments are presumably also affected by social desirability bias. It may be that in these more polarized times, mean sentiment will remain representative even as both sides are driven to extremes by social pressures, but it will nevertheless be important to measure and correct for these effects using existing polling measures as ground-truth tests.

Marko Klašnja

Government, Georgetown University

Georgetown University

Pablo Barberá

International Relations, University of Southern California

University of Southern California

Nick Beauchamp

Political Science, Northeastern University

Northeastern University

Jonathan Nagler

Politics, New York University

Jonathan Nagler is Professor of Politics, Wilf Family Department of Politics, New York University.

Joshua Tucker

Politics, New York University

New York University