Special Issue

Social Media as an Alternative to Surveys of Opinions About the Economy

Social Science Computer Review
2021, Vol. 39(4) 489-508

DOI: 10.1177/0894439319875692
journals.sagepub.com/home/ssc
SAGE

Frederick G. Conrad[1,*], Johann A. Gagnon-Bartsch[1,*], Robyn A. Ferg[1], Michael F. Schober[2], Josh Pasek[1], and Elizabeth Hou[1]

Abstract

There is interest in using social media content to supplement or even substitute for survey data. In one of the first studies to test the feasibility of this idea, O'Connor, Balasubramanyan, Routledge, and Smith report reasonably high correlations between the sentiment of tweets containing the word "jobs" and survey-based measures of consumer confidence in 2008–2009. Other researchers report a similar relationship through 2011, but after that time it is no longer observed, suggesting such tweets may not be as promising an alternative to survey responses as originally hoped. But, it's possible that with the right analytic techniques, the sentiment of "jobs" tweets might still be an acceptable alternative. To explore this, we first classify "jobs" tweets into categories whose content is either related to employment or not, to see whether sentiment of the former correlates more highly with a survey-based measure of consumer sentiment. We then compare the relationship when sentiment is determined with traditional dictionary-based methods versus newer machine learning-based tools developed for Twitter-like texts. We calculated daily sentiment in three different ways and used a measure of association less sensitive to outliers than correlation. None of these approaches improved the size of the relationship in the original or more recent data. We found that the many micro-decisions these analyses require, such as the size of the smoothing interval and the length of the lag between the two series, can significantly affect the outcomes. In the end, despite the earlier promise of tweets as an alternative to survey responses, we find no evidence that the original relationship in these data was more than a chance occurrence.

Keywords

social media, Twitter, surveys, Big data and surveys, Twitter sentiment, consumer sentiment

This article is part of the SSCR special issue on "Big Data and Survey Science," guest edited by Adam Eck (Oberlin College), Ana Lucia Cordova-Cazar (Universidad San Francisco de Quito), Mario Callegaro (Google Ltd.), and Paul Biemer (UNC-CH).

[1] University of Michigan, Ann Arbor, MI, USA
[2] The New School for Social Research, New York, NY, USA
* These authors contributed equally to this work.

Corresponding Author:
Frederick Conrad, University of Michigan, Ann Arbor, MI 48103, USA.
Email: fconrad@umich.edu.

Social scientists have made important progress in understanding social, political, and economic problems by collecting and analyzing survey data. Sample surveys that allow generalization to a larger population have been at the heart of the social research paradigm for 70 or more years. Recently, there has been considerable enthusiasm among researchers about new types of data, such as social media content, which may be timelier and less expensive than traditional survey data (e.g., Hsieh & Murphy, 2017; Schober, Pasek, Guggenheim, Lampe, & Conrad, 2016) and which could supplement or even substitute for surveys in a range of domains. The general approach has so far involved transforming social media content into a form that can be compared to or combined with survey data. For example, the content of a social media post might be transformed into sentiment scores. These scores are then aggregated over the posts, and, if the goal is to enhance survey data, they may be added to the statistical models built on the survey data; if the goal is to match survey results as a demonstration that social media content can potentially be substituted for survey data, the correspondence between the two data sources is calculated over a given time period.

Smith and Gustafson (2017) conducted a study that illustrates the first approach (augmenting survey data). They incorporated Wikipedia page views for candidates in about 100 U.S. Senate races between 2008 and 2012 into models predicting the election outcomes based on preelection polls. The authors made the critical assumption that visiting a candidate's page was associated with increased likelihood of voting for the candidate. They compared *simple* models consisting of the proportion of likely vote share for the Democratic candidate predicted by the poll as well as "fundamentals" (e.g., presidential approval, incumbency, and economic indicators) to *synthesized* models consisting of the simple model plus log-transformed counts of candidate page views. The simple models predicted election outcomes quite accurately, but the synthesized models performed significantly better.

The second approach (substituting social media content for survey data) is more ambitious than the first approach yet, paradoxically, has been tested more often (Schober et al., 2016). O'Connor, Balasubramanyan, Routledge, and Smith (2010) conducted one of the first and, to date, most influential studies of the correspondence between sentiment in social media and survey responses. They compared the daily sentiment of tweets containing a particular key word between 2008 and 2009 to survey estimates of U.S. consumer confidence and public opinion (presidential job approval and preelection polls) over the same time period. For consumer confidence, O'Connor et al. (2010) selected tweets containing the word "jobs" and compared the sentiment of these tweets to Gallup's Economic Confidence Index and the University of Michigan's Index of Consumer Sentiment (ICS). For presidential job approval, the authors compared the sentiment of tweets containing "Obama" to support for Obama derived from Gallup's Daily Tracking Poll. For election prediction, they compared the sentiment of tweets containing either "Obama" or "McCain" to the percent support for Obama in a compilation of 2008 U.S. presidential preelection polls prepared by Pollster.com.

The results included some reasonably high correlations between Twitter sentiment and survey data. For example, for consumer confidence, Twitter sentiment correlated with the Gallup index $r = .79$ and with the Michigan index $r = .64$. These particular correlations benefited from smoothing the Twitter sentiment and imposing a lag between the dates of the tweets and of the survey data (the correlations were higher when the tweets preceded the survey data). The particular smoothing intervals and lags that resulted in optimal correlations were different for the different sets of survey data (e.g., 30-day smoothing and a 20-day lag produced the highest correlation with the Gallup index and 30-day smoothing and a 50-day lag produced the highest correlation with the Michigan index). And while the correlation was reasonably strong for presidential approval ($r = .75$ with a 15-day smoothing interval), there was virtually no correlation between sentiment and preelection support for the candidates ($r = -.08$) with 7-day smoothing. Thus, while the results are mixed, there is indeed some correspondence, and the correspondences that are observed seem to depend on how the data are adjusted (i.e., smoothed and lagged).

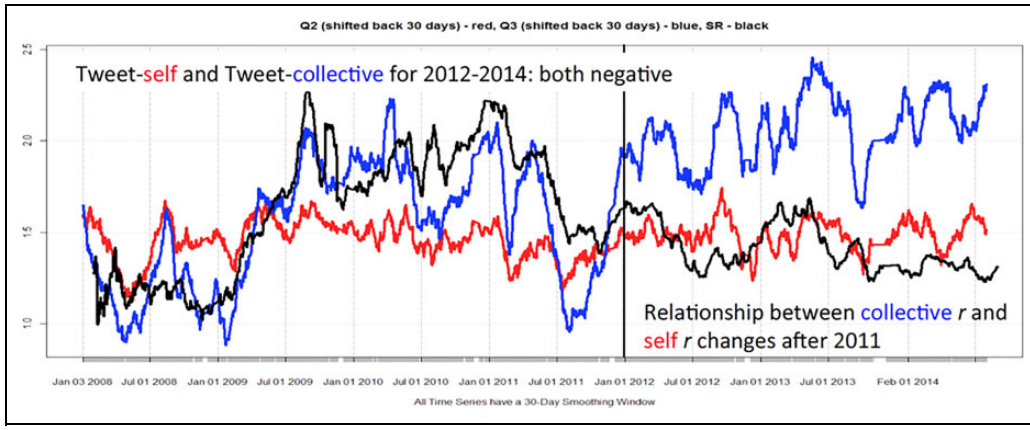

Figure 1. Sentiment in tweets containing "jobs," responses to "self" question in SCA ("Now looking ahead—do you think that a year from now you (and your family living there) will be better off financially, or worse off, or just about the same as now?") and "collective" question in SCA ("Now turning to business conditions in the country as a whole—do you think that during the next twelve months we'll have good times financially, or bad times, or what?"). Before 2012, the correlation between the self-question and Twitter sentiment was $r = .39$ and between the collective question and Twitter sentiment was $r = .84$; after 2012, both correlations were small and negative.

Motivation for Current Study

A number of other studies reported in the literature have compared social media content to survey data or related measures and found some correspondence (e.g., Antenucci, Cafarella, Levenstein, Ré, & Shapiro, 2014; Ceron, Curini, Iacus, & Porro, 2014; Fu & Chan, 2013; Jensen & Anstead, 2013; Pasek, Yan, Conrad, Newport, & Marken, 2018; Tumasjan et al., 2010). But not all of these initial success stories have held up. For example, in our own work (Conrad et al., 2015), the relationship between sentiment of tweets containing "jobs" and the Michigan ICS reported by O'Connor et al. (2010) was replicated through 2011, and, in fact, that study increased the correlation by focusing on individual survey questions from among the five questions on which the ICS rather than the global ICS measure. However, in subsequent analyses that included data collected after 2011, the relationship degraded rapidly, becoming small and negative (see Figure 1). Pasek, Yan, Conrad, Newport, and Marken (2018) also examined the relationship between sentiment of "jobs" tweets and the Gallup and Michigan indexes. They observed the largest correlations prior to 2012 and more negative correlations afterward.

Finally, Antenucci, Cafarella, Levenstein, Ré, and Shapiro (2014) successfully predicted U.S. unemployment, measured by initial claims for unemployment insurance, by computing a "job loss" index based on the frequency of tweets containing words and phrases such as "fired," "axed," and "canned". Between 2011 and 2014, the job loss index and unemployment insurance claims tracked one another closely. However, starting in mid-2014, the predictions and actual claims began to diverge and have not returned to previous levels of agreement (see http://econprediction.eecs.umich.edu/).

This pattern of relatively strong relationships over early years followed by highly attenuated relationships in more recent years raises serious questions about the viability of using social media content in place of survey data. In the current article, we investigate why this relationship might have weakened, whether it might be restored through different analytical methods, or—to foreshadow some of our findings—whether it might have been spurious.

Method

Overview

Our investigation examines the relationship between responses to the ICS survey, based on data from five questions asked in the University of Michigan Survey of Consumers, and sentiment of tweets containing "jobs" between 2008 and 2014. We first attempt to reproduce the key findings of O'Connor et al. (2010) for tweets from 2008 to 2009, and we obtain nearly identical results. In fact, Twitter sentiment and the ICS correlate strongly from 2008 to 2011. However, when we extend the analysis to the years 2012–2014, the relationship disappears. Our goals are to (1) learn why the relationship disappears in the later years and, if possible, to recover it and (2) introduce methods to help evaluate the believability of these results.

To address the first goal, we identify, as exhaustively as possible, the many analytic decisions required to find relationships between sentiment of tweets and survey responses. One such decision is filtering and categorizing tweets. "Jobs" tweets unrelated to employment presumably introduce noise and probably also introduce bias. Other decisions include how sentiment is calculated, the size of the smoothing interval, and lag between the date of survey responses and tweets. Finally, we consider the measure of association between Twitter sentiment and survey responses: Correlations of time series are often misleadingly large and also sensitive to outliers, so we consider an alternative measure that captures how often Twitter sentiment and consumer confidence move in the same direction.

To address the second goal, we first evaluate the stability of the observed relationships across changes to analytic approaches: We make small adjustments to the sentiment calculation, smoothing, and lag and observe how our results change. If the relationship changes substantially when we, for example, change smoothing and lag by only a few days, we would conclude that the relationship might be spurious. As an additional check on the credibility of a result, we repeat the analysis on "jobs" tweets that are relevant to employment and irrelevant to employment and compare the results.

Data Sources

Twitter data. Tweets from January 1, 2007 through June 27, 2014 containing the word "jobs" were collected from Topsy, which was purchased by Apple in 2015 and no longer provides this service. Topsy removed spam tweets, but the exact method of detecting spam tweets is not publicly available. There were many days with only a handful of "jobs" tweets in 2007, so we omit tweets from 2007 from the analysis. In order to reduce computational burden, we use a random sample of 500 tweets per day, although the number of "jobs" tweets is far greater than 500.

To reduce noise from the daily Twitter sentiment (see "Dictionary-based" scoring and "Machine learning–based" scoring subsections), we smooth these data over different temporal intervals, for example 30 days or 50 days, using the same approach used by O'Connor et al. (2010), giving a moving average (see "Correlation" subsection for more detailed description). Note that in analyses that examine O'Connor's findings, we use data from 2008 to 2009; in other analyses, to investigate whether the early results persist in later years, we include data through 2014. We smooth data when replicating the conditions used in O'Connor et al. (2010), but in some analyses, such as those involving comovement, we do not smooth the data because comovement produces a single sentiment score for the entire month.

Survey data. The survey data come from responses to five questions asked in the University of Michigan Surveys of Consumers (SCA). Each month, the SCA conducts approximately 500 (mostly) telephone interviews with a national sample of U.S. adults. The interviewers ask a series of questions to elicit respondents' attitudes about economic and business conditions—their personal

circumstances and the national situation. Responses to five of the questions are used to compute the ICS. The five questions are presented in Online Appendix A.

Many studies including O'Connor et al. (2010) use the final ICS, released *monthly* by the University of Michigan. Instead, we use *daily* aggregate consumer confidence responses. Having access to daily survey responses allows us to add lag and smoothing at whichever level we choose, discussed in greater detail in the "Measures of association" subsection.

Classifying Tweets

From discussing one's own job to mentioning Steve Jobs to posting about jobs of a sexual nature, the meaning of the word "jobs" varies considerably in our corpus. For our purposes, tweets about Steve Jobs and jobs of a sexual nature are irrelevant; there is no particular reason they should reflect people's thoughts and feelings on employment. Even tweets that are related to employment may capture sentiment about substantially different topics, for example, users discussing their own jobs versus users voicing an opinion about the government's role in job creation.

We created a classification system consisting of five broad categories of "jobs" tweets. The categories are *news/politics*, *personal*, *advertisement*, *irrelevant*, and *other*. Each of these categories is described below. Examples of actual tweets that belong in each category appear in Online Appendix C.

1. *News/politics*: This type of tweet generally refers to either current events on the national level or political opinions. Many of these tweets have to do with the U.S. economy as a whole.
2. *Personal*: Tweets in this category refer to an individual's job, often commenting on job satisfaction or change in employment status.
3. *Advertisements*: Tweets in this category display available jobs in various fields and cities. Advertisements consist mostly of tweets from third-party services such as Tweet My Jobs, the online job posting site Indeed, and so on.
4. *Irrelevant*: The jobs mentioned in these tweets are unrelated to employment or the economy. The most common "jobs" references in irrelevant tweets concern Steve Jobs (and biographical movies about him), the TV show *Dirty Jobs*, and jobs of a sexual nature.
5. *Other*: Tweets in this category are usually links to articles or lists posted online. These tweets are generally unrelated to recent economic events but not necessarily unrelated to the state of the economy. For example, more articles may be written about recession-proof jobs during a recession.

These five categories are not necessarily distinct; not every tweet fits unambiguously into one category. However, for simplicity, we assign each tweet to only one category.

We created an algorithm to automatically sort tweets into these categories; details are in Online Appendix B. To verify the accuracy of the algorithm, we randomly sampled 500 tweets and hand-classified them into one of the above five categories. Online Appendix Table B1 compares the hand classification to the classification as given by the algorithm. About 75% of these tweets were classified correctly (Cohen's $\kappa = .67$). The most difficult category for the algorithm was *other*. If we remove the *other* category, the algorithm accuracy increases to 89% (Cohen's $\kappa = .83$).

The exact proportion of "jobs" tweets in each category varies from year to year (see Online Appendix Figure B1), but on average, 8% of all tweets were about news or politics, 28% were *personal*, 27% were *advertisements*, 12% were *irrelevant*, and 24% concerned other material.

Twitter Sentiment

Although the survey responses directly measure sentiment, tweets do not. Thus, sentiment of tweets needs to be scored by analyzing the tone of words comprising the tweets. We use two broad strategies, *aggregate scoring* and *individual scoring*.

Aggregate sentiment scores are not meaningful at the tweet level but only meaningful when we examine many tweets across a given time period. Aggregate methods often use a dictionary of words, each labeled as either positive or negative. We check whether each word in a tweet is contained in either the positive or negative dictionary. Because tweets contain relatively few words, many tweets contain no words in any particular dictionary. Even if a tweet contains one or two words found in a dictionary, it is difficult to assign such tweets a meaningful continuous sentiment score. Instead, aggregate methods can sort such tweets into a sentiment categories, for example, positive tweet or negative tweet (or neither) depending on the number of positive and negative words it contains, or count the total number of positive and negative words in multiple tweets. The overall sentiment for a given day can then be calculated using either the number of positive and negative tweets from that day or number of positive and negative words from that day.

Individual scoring methods, on the other hand, assign a continuous sentiment score to each individual tweet. Individual scores are designed to be meaningful at the tweet level; they indicate not just whether a tweet is positive or negative but to what degree. Scores are assigned using lexical features of tweets and rule-based models. Tweets without any words in the dictionary were not assigned a sentiment score and only figured into the calculation of daily sentiment when the particular formula included total number of tweets, irrespective of sentiment. To calculate the overall sentiment for a given day, we use the mean sentiment score of all tweets from that day.

Dictionary-based scoring. Aggregate scoring methods are typically dictionary-based and are the most straightforward method of sentiment analysis. Each word in a text is compared to dictionaries of positive or negative words. The dictionaries are typically built on existing dictionaries (e.g., Roget's Thesaurus) and updated on the basis of human coders' judgments about the polarity of additional words in texts from a particular domain. Numerous limitations have been pointed out concerning sentiment analysis; for example, a word can be both positive and negative, the approach is not well equipped to detect sarcasm or irony, and the entries are typically single words which are insensitive to negation and the ways meaning and tone can be changed by adjacent words—e.g., the meaning of "lie" is quite different when considered in isolation than when followed by "down." Several dictionaries have been improved to address certain limitations, especially challenges resulting from negation (e.g., Young and Siroka, 2012).

Positive and negative sentiment in tweets can be quantified in one of (at least) two ways. The first is counting the total number of positive and negative words in "jobs" tweets from that day. The second, used by O'Connor et al. (2010), is counting the number of positive and negative tweets. Following this approach, if a tweet contains at least one positive word, it is considered a positive tweet, and if it contains at least one negative word, it is considered a negative tweet; a single tweet can thus be positive, negative, both positive and negative, or neither. Since tweets contain relatively few words, one would assume the difference between counting the number of positive and negative words versus the number of positive and negative tweets would be negligible. Here, we include both options to reproduce previous analyses and to see what effect a seemingly small change may have on the results.

Once we have the number of positives and negatives (either tweets or words) for a single day, overall sentiment for that day can be calculated in one of three ways: (1) $\frac{\text{positives}}{\text{negatives}}$, (2) $\frac{\text{positives} - \text{negatives}}{\text{total}}$, or (3) $\frac{\text{positives}}{\text{positives} + \text{negatives}}$. Considering these multiple metrics allows us to both reproduce previous analyses and to see whether a seemingly small change in sentiment calculation affects the results.

We use three dictionaries:

Lexicoder. Lexicoder (Young & Soroka, 2012) was developed for measuring tone in news content. It has been shown to perform well compared to manually coding newspaper articles regarding public policy and election campaigns. Lexicoder consists of 1,700 positive words and 2,857 negative words. One difference between Lexicoder and the other two dictionaries is the inclusion of negated words: for every positive and negative word (e.g., "happy," "sad"), there is an associated negated entry (e.g., "not happy," "not sad"). The final positive dictionary consists of the positive words and negated negative words, and the negative dictionary consists of the negative words and negated positive words. The Lexicoder dictionary is available for download (http://lexicoder.com/) and through the quanteda package in R (Benoit, 2018).

Liu–Hu. The Liu–Hu dictionary (Liu, Hu, & Cheng, 2005; Hu & Liu, 2004) was created to assess customer sentiment contained in online product reviews, where it is common to mention both positive and negative features of products. Because the data come from customers' reviews, the list includes common misspellings and therefore might be better suited to analyzing sentiment of opinions expressed on Twitter than dictionaries developed from more professionally created texts. The word lists consist of 4,783 negative words and 2,006 positive words. The Liu–Hu dictionary is available through the Sentiment Analysis package in R (Feuerriegel & Proellochs, 2018).

OpinionFinder. This dictionary was developed to evaluate a theory of polarity in lexical semantics. The OpinionFinder word lists (Wilson, Wiebe, & Hoffmann, 2005) consist of 1,600 positive and 1,200 negative words. Similar to Lexicoder, these lists do not contain slang or misspellings. O'Connor et al. (2010) used OpinionFinder.

Machine learning–based scoring. Individual scoring methods typically make use of machine learning algorithms. Rules for scoring sentiment of individual tweets are created via a machine learning algorithm trained on a corpus of texts whose sentiments were hand-scored. The machine learning–based methods construct lexicons that contain content words, function words, for example, negations and intensifiers like "very," and nonword entries, for example, emojis and punctuation. Each entry in the lexicon has either an associated sentiment and intensity score or an associated rule. We include two machine learning-based methods in our study.

Vader (Hutto & Gilbert, 2014) assigns each individual tweet a sentiment score between -1 and 1, taking into account text features commonly found in short social media messages such as words, slang, negations, intensifiers and punctuation (e.g., exclamation points and capitalization), and emoticons. As Hutto and Gilbert demonstrate, Vader performs well across various contexts of social media messages. Note that Vader was trained on tweets.

TextBlob (Loria, 2018) sentiment analysis is a Naive Bayes classifier trained on the Stanford NLTK data set of movie reviews. TextBlob outputs a sentiment score from -1 to 1 for each individual tweet. Similar to Vader, TextBlob incorporates negations and intensifiers when calculating sentiment of a tweet.

Measures of Association

We measure the relationship between Twitter sentiment and consumer confidence in two ways: (1) correlation and (2) comovement, which is a measure of how often both time series move in the same direction.

Correlation. Pearson's correlation is commonly used for assessing relationships between survey responses and Twitter sentiment (both O'Connor, Balasubramanyan, Routledge, & Smith, 2010 and Conrad et al., 2015 used Pearson's correlation for this purpose). There are many commonly known strengths and weaknesses of correlation. For example, one particular weakness of correlation is sensitivity to outliers. One feature of correlation, both a strength and a weakness when assessing the similarity of two time series, is the ability to capture similarity in long-term trends. For example, if two time series both exhibit a long-term increasing trend, they will be highly correlated. Such examples can easily occur spuriously, however, and in fact, spurious correlations are common in time series data (http://www.tylervigen.com/spurious-correlations).

Before computing the correlation between Twitter sentiment and survey responses, we first smooth both Twitter sentiment and survey responses. Each day in our data has a Twitter sentiment score (calculated using one of the various methods described above) and an ICS score. These are both very noisy day-to-day. Similar to O'Connor et al. (2010), we calculate the smoothed daily sentiment and ICS score by taking the average of the current and previous $K - 1$ days. The same K value is used for both time series. We add a shift (or lag) of L days to the survey responses, which indicates by how many days Twitter sentiment leads or lags survey responses. A positive L means Twitter sentiment lags survey responses, and a negative L means Twitter sentiment leads survey responses. We then find the correlation between the smoothed Twitter sentiment and smoothed and lagged survey responses. In other words, we compute the cross-correlation between smoothed Twitter sentiment and smoothed survey responses.

Comovement. The second measure of association we use is comovement, which measures how often two time series move in the same direction from one time period to the next. While correlation uses the actual values of each time series, comovement uses the direction of the differences. The notion of comovement dates back to the late 1800s (Fechner, 1897), with further developments by Moore and Wallis (1943) and Goodman and Grunfeld (1961). More precisely, if we have T time units and two time series $x_1, x_2, \ldots, x_T$ and $y_1, y_2, \ldots y_T$,

$$comovement(x, y) = \frac{1}{T-1}\sum_{t=2}^{T} 1[\text{sgn}\,(x_t - x_{t-1}) = \text{sgn}(y_t - y_{t-1})]$$

where $1[\text{sgn}(x_t - x_{t-1}) = \text{sgn}(y_t - y_{t-1})]$ is 1 if x and y move in the same direction from time period $t - 1$ to t and 0 if x and y move in opposite directions from time period $t - 1$ to t.

Comovement has several advantages over correlation. First, because comovement only deals with differences, it is not overly sensitive to long-term trends in the data. If two unrelated time series happen to have similar long-term trends, comovement will not necessarily indicate a strong relationship between them. Secondly, because comovement uses only the direction of the differences and not the magnitudes, it is more robust to outliers. Lastly, comovement is easily interpretable. For example, if the comovement between two time series is 0.9, then the two time series move in the same direction from one time period to the next 90% of the time.

After shifting survey responses by L days, we calculate comovement on various timescales: daily, weekly, or monthly. Note that because the comovement value measures the percent of time two time series move in the same direction from one time period to the next, each time series can only have one value per time period. That is, depending on the time unit chosen, for each time series, we have one value per day, one value per week, or one value per month.

We have some freedom in how exactly to calculate weekly and monthly sentiment by choosing which day of the week or month to start on. For example, in calculating weekly sentiment, we can start the week on Sunday, meaning sentiment for a week is the average sentiment from Sunday through the following Saturday, or we can choose to start the week on Monday, meaning sentiment for a week is

Table 1. Correlations Between Sentiment of Tweet Categories and Index of Consumer Sentiment by How Daily Sentiment Is Calculated, Based on O'Connor, Balasubramanyan, Routledge, and Smith (2010). Tweets Are From 2008 to 2009, Using 30-Day Smoothing and 50-Day Lag.

Category of Tweet	$\frac{\text{positive tweets}}{\text{negative tweets}}$	$\frac{\text{positive tweets} - \text{negative tweets}}{\text{total tweets}}$	$\frac{\text{positive tweets}}{\text{positive tweets} + \text{negative tweets}}$
All tweets	.65	.00	.48
News/politics	.17	.30	.19
Personal	−.23	−.30	−.26
Advertisements	.71	−.24	.32
Irrelevant	.42	.16	.32
Other	.19	.43	.52

the average sentiment from Monday through the following Sunday, and so on. As a robustness check, we therefore compute comovement starting on various days and compare the results.

Results

Replication of O'Connor et al.

We begin by replicating the analysis from O'Connor et al. (2010), in which sentiment of "jobs" tweets from 2008 to 2009 is compared to consumer confidence as measured by ICS. We use the same settings and time frame. The differences in our analysis are (1) a different corpus of "jobs" tweets (O'Connor et al., 2010, obtained their tweets from the Twitter API; our corpus is from Topsy) and (2) ICS is computed daily in our study but monthly for O'Connor et al. (2010). Following O'Connor et al. (2010), we calculate daily sentiment as the ratio of positive to negative tweets, using the OpinionFinder dictionary. As in O'Connor et al. (2010), Twitter sentiment is smoothed by $K = 30$ days and shifted by $L = -50$ days. O'Connor et al. (2010) find a correlation of .64 and we find one of .65 (see top left cell of Table 1), suggesting that our replication succeeded.

Sorting by Twitter Category

We calculate sentiment for the tweets assigned to each of the five content categories and, using the same settings as above, find the correlation between each of the categories and ICS. Results can be seen in column 1 of Table 1. We expected the sentiment of tweets from the *irrelevant* and *advertisements* categories to have the lowest correlations with survey responses and *news/politics* to have the highest but find the opposite. Correlation with *advertisements* (.71) and *irrelevant* (.42) is far higher than with *news/politics* (.17). The correlations with *personal* (−.23) and *other* (.19) are also not particularly strong. We assumed *advertisements* and *irrelevant* tweets would be unrelated to employment, so these two high correlations may well be spurious.

Robustness of Results

With many researcher decisions (e.g., choice of dictionary, the determination to count words vs. tweets, the particular smoothing interval chosen) contributing to the resulting correlation of .65 above, we are interested in how these decisions might affect the outcome. To assess this, we adjust these analysis parameters and compare the resulting correlations.

We first adjust the method used to calculate sentiment, using the three formulas given in the "Dictionary-based scoring" subsection. Results can be seen in Table 1. We find that the choice of

Table 2. Correlations With Index of Consumer Sentiment, Using Tweets From 2008 to 2009, 30-Day Smoothing, and 50-Day Lag, Based on O'Connor, Balasubramanyan, Routledge, and Smith (2010).

Category of Tweet	OpinionFinder		Lexicoder		Liu–Hu	
	positive tweets / negative tweets	positive words / negative words	positive tweets / negative tweets	positive words / negative words	positive tweets / negative tweets	positive words / negative words
All tweets	.65	.64	.56	.56	.66	.61
News/politics	.17	.10	.30	.39	.15	.18
Personal	−.23	−.25	−.07	.02	.07	.11
Advertisements	.71	.67	.29	.19	.57	.53
Irrelevant	.19	.19	.51	.48	.28	.30
Other	.42	.33	.56	.36	.49	.43

Table 3. Correlations Between Sentiment of Tweets From Vader and TextBlob and Index of Consumer Sentiment for 2008–2009.

Category of Tweet	Vader	TextBlob
All tweets	.54	.18
News/politics	.51	.10
Personal	−.05	.22
Advertisements	−.24	−.39
Irrelevant	.45	.05
Other	.64	.60

scoring formula can drastically change the results. Most striking is the change in correlation with all tweets, dropping from 0.65 to 0.00 when we calculate sentiment as $\frac{\text{positive tweets} - \text{negative tweets}}{\text{total tweets}}$. Had O'Connor et al. (2010) used this scoring formula, they would have reached a different conclusion: Instead of a fairly strong relationship, there would have been no relationship. While this is a dramatic change in results, the correlation with *advertisements* changes even more, from a strong positive correlation (0.71) to moderately negative (−0.24) and moderately positive (0.32) depending on the formula used. Correlations with *news/politics* and *personal* tweets remained relatively constant (and small).

We next explore the choice of dictionary and the difference between counting tweets versus words. Results can be seen in Table 2. We see these two decisions do not have a dramatic effect on the correlation with all tweets, with the correlations hovering around 0.6. As above, the most dramatic effect of the dictionary choice is seen for *advertisements*, with the correlations ranging from 0.71 when counting tweets with OpinionFinder to 0.19 when counting words in Lexicoder. In general, counting words versus tweets did not make a large difference in correlation. The largest differences in counting words versus tweets occurred when using the Lexicoder dictionary, where the correlation with sentiment from irrelevant material changed from 0.56 (when counting tweets) to 0.36 (when counting words).

Table 3 presents the correlations between sentiment of tweets based on Vader and TextBlob and ICS for 2008–2009. While both Vader and TextBlob are machine learning–based methods, they nonetheless differ from each other, and results based on their scores also differ. Correlation between all tweets and ICS is 0.54 using Vader and 0.18 using TextBlob, correlation with *news/politics* tweets is 0.51 using Vader and 0.10 using TextBlob, and correlation with *irrelevant* tweets is 0.45 with Vader and 0.05 with TextBlob. The difference in the size of these correlations presumably reflects disagreement in the actual sentiment scores that Vader and TextBlob assigned to some

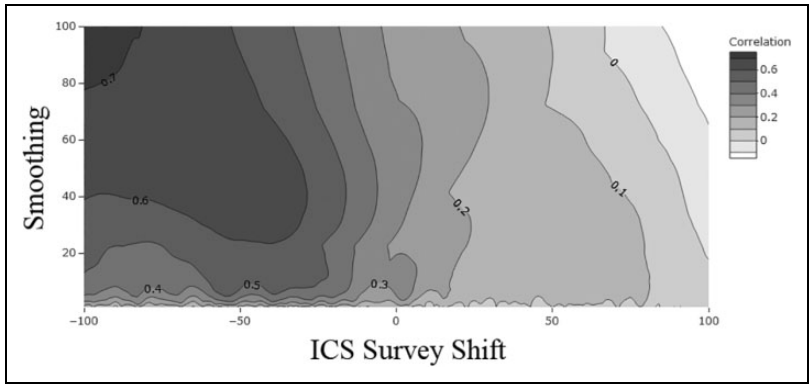

Figure 2. Effect of smoothing and lag parameters on correlation between all tweets from 2008 to 2009 and ICS. Sentiment calculated as $\frac{\#\text{ positive tweets}}{\#\text{ negative tweets}}$ using the OpinionFinder dictionary. Contour lines reflect correlation boundaries.

tweets. This is evident in the modest correlation between Vader's and TextBlob's sentiment scores, $r = .54$, for 2008–2009, suggesting that using Vader versus TextBlob can potentially lead to quite different results. The correlations between all five sentiment scoring tools can be found in Online Appendix D for 2008–2009 (Online Appendix Table D1), the years for which O'Connor et al. (2010) reported the relationship, and 2008–2014 (Online Appendix Table D2), the years for which we explored the extensibility of the O'Connor et al. (2010) results over time. Many of the correlations between the tools are modest or low, suggesting that sentiment scoring tools—at least these five—are not interchangeable.

Lastly, we compare results with different levels of smoothing and lag. Using the original conditions as in O'Connor et al. (2010), we can see how the correlation between all tweets and survey responses changes as we adjust smoothing from $K = 1$ to 100 and the shift from $L = -100$ to 100 days. The resulting contour map is displayed in Figure 2. In general, under these settings, correlation increases as L changes from positive to negative, becoming largest (darkest regions) when the lag is most negative (social media precedes survey data by 100 days) and the smoothing interval is largest (upper left area). However, we have no theoretical explanation for why such a large lag leads to a stronger correlation under the given settings. In particular, we would expect daily Twitter sentiment and daily ICS to be more or less aligned. While it may be the case that Twitter users form opinions somewhat faster (or slower) than the general population, we would expect this difference to be on the order of days, not months; the fact that we see the largest correlations with an L on the order of months suggests to us that one should not read too much into these correlations as they may well be spurious.

We note also that while correlation tends to increase as smoothing increases, high levels of smoothing can artificially inflate correlation between two time series.

Comovement

We compute comovement using the same settings as in O'Connor et al. (2010). Results appear in Table 4. There are no strong relationships between any Twitter categories and survey responses at the daily or weekly levels, all hovering around 0.5 (what we would expect by chance). When calculating monthly comovement using 2 years of data, there are only 23 monthly differences. With so few data points, it is easier to obtain more extreme comovements due to chance. At the monthly level, comovement varies depending on which day of the month we start on. If there is a genuine

Table 4. Comovement Between Sentiment of Tweets and Index of Consumer Sentiment from 2008 to 2009. Comovement Calculated Daily, Weekly, and Monthly Starting on the First, Monthly Starting on the Second, and Monthly Starting on the Fourth Day of the Month.

Category of Tweet	Daily	Weekly	Monthly First	Monthly Second	Monthly Fourth
All tweets	.47	.52	.70	.61	.65
News/Politics	.46	.53	.39	.35	.35
Personal	.46	.47	.65	.65	.61
Advertisements	.43	.40	.48	.52	.57
Irrelevant	.52	.54	.57	.70	.57
Other	.50	.46	.39	.43	.70

relationship between survey responses and Twitter sentiment, we would expect comovement to be not only large but also robust to starting date. However, we find different results depending on the starting day. Starting the month on the first results in a comovement of 0.70, but moving the start day by just one day to the second results in a comovement of 0.61. The most striking example is seen for *other*, jumping from 0.39 and 0.43 starting on the first and second, respectively, to 0.70 when starting on the fourth. As another example, *irrelevant* tweets jump from 0.57 to 0.70 to 0.57 when starting on the first, second, and fourth, respectively.

Finally, note that in the cases of *news/politics* and *personal* tweets, the two categories we most expected to be associated with survey responses, the comovement was not particularly large for any start date.

Extension in Time

When we began this study, our motivation was to understand why the relationship between Twitter sentiment and survey responses deteriorated over time. This raises the question of whether the relationship actually does weaken over time or simply started and remained volatile. To examine this, we compute the correlation for each year from 2008 through mid-2014 under the settings originally used by O'Connor et al. (2010) for 2008 through 2009. These correlations are displayed in Table 5. In some years, there is a very high correlation, which disappears or moves in the opposite direction the following year. Additional results for Vader and TextBlob appear in Tables 5, and the patterns are relatively similar. Furthermore, there is no discernible pattern throughout. In particular, correlations do not slowly deteriorate over the years. If this had been the case, it could have suggested some systematic change in the Twitter data. Instead, the evidence suggests there never was a relationship to begin with.

Personal Versus Collective Hypotheses

Conrad et al. (2015) hypothesized that comparing Twitter sentiment to responses for individual survey questions—as opposed to the overall ICS—might strengthen the relationship. More specifically, they reasoned that if people create tweets for others to read, like, and retweet, the sentiment might be more similar to answers to questions about the national economy (collective) than about one's personal financial circumstances (self). In fact, they observed a higher correlation with a collective question concerning "business conditions in the country as a whole" (Q3 in Online Appendix A), $r = .84$, than with a self-question concerning the financial well-being of "you and your family" (Q2 in Online Appendix A), $r = .39$. The correlations were measured over the years 2008–2011 so entailed the years originally investigated by O'Connor et al. (2010), that is, 2008–2009, as well as the following two years.

Table 5. Correlations Between Index of Consumer Sentiment and Twitter Categories by Year for 2008–2014 Under Settings Used Originally by O'Connor, Balasubramanyan, Routledge, and Smith (2010) for 2008–2009.

Category of Tweet	2008	2009	2010	2011	2012	2013	2014
All tweets	.21	.66	−.03	.54	.02	.28	.41
News/politics	−.05	.18	.22	.37	−.02	.02	−.61
Personal	−.10	.36	.08	.23	−.07	.09	−.24
Advertisements	−.02	.64	.01	.59	−.17	.29	.84
Irrelevant	.06	.29	−.21	−.21	−.16	−.16	.16
Other	−.38	.46	−.57	.67	.02	.53	−.25
Correlations Between Index of Consumer Sentiment and Twitter Categories by Year Using Vader							
	2008	2009	2010	2011	2012	2013	2014
All tweets	−.18	.71	.21	.62	−.04	−.10	.35
News/politics	.21	.45	.75	.47	.20	−.15	.48
Personal	−.32	.11	.49	.44	.09	−.21	.34
Advertisements	−.14	−.69	.28	−.07	−.37	.20	−.74
Irrelevant	.16	.51	.52	−.31	−.32	−.16	−.71
Other	−.05	.78	−.49	.76	.03	.17	−.33
Correlations Between Index of Consumer Sentiment and Twitter Categories by Year Using TextBlob							
	2008	2009	2010	2011	2012	2013	2014
All tweets	−.45	.39	−.44	.37	.23	−.21	.07
News/politics	−.10	.28	.40	.16	.33	−.41	.40
Personal	−.24	.18	.05	.28	.02	−.20	.27
Advertisements	−.25	−.41	−.23	−.21	−.06	.13	−.46
Irrelevant	−.13	.01	−.24	−.51	.18	−.10	−.39
Other	.41	.45	−.43	.59	.27	.07	.61

Here, we revisit the main hypothesis proposed by Conrad et al. (2015) over a larger time period (2008–2014) and by correlating survey responses with particular categories of tweets whose content may be relevant to the two survey questions they examined. In particular, we correlate responses to the collective questions with *news/politics* jobs tweets and the responses to the self-question with *personal* "jobs" tweets. Similar to Conrad et al., we calculate sentiment as $\frac{\#\text{positive words}}{\#\text{negative words}}$, where positive and negative words are defined by the Lexicoder dictionary, with 30-day smoothing and lag of 50 days as in O'Connor et al. (2010). Table 6 presents the correlations between each category and the collective-and self-survey questions. As with the overall ICS by year (top row), there is no clear pattern. Tweets about the U.S. economy (*news/politics*) were no more related to survey responses about the direction of the national economy than any other category of tweets, and tweets about one's own job (*personal*) were no more related to survey responses about one's personal finances than any other category of tweets.

Discussion

We began our investigation hoping to explain why the previously observed relationship between Twitter sentiment and survey responses (e.g., O'Connor et al., 2010) disappeared when more recent data were included in the analyses (Conrad et al., 2015; Daas, Puts, Buelens, & van den Hurk, 2015). We conducted a series of analyses to help explain this loss of relationship. Despite our initial

Table 6. Correlations Between Twitter Categories and Collective Question by Year.

Category of Tweet	2008	2009	2010	2011	2012	2013	2014
All tweets	.18	.84	.24	.44	.18	.07	.70
News/politics	.16	.25	.68	.50	.21	−.41	.80
Personal	.44	−.14	.60	.33	.35	−.21	−.19
Advertisements	.06	.65	.39	.20	.15	.12	−.66
Irrelevant	.08	.62	.34	.03	−.38	.14	.16
Other	−.30	.72	−.53	.60	.30	.14	.23
Correlations Between Twitter Categories and Self-Question by Year							
	2008	2009	2010	2011	2012	2013	2014
All tweets	.19	.52	.27	.02	.03	.13	−.36
News/politics	.17	−.16	.59	.28	.14	−.14	−.53
Personal	.15	.20	.37	−.03	.08	−.17	−.15
Advertisements	.01	.52	.42	−.02	−.08	.01	.35
Irrelevant	−.08	.41	.40	.32	−.31	.14	.37
Other	−.61	.29	.33	.07	.29	.21	−.04

optimism that the right classification of the tweets, the right sentiment dictionary, or a more robust measure of association would restore the relationship, our findings have ultimately cast doubt on the validity of the initially reported relationship. While not the outcome we anticipated, these results have helped expose inherent obstacles to fruitfully using social media data for social research. Below we summarize the challenges we encountered and discuss possible approaches to address them.

Summary of Results

Initially, we believed that the signal apparent in the early analyses was simply obscured in the more recent data, perhaps by a larger proportion of *advertisements* or *irrelevant* tweets containing "jobs." To test this idea, we employed a content-based classification process to isolate "jobs" tweets relevant to economic opinion. However, this did not recover the relationship, and what we actually found was that the highest correlations emerged for tweets we classified as *advertisements* and *irrelevant*, which we had no reason to expect would correlate with the survey results. It is conceivable that advertisement tweets might have correlated highly with consumer sentiment, if consumer sentiment reflects in part the number of job opportunities at any moment, which sentiment in job ads could also reflect. The fact that the correlations for advertisement tweets vary to such a great extent with different analytic strategies suggests this is unlikely to be the case. The correlations for tweets we classified as *news/politics* and *personal*—which we did expect to correlate with the survey data—were both *lower* than those of the *irrelevant* and *advertisement* tweets, casting further doubt on the credibility of the original results.

Second, we hypothesized that, as Twitter became more mainstream, the writing style in tweets may have evolved and that the relationship could be recovered by using more Twitter-specific methods to score the sentiment of tweets. Thus, we compared results across traditional dictionary methods with more sophisticated machine learning–based tools developed for Twitter and similar texts. Different methods for calculating sentiment often produced vastly different results (see Table 1). However, using these methods, we were unable to recover the original relationship.

We next considered the possibility that the signal was in fact present in the Twitter data but that we were using an improper measure of association to relate the Twitter and survey data; in particular, Pearson correlation could be a poor measure of association to use in this context. We considered instead comovement, a nonparametric measure of the strength of a relationship of two time series. However, when using this metric, we found no strong associations.

When we were unable to restore the underlying relationship, we considered the possibility that the initial findings, despite appearing quite strong, might have been spurious. Our strategy at this point was to see how easily spurious relationships might be produced within this framework, and we focused on the many micro-decisions that these analyses require, such as whether sentiment is computed per tweet or per day, the size of the smoothing interval, and the length of the lag. We found that what seem like minor decisions can have major impacts on the outcomes. Moreover, we found that it is not difficult to create relatively large correlations by arbitrarily adjusting these parameters, leading to deceptively encouraging results.

We have found in particular that the ordinary Pearson correlation, as a measure of the association of Twitter sentiment and survey responses, is especially prone to produce such deceptively encouraging results. The primary reasons for this are that both the Twitter time series and the survey time series individually exhibit autocorrelation (reducing effective sample size and increasing chance error) and that both the Twitter time series and the survey time series individually exhibit overall trends (increasing or decreasing); this alone may induce a correlation. Indeed, spurious correlations between time series are common. Although we considered comovement, an alternative metric of relatedness which is much less sensitive to background trends (as well as outliers) than Pearson correlation, we were unable to find any evidence with this measure that the Twitter and survey time series were related.

On balance, we now believe that what appeared to be a reasonably strong relationship between the tweets and survey responses in the early years of Twitter was likely spurious and efforts—including work by some of the current authors—to extend those results in time are equally likely to have produced spurious results. In the current study, we have asked many questions of these data, and none has produced evidence that there really is a signal. It is of course possible that additional analyses might detect a signal, but it is not clear to us what those additional analyses might be. Rather than investing effort in further analyses, it seems appropriate instead to take a hard look at the general approach. In particular, we feel it is necessary to evaluate the plausibility of key assumptions required for this approach to succeed.

Challenges of Future Work

There are two major questions that, in our view, require investigation in order to move forward. First, as others have noted (e.g., Duggan, Ellison, Lampe, Lenhart, & Madden, 2015; Graham, Hale, & Gaffney, 2014; Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011), the users who create social media and other types of found or organic data (Groves, 2011) are unlikely to represent the population of interest to social researchers. But it is unclear just how nonrepresentative these data are, which is an empirical question.

The sentiment expressed in any Twitter corpus represents the sentiment of Twitter users at a particular point in time but cannot be assumed to represent the mood in any other population (Baker, 2017, p. 59). If the research goal is to generalize the sentiment in a Twitter corpus to a national population, for example, the U.S. adult population, then it is worth seeing whether the methods that survey researchers have developed to produce population estimates from nonprobability samples of survey respondents can be applied to Twitter content. Of course, survey researchers have found that nonprobability samples are often biased (e.g., Tourangeau, Conrad, & Couper, 2013), and adjustments can make matters worse (Baker et al., 2013). Nonetheless, an appropriate next step is to

explore the potential benefits of treating a Twitter corpus as a nonprobability sample (cf., Diaz, Gamon, Hofman, Kıcıman, & Rothschild, 2016; Pasek et al., 2018).

Most efforts to derive population estimates from nonprobability surveys involve reweighting the data. The problem with adapting this general method to tweets is that they generally lack the kind of covariates required for reweighting, primarily demographic characteristics of the users who have posted particular content. One approach is to infer the users' characteristics based on the content of their posts and associated metadata (such as geotags). The kinds of characteristics that have been extracted from social media content include location (Ajao, Hong, & Liu, 2015; Jurgens, Finethy, McCorriston, Xu, & Ruths, 2015; Schulz, Hadjakos, Paulheim, Nachtwey, & Muhlhauser, 2013), political affiliation (Barberá, 2014; Barberá, 2016; Cohen & Ruths, 2013), income (Preoţiuc-Pietro, Volkova, Lampos, Bachrach, & Aletras, 2015), and computer usage (Blank, 2017). These are early efforts and the results are mixed, but the approach may ultimately allow researchers to reweight according to inferred user characteristics with acceptable confidence (though see Freelon, 2019).

Alternatively, one might reweight the survey data to better match characteristics of the Twitter user base. While such an analysis does not address whether opinions of Twitter users can be extrapolated to the larger population, it can at least be used to test whether sentiment extracted from tweets can be reliably correlated with survey results from the Twitter population. Characteristics of Twitter users have been assessed in surveys (e.g., Greenwood, Perrin, & Duggan, 2016) which both ask a random sample of respondents if they use or have used Twitter and also measure their demographic information. The efforts that we are aware of that used this approach (Pasek & Dailey, 2019; Pasek et al., 2018) did not improve the relationship between Twitter sentiment and survey responses, but the approach seems promising to us.

Another approach is to conduct a survey of Twitter users who have tweeted on topics of interest to the researcher, asking questions about demographics and the study topics while simultaneously conducting a calibration survey, that is, a survey of a representative sample of the population to which the researchers wish to generalize the Twitter results. This approach should enable researchers to reweight the data extracted from the Twitter corpus based on the discrepancies between the demographics of the Twitter survey sample and the calibration sample. This approach assumes that the sample of Twitter users represents all Twitter users posting on the same topic(s) at that time. In addition to enabling reweighting of the Twitter data, this approach has the benefit of allowing a comparison of responses to questions on the substantive topic(s) of interest between the Twitter and representative samples and adjusting accordingly.

In retrospect, we should probably not be surprised about the lack of correspondence between survey responses and Twitter sentiment because, aside from the representational issues just mentioned, there are differences in how respondents and users create data in their respective tasks (see Schober et al., 2016). Survey researchers determine the topic about which respondents are asked while social media users themselves determine the content about which they post. Survey researchers present exactly the same stimulus (the question) to all respondents, so the data are created under relatively comparable circumstances. Social media users, in contrast, post content in response to unknown—but presumably highly varied—stimuli. For example, Naaman, Boase, and Lai (2010) classified 3,379 personal tweets, making it clear that users post about what is on their mind at the time: most prevalent was *Me now* (41%), for example, "tired and upset," followed by *Random Thoughts* (25%), for example, "I miss New York but I love LA . . . ," followed in turn by *Opinions/Complaints* (24%), for example, "Illmatic = greatest rap album ever." This could not be more different from the direct connection between survey questions and responses. When we look at the two sources of data in the current study from this perspective, we are unable to identify a reason why they *should* contain the same story or sentiment. But for some topics on some occasions, tweets and responses might overlap. For example, Pasek and Dailey (2019) found that the sentiment of "jobs" tweets is negatively correlated with volatility in the stock market. But in general, there is not sufficient evidence in the literature to know when such

overlap is likely to occur. Similarly, several studies have found credible relationships between Google search data and survey results (e.g., Baker & Fradkin, 2011; Choi & Varían, 2012; Jun, Yoo, & Choi, 2018), but the reversal of Google Flu Trends' initial success predicting U.S. government surveillance data is well known (e.g., Butler, D., 2013).

To the extent that survey respondents and Twitter users create data with particular audiences in mind, respondents consider a far narrower range of audiences than do Twitter users. Some survey respondents edit their answers to present themselves more favorably to the interviewer, but when respondents self-administer questions, as in online questionnaires, there is much less evidence of social desirability bias in their answers, (e.g., Tourangeau & Smith, 1996; Kreuter, Presser, & Tourangeau, 2008; Lind, Schober, Conrad, & Reichert, 2013) suggesting they do not design their responses for any. Twitter users, in contrast, seem to design their posts for multiple, specific audiences. Marwick and boyd (2011) asked users who they imagined they were tweeting to. The 226 responses they received indicated some users do not consider the audience when posting content ("I don't tweet to anybody; I just do it to do it"), but others described particular imagined audiences ("I think of a room filled with friends when I tweet. I assume people like me that are reading my tweets."). It was clear that users adjust content on the basis of their imagined audience ("i'm very conscious that twitter is public. i wouldn't tweet anything i didn't want my mother/employer/professor to see"). Again, it is possible that for some topics on some occasions, the differing role of imagined audiences has no impact on the comparability of the two data sources. But we are not aware of any research that might provide guidance on when this might be the case.

In short, it may be possible under some circumstances to capture sentiment from social media data that are genuinely associated with the sentiment of the entire population. But our concern is that, as Groves (2011) said about nonprobability surveys, "such designs work well until they don't; there is little theory undergirding their key features." Developing such theory, if it can be developed, should be a priority.

Data Availability

The average daily ICS from 2008 to 2014 as well as the average daily measures from the five SCA questions on which the ICS is based can be found at https://www.openicpsr.org/openicpsr/project/109581/version/V1/view/. The daily average sentiment scores for jobs tweets from 2008 to 2014 computed with five different tools can be found at https://www.openicpsr.org/openicpsr/project/109581/version/V1/view/

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 disclosed receipt of the following financial support for the research, authorship and/or publication of this article: Robyn Ferg was supported by the National Science Foundation under Grant No. 1646108.

Software Information

A script for a Shiny app that allows a user to assess the relationship between the sentiment data and survey responses, along with a script that allows the user to reproduce all results and figures reported in these analyses, can be found at https://github.com/robynferg/Twitter_ICS. R code for all results and figures can be found at https://github.com/robynferg/Twitter_ICS/blob/master/ResultsFigures.R

Supplemental Material

Supplementary material for this article is available online.

References

Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science*, *41*, 855–864.

Antenucci, D., Cafarella, M., Levenstein, M. C., Ré, C., & Shapiro, M. D. (2014). Using social media to measure labor market flows." Manuscript under review.

Baker, R. (2017). Big data: A survey research perspective. In P. Biemer, et al. (Eds.), *Total survey error in practice. Total survey error in practice* (pp. 47–69). Hoboken, NJ: Wiley.

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.

Baker, S., & Fradkin, A. (2011). *What drives job search? evidence from Google search data*. Technical report, Stanford University. Retrieved from https://econpapers.repec.org/paper/sipdpaper/10-020.htm

Barberá, P. (2014). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*, 76–91.

Barberá, P. (2016). *Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data*. Working paper. Retrieved from http://pablobarbera.com/static/barbera_polarization_APSA.pdf

Benoit, K. (2018). Quanteda: Quantitative analysis of textual data (R package). Retrieved from http://quanteda.io

Blank, G. (2017). The digital divide among Twitter users and its implications for social research. *Social Science Computer Review*, *35*, 679–697.

Butler, D. (2013). When Google got flu wrong. *Nature News*, *494*, 155–156.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (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*, 340–358.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2–9.

Cohen, R., & Ruths, D. (2013, June). Classifying political orientation on Twitter: It's not easy! In *International Conference on Web and Social Media*. Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Conrad, F. G., Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Hou, E. (May, 2015). *A "collective-vs-self" hypothesis for when Twitter and survey data tell the same story*. Paper presented at the annual conference of the American Association for Public Opinion Research, Hollywood, FL.

Daas, P. J., Puts, M. J., Buelens, B., & van den Hurk, P. A. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, *31*, 249–262.

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

Duggan, M., Ellison, N., Lampe, C., Lenhart, A., & Madden, M. (2015). *Social media update 2014*. Washington, DC: Pew Internet and American Life Project.

Fechner, G. T. (1897). *Kollektivmasslehre* [The science of measuring collectives]. Leipzig, Germany: Wilhelm Engelmann

Feuerriegel, S., & Proellochs, N. (2018). Sentiment analysis: Dictionary-based sentiment analysis. R package version 1.3-2. Retrieved from https://CRAN.R-project.org/package=SentimentAnalysis

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* (pp. 96–110). New York, NY: Routledge.

Fu, K. W., & Chan, C. H. (2013). Analyzing online sentiment to predict telephone poll results. *Cyberpsychology, Behavior, and Social Networking*, *16*, 702–707.

Goodman, L. A., & Grunfeld, Y. (1961). Some nonparametric tests for comovements between time series. *Journal of the American Statistical Association*, *56*, 11–26.

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, *66*, 568–578.

Greenwood, S., Perrin, A., & Duggan, M. (2016). *Social media update 2016*. Washington, DC: Pew Research Center

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*, 861–871.

Hsieh, Y. P., & Murphy, J. (2017). Total Twitter error. In P. Biemer, et al. (Eds.), *Total survey error in practice* (pp. 23–46). Hoboken, NJ: Wiley.

Hu, M., & Liu, B. (2004). *Mining and summarizing customer reviews*. *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD 04*. doi:10.1145/1014052.1014073

Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth international AAAI conference on weblogs and social media*. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109/8122

Jensen, M. J., & Anstead, N. (2013). Psephological investigations: Tweets, votes, and unknown unknowns in the Republican nomination process. *Policy & Internet*, *5*, 161–182.

Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change*, *130*, 69–87. doi:10.1016/j.techfore.2017.11.009.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015). Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. *Ninth AAAI Conference on Weblogs and Social Media* (pp. 188–197). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*, 847–865.

Lind, L. H., Schober, M. F., Conrad, F. G., & Reichert, H. (2013). Why do survey respondents disclose more when computers ask the questions? *Public Opinion Quarterly*, *77*, 888–935.

Liu, B., Hu, M., & Cheng, J. (2005). Opinion observer. *Proceedings of the 14th International Conference on World Wide Web—WWW 05*. doi:10.1145/1060745.1060797

Loria, S. (2018). TextBlob (Python Package). Retrieved from https://textblob.readthedocs.io

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.

Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., & Rosenquist, J. N. (2011). Understanding the demographics of twitter users. *Fifth AAAI Conference on Weblogs and Social Media* (pp. 554–557). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Moore, G. H., & Wallis, W. A. (1943). Time series significance tests based on signs of differences. *Journal of the American Statistical Association*, *38*, 153–164.

Naaman, M., Boase, J., & Lai, C. H. (2010, February). Is it really about me? Message content in social awareness streams. In *Proceedings of the 2010 ACM Conference on Computer-Supported Cooperative Work*, Savannah, GA, USA, 6–10 February 2010 (pp. 189–192). New York, USA: Association for Computing Machinery.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *Fourth AAAI Conference on Weblogs and Social Media* (pp. 122–129). Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Pasek, J., & Dailey, J. (2019). Why don't tweets consistently track elections? Lessons from linking Twitter and survey data streams. In N. J. Stroud & S. McGregor (Eds.), *Digital discussions: How big data informs political communication* (pp. 68–95). New York, NY: Routledge. doi:10.4324/9781351209434-5

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*, *82*, 470–492.

Preoţiuc-Pietro, D., Volkova, S., Lampos, V., Bachrach, Y., & Aletras, N. (2015). Studying user income through language, behaviour and affect in social media. *PLoS One*, *10*, e0138717.

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.

Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., & Muhlhauser, M. (2013). *A multi-indicator approach for geolocalization of tweets*. In *International Conference on Web and Social Media*. Menlo Park, CA: Association for the Advancement of Artificial Intelligence.

Smith, B. K., & Gustafson, A. (2017). Using Wikipedia to predict election outcomes: Online behavior as a predictor of voting. *Public Opinion Quarterly*, *81*, 714–735.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The science of web surveys*. Oxford, England: Oxford University Press.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60*, 275–304.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010, May). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Fourth international AAAI conference on weblogs and social media* (pp. 178–185).

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Technology and Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics.

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

Author Biographies

Frederick G. Conrad, PhD, University of Chicago, is a research professor of survey methodology and professor of psychology at the University of Michigan, where he directs the Michigan Program in Survey Methodology. His research concerns new methods for collecting survey data and new types of data for conducting social and behavioral research. E-mail: fconrad@umich.edu

Johann A. Gagnon-Bartsch, PhD, UC Berkeley, is an assistant professor of statistics at the University of Michigan. His research focuses on causal inference, machine learning, and nonparametric methods with applications in the biological and social sciences. E-mail: johanngb@umich.edu

Robyn A. Ferg is a PhD candidate in statistics at the University of Michigan. Her research interests include finding and evaluating relationships between social media data and public opinion surveys. E-mail: fergr@umich.edu

Michael F. Schober, PhD, Stanford University, is a professor of psychology and Vice Provost for Research at The New School. His research explores interaction and miscommunication in survey interviews, collaborative music-making, and everyday conversation; how new communication technologies are changing interaction; and alternative methods for large-scale data collection. E-mail: schober@newschool.edu

Josh Pasek, PhD, Stanford University, is an 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. His research explores how new media and psychological processes each shape political attitudes, public opinion, and political behaviors as well as tools to measure these processes. E-mail: jpasek@umich.edu

Elizabeth Hou, PhD, University of Michigan, is a researcher at Systems and Technologies Research in Woburn, MA. Her research interests include statistical machine learning, sequential learning, Bayesian inference, and anomaly detection. E-mail: emhou@umich.edu.