Al Baghal et al. *EPJ Data Science* (2021) 10:32
https://doi.org/10.1140/epjds/s13688-021-00286-7

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Linking Twitter and survey data: asymmetry in quantity and its impact

Tarek Al Baghal[1*], Alexander Wenz[2], Luke Sloan[3] and Curtis Jessop[4]

*Correspondence:
talbag@essex.ac.uk
[1]University of Essex, Colchester, UK
Full list of author information is available at the end of the article

Abstract

Linked social media and survey data have the potential to be a unique source of information for social research. While the potential usefulness of this methodology is widely acknowledged, very few studies have explored methodological aspects of such linkage. Respondents produce planned amounts of survey data, but highly variant amounts of social media data. This study explores this asymmetry by examining the amount of social media data available to link to surveys. The extent of variation in the amount of data collected from social media could affect the ability to derive meaningful linked indicators and could introduce possible biases. Linked Twitter data from respondents to two longitudinal surveys representative of Great Britain, the Innovation Panel and the NatCen Panel, show that there is indeed substantial variation in the number of tweets posted and the number of followers and friends respondents have. Multivariate analyses of both data sources show that only a few respondent characteristics have a statistically significant effect on the number of tweets posted, with the number of followers being the strongest predictor of posting in both panels, women posting less than men, and some evidence that people with higher education post less, but only in the Innovation Panel. We use sentiment analyses of tweets to provide an example of how the amount of Twitter data collected can impact outcomes using these linked data sources. Results show that more negatively coded tweets are related to general happiness, but not the number of positive tweets. Taken together, the findings suggest that the amount of data collected from social media which can be linked to surveys is an important factor to consider and indicate the potential for such linked data sources in social research.

Keywords: Data linkage; Social media; Survey; Quantity; Measurement

1 Introduction

Linking social media and survey data may be useful for studying social phenomena and improving survey processes (Murphy et al. [14]) and can address methodological issues of research projects that rely on either data source exclusively. However, multiple practical considerations come before being able to conduct analyses on linked datasets. First, it is important to consider the population that the data represents, i.e., data can only be linked for those using a particular social media platform. For example, a recent Ofcom (UK) survey showed that 25% of internet users have a Twitter account, while 91% had a Facebook account (Ofcom [16]). In addition, there are important and unique ethical

Springer

concerns relating to obtaining informed consent, and the usage and storage of such data that need to be addressed (Sloan et al. [24]). This consent must be obtained within a survey, not only for ethical but also logistical reasons, so that the data sources can be linked.

After consent is obtained, an additional concern is the asymmetry in data that comes from the linked survey and social media data, specifically the variability in the data quantity obtained from social media. Survey data structure is planned, with similar data outcomes expected for most respondents. Conversely, as posting on social media platforms is unconstrained, there can be highly variant amounts of data collected per person, and the variability can affect analyses of these data (Amaya et al. [2]; Dahal, Kumar and Li [4]). If there are differences in measures derived from social media posts which are systematically related to the quantity posted, questions arise in regard to the relationship between the amount of content and its impact on analyses. Most basically, if there is a sample of respondents that post more (or less), then the results may be different than a sample posting less (or more). More broadly, the question also arises as to why these differences exist. If differences across amounts of content exist, it is unclear whether differences that are identified in substantive analyses (such as sentiment analyses) can be attributed to actual differences in attitudes or behavior, or just to differential ability to identify the signal in the varying amounts of content. The present research is, to our knowledge, the first to explore these issues as a methodological concern.

For the methodology of linking survey and social media data to further develop, it is important to explore the nature and impact of this asymmetry in data. As no previous research has been conducted in this area, a baseline understanding is required for future research to build upon. We use linked survey and Twitter data to show the extent of this data asymmetry and importance it can have on analyses. However, our research questions and methods apply across other instances of linked social media and survey data. We ask the following two research questions:

RQ1) Does the amount of data available from Twitter vary by individual characteristics identifiable through their survey and social media data?

RQ2) Does variation between individuals in the amount of collected Twitter data affect the relationship between the linked survey and social media data?

The answers to these questions inform the nature of these types of linked data broadly and the considerations that researchers need to make in conducting analyses using this methodology.

2 Linked survey and social media data and its asymmetry

Social media data, such as from Twitter, rarely contain socio-demographic information on users; in response, researchers classify users into socio-demographic categorizations, such as gender and socio-economic status, based on users' social media data (Sloan [23]). Evidence suggests that these classifications are frequently erroneous (Sloan [23]), but linked survey and social media data have the potential to improve these classification methods by testing outcomes against a gold standard (i.e., survey data). Similarly, the method of generating indicators for attitudes and behaviors from social media data may be improved through linkage with surveys. Linked survey and social media data that have corresponding types of constructs and measures can be compared to identify if outcomes are related in expected ways.

Survey research can likewise benefit from such linkage. Text analysis of social media posts (e.g., tweets) can generate indicators to match the domains measured in the survey.

Sentiment analysis can provide indicators of the valence of social media content, which may provide additional measures of respondent views and feelings. Analysis of social media posts using available lexicon dictionaries can also automatically output a number of new variables relating to, among others, linguistic, psychological, social and biological processes, beliefs, and socio-economic issues (e.g., Pennebaker et al. [18]) that can be linked to survey data. Indicators of social networks, not easily mapped in a conventional survey, may be obtained through social media metadata. These additional measures can bolster the available measures in survey data, without additional burden to respondents, and can be used for a variety of purposes, in methodological or substantive research. Given the dynamic and fluid nature of social media, these data can be generated on a continuous basis, as opposed to the static nature of a survey interview.

In order to link these data, consent is needed to be asked within the survey context, for both ethical and logistic (e.g., to obtain user handles) reasons. In the limited number of studies linking survey and social media data, consent to link survey and social media data has ranged between 27% and 90% (Murphy et al. [13]; Karlsen and Enjolras [8]; Guess et al. [5]; Al Baghal et al. [1]; Henderson et al. [6]; Wojcik and Hughes [26]). Low consent rates found in some of these studies can, in some instances, introduce bias during data analysis (Sakshaug and Kreuter [21]; Sakshaug et al. [20]). One study has explored consent rates and biases in the request to link Twitter data to survey responses (Al Baghal et al. [1]). Using the same data as the present research, this study found that almost 26% of NatCen Panel (NCP) respondents were Twitter users, and 27% of these consented to link their data. The *Understanding Society* Innovation Panel (IP) obtained a 31% consent rate among the 22% of the respondents identifying as Twitter users. Further analyses found that in the NCP women and younger respondents consented at a lower rate than men and older respondents; the only statistically significant impact in the IP was that web respondents consented less than face-to-face respondents in its mixed-mode design.

While these findings show differences between those consenting and not, there may be further error introduced *within* individuals consenting. In a survey, while respondents can answer a question or not, the design controls the amount of data a person can/should provide. No such constraint exists in an individual's use of social media, such as Twitter: they can post as much or as little as they like and control who has access to these data. This unconstrained use can result in variation in the amount of data available to link. Some consenters may post nothing, others may post but make their data private, while others may have plenty of posts available for analysis. This asymmetry in data obtained from survey and social media sources has a potentially substantial impact on analyses utilizing this novel data source.

Analyses of Twitter posts can be made using indicators that are generated to match or supplement the domains measured in the survey. For example, sentiment analysis can provide indicators of the valence of a post's content which can potentially indicate the mental wellbeing of the person who has written the post. This type of information can be either at the individual level or at the level of Twitter posts. Individual-level analyses are analogous to the survey approach: Twitter users are treated as individual cases and summary measures are calculated for each user. Conversely, tweets can be analyzed as units nested within respondents.

For either type of analysis, there is a potential difficulty in using Twitter posts due to the potentially large variation in the number of posts across individuals. A smaller number

of Twitter posts can increase within-individual variation, decreasing the precision of estimates, and may limit the possibilities of conducting complex multivariate analyses (Amaya et al. [2]). The quantity of available posts, however, also matters for individual-level analyses of Twitter data since the summary measures that are aggregated at the individual level are a function of the tweets themselves. Research using sentiment analysis shows that such techniques may also produce biased estimates as the number of tweets included decrease (Dahal, Kumar, and Li [4]). As such, measures that are based on a larger number of posts are likely to capture sentiment more accurately, and researchers may be more confident in providing summary measures for individuals who produce more content on Twitter than those posting less.

However, if the amount of data posted is related to the measures being derived from this content in a direct, systematic way, the analysis may be impacted in currently unknown ways. Those who provided more (or less) social media data may be different on the outcome of interest than those providing less (or more). Whether there is such a difference has yet to be ascertained. To the extent that differences exist, questions arise as to whether these differences are related to respondent characteristics and if there is any evidence as to if these differences would impact analyses in a detectable way.

In addition to the textual data, each Twitter post also contains associated metadata which could be linked to survey data. For example, tweets may have a location tag and more posts would provide a fuller picture of place and movement of survey respondents. Researchers could also collect information about social networks through the respondents' followers and people they follow. The amount of information is increasingly important when thinking about social media data in a real-time, longitudinal manner, as the data becomes more diffuse. To study social phenomena across time, data must be collected during the time interval of interest. Fewer posts reduces the likelihood that data will be available for that time interval, which decreases the information available for analysis or requires an increase in the size of the interval studied. Increasing the time interval, however, reduces the real-time nature and the amount of between-wave analysis available for a longitudinal study.

The present study is the first to explore the differences in the quantity of data produced across respondents in regard to how it may impact the data and methodology of linking social media and survey data. In particular, we explore the impact that this asymmetry may have when analyzing linked Twitter data and survey responses from two nationally representative panel studies. The current study analyzes 2017 NatCen and Innovation Panel data, where Twitter data of consenting respondents are collected through the Twitter application programming interface (API) and linked to their survey responses. We show how the volume of data collected from social media can vary between individuals and discuss how systematic differences across respondent characteristics may suggest possible information bias. Importantly, we provide an example of how differences in the amount of Twitter data collected from individuals can impact analyses by examining the relationship between the quantity of sentiment-scored tweets and a measure of wellbeing recorded in the IP survey.

3 Data

Respondents gave consent to link their Twitter and survey data in the tenth wave of the IP, fielded during 2017. The IP is a vehicle for methodological experimentation in a longitudinal survey design and is conducted annually. Interviews are attempted with all household

members 16+ years of age. It uses a multi-stage probability sample of persons and households in Great Britain. Refreshment samples were also selected at the fourth (IP4), seventh (IP7), and tenth wave (IP10) (University of Essex [25]). The IP is a mixed-mode survey, including web and face-to-face modes.

Individual response rates for the IP are calculated as completion rates among those responding at their initial wave of interview. At the initial wave (IP1), conducted in 2008, the individual response rate (AAPOR RR3) by sample members was 52.4%. In 2011, for the IP4 refreshment sample, the initial response rate was 44.1%, and the initial IP7 individual response rate was 24.3% in 2014. The reinterview rate at IP10 for those interviewed at IP1 was 31.2%, resulting in an overall response rate of 16.3%; for the IP4 refreshment sample, the reinterview rate at IP10 was 48.4%, resulting in an overall response rate of 21.3%; and the reinterview rate for the IP7 refreshment sample at IP10 was 61.8%, resulting in an overall response rate of 15.0%.

The NCP asked for consent to link Twitter and survey data in the July 2017 wave. It is a probability-based mixed-mode panel (web and telephone) designed to represent the British adult (18+) population (Jessop [7]). Panel members were recruited from all respondents completing the 2015 and 2016 British Social Attitudes (BSA) cross-sectional surveys, where households were selected in a three-stage design, using the Postcode Address File (PAF), a list of addresses (or postal delivery points) compiled by the UK Post Office. The July 2017 wave was the ninth fieldwork wave of the NCP since November 2015, which is conducted at irregular intervals (no more than one in any month) to address specific research issues. For the July 2017 NCP, the survey response rate (AAPOR RR1), i.e. the proportion of people issued to the survey that took part, was 59.6%. The overall response rate, i.e. the participation from the original sample frame using the proportion of participants eligible for an interview in the BSA, was 14.7%. Since the goal of our analysis is to study the sample of survey respondents who consented to link their Twitter data rather than making generalizable estimates, our analyses are unweighted.

3.1 Data collection

We used the *rtweet* package (Kearney [9]) in R (R Core Team [19]) to collect Twitter data from IP and NCP respondents who provided their Twitter username. While other packages are available in R to retrieve data from the Twitter API, *rtweet* is the only package to our knowledge that allows capturing tweets and metadata for a set of usernames and is currently maintained (see https://cran.r-project.org/web/packages/rtweet/readme/README.html). For a given user, the package extracts the following data from Twitter's Representational State Transfer (REST) API, which provides data in JavaScript Object Notation (JSON) format and is converted to an R dataset:

- List of tweets posted by the user: original tweets, retweets (user re-posts tweet from another user), or quoted tweets (user re-posts tweet from another user plus their comment)
- Metadata about each tweet (see Sloan et al. [24] for further details): date and time of the tweet posting; location of the tweet posting; application posting the tweet (e.g. iPhone App); number of retweets for posting; and number of times the tweet has been liked ("favorited")
- Metadata about the user: name; profile description; location; website; number of accounts they follow ("friends"); number of accounts following them ("followers");

Table 1 Outcome of Twitter Data Collection in IP10 and NCP. Conditional Percentages

	IP10		NCP	
	N	%	*N*	%
Total respondents	1945	100	2184	100
Have personal Twitter account	428	22	558	26
Consented to link Twitter and survey data	171	40	151	27
Provided username	163	100	150	100
Have public account and tweets > 0	127	78	113	75
Have public account and tweets = 0	6	4	5	3
Have private account	16	10	8	5
Invalid username	14	9	24	16

number of tweets ever posted ("statuses count"); date and time of account creation; whether account is public or private (tweets only visible to approved followers); and whether the account is verified.

Note that Twitter data can only be captured from users with a public account, i.e. where tweets are publicly visible. Furthermore, the Twitter API has data capture restrictions specifying that for each user only the most recent 3200 tweets are retrievable (https://developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/overview). There are also limits to the number of requests per time interval: up to 900 calls can be made to the Twitter API per 15-minute window (i.e., Twitter data can be requested for up to 900 users), and up to 100,000 calls per 24-hour window.

In the IP, 428 respondents (22% of the full sample) indicated they have a personal Twitter account, 171 respondents (33% of those with a Twitter account) consented for their data collected from Twitter to be linked with their survey data, and 163 respondents provided their username. Most of those respondents (78%) have a public Twitter account and posted at least one tweet on their timeline. Another 14% of usernames provided either have a private account (i.e. the content of tweets is not publicly available) or they have a public account but have posted no tweets. For these users, we can only collect the account metadata. The remaining 9% of usernames provided are invalid, and we cannot collect any Twitter data. We manually looked up these usernames using the search function on the Twitter website and found that while some of these names are non-existent, others relate to a deleted or suspended account. Results are similar for the NCP (Table 1). Overall, 29.7% of Twitter-using respondents (the reference population) in the IP sample and 20.3% in the NCP sample have usable Twitter data, i.e., provided a valid username and have a public Twitter account with at least one tweet.

3.2 Twitter data quantity analyses

Of central interest for this paper is the amount of data provided by individuals in their tweet content, one indication of which is the total number of posts a person makes ("statuses count"). Over time, posts can be extracted repeatedly, and the data available to link can include the most recent 3200 posts for the initial collection plus all posts from that point on. To reflect that studies could collect many more posts over time and to better reflect the true variation in amount of data generated by respondents, we use the total number of posts ever made as the outcome variable. Since data would only be available for public accounts with at least one tweet, only these are further analyzed, leaving 127 cases from the IP and 113 from the NCP.

Additional variables from the metadata can be informative to the amount of data that is available to link to survey responses; here we focus on the number of followers and friends (accounts followed) respondents have on Twitter. Both can be useful for studying social networks and are additional data that may be linked. Importantly for the current study, both number of followers and friends are also useful to understand Twitter data generation, i.e. via posting. Prior research has identified a strong positive relationship between the number of tweets and the number of followers (Kwak et al. [11]), with the number of friends being an even stronger predictor of posting behaviour than followers (Yang and Counts [27]).

3.3 Sentiment analysis

Given the availability of well-being survey measures in the IP only, this particular example does not include the NCP sample. We performed sentiment analysis on individual tweets collected from IP respondents using both the Afinn (Nielsen [15]) and Bing (Liu [12]) lexicons, which are dictionaries of words with sentiment ratings included in the R *tidyverse* package (Silge and Robinson [22]). We use both lexicons in the analyses to increase the chances of reliable results. The Afinn lexicon scores words on a −5 (most negative) to +5 (most positive) scale, while the Bing lexicon simply indicates a word as either positive or negative. However, this simpler scoring allows for more words to be included in the analysis. Both lexicons focus on a set of affective words; 2477 words are coded in the Afinn lexicon and 6786 in the Bing lexicon. Non-textual data (e.g. emojis), symbols or links are not included in the lexical sentiment analysis. Using these two lexicons, two sentiment scores are created per tweet: one each for the Afinn and Bing lexicons. The Afinn sentiment score is the summation of word values coded from the lexicon. For the Bing sentiment score, the number of positive and negative words are each counted, and then the difference between these are taken. Positive values mean that more positive words are used in the tweets and indicate an overall positive sentiment; negative values represent the opposite. We do not include retweets in our sentiment analysis, only original tweets by the respondent.

As not all words are coded, not all tweets receive a sentiment score for each lexicon. As such, not every respondent had a sentiment-coded tweet, and given the larger number of words in the Bing lexicon, slightly more respondents had a Bing-scored tweet than Afinn-scored. A total of 119 respondents (of the 127 which we were able to collect tweets from) had at least one tweet with a Bing score, while 117 had at least one Afinn-scored tweet. Where sentiment scores are calculated, we code the tweet as being positive, negative, or neutral on each sentiment score separately. We then count the number of positive, negative, and neutral tweets each respondent made. Respondents with at least one collected tweet but no sentiment score were counted as having zero positive, negative, and neutral tweets.

For comparison to the survey data, which contains an equivalent amount of data across respondents, we link sentiment scores for respondents' tweets to their survey responses, focusing on indicators of well-being. Sentiment scores are intended to measure positivity or negativity, and without limitations on what content is being included, will reflect sentiment across domains generally. For this reason, we focus on a measure of general mental well-being. In particular, we use answers to a question about general happiness from the General Health Questionnaire (GHQ) (Cox et al. [3]). The question asks "Have you recently been feeling reasonably happy, all things considered?" and we collapse the

response options "More so than usual" and "About the same as usual", considering these "happy" responses. The remaining response options "Less so than usual" and "Much less than usual" are collapsed as the "unhappy" category, creating a dichotomous outcome. Of the respondents providing at least one public tweet (and hence available for sentiment coding), 82.1% (n = 103) are in the "happy" and 17.9% (n = 23) are in the "unhappy" category; one respondent declined to answer.

3.4 Additional covariates

The remainder of the covariates explored come from the survey data. While these data could be collected using other methods (e.g. non-probability methods), we are evaluating the methodology of linked data, and hence the necessity of using these large probability surveys. From these surveys, we select additional variables that are available in both the IP and NCP, including objective and subjective measures to indicate possible biases in data collected across a range of respondent characteristics. We use these variables for multivariate analyses. Education is indicated in three categories (precoded in the data): higher education degree, professional degree/A-levels, and anything lower. Analyses compare employed to unemployed, females to males, married to non-married, frequent internet use to less frequent and age is measured continuously in years.

Income is categorized by monthly income and is dichotomized as more or less than £2400 per month, the (precoded) categorization closest to the median earnings in the UK (ONS [17]). Any missing data on income in the IP is imputed using several techniques (Knies [10]). There was one respondent in the subset of NCP data for which income data was missing who was excluded from the analysis. Ethnicity is indicated as white British identity or not in the IP, although it is more basically indicated as white or not in the NCP. Subjective current financial wellbeing is one of the few subjective measures captured for all IP respondents, and the only one that is measured identically in both surveys. Those saying they are "living comfortably" or "doing alright" financially are combined and compared to respondents who say they are less well off (all questions used are in the Appendix). Given the smaller sample sizes, we use both the $p < 0.10$ and $p < 0.05$ cut-offs to indicate statistical significance.

4 Results

Calls to the Twitter API collected a total of 118,593 and 132,687 tweets from IP and NCP respondents, respectively. Table 2 presents descriptive statistics for tweets we collected, tweets ever posted (total posts), followers, and friends. In both datasets, there are similar numbers of posts available per respondent, with slightly larger amounts available in the NCP: While we collected a median number of 304 posts per IP respondent, we collected a median number of 485 posts per NCP respondent. Besides the amount of data collected from tweets, information about friends and followers is also available. There are larger numbers of friends than followers per person: in the IP, for example, respondents have a median number of 182 friends and 71 followers. But even for the median number of followers, the data suggest that most respondents bring a wider set of data beyond their own posts.

There is substantial variation in the amount of posting from these respondents' Twitter accounts. In the IP data, although 19 respondents had over 3200 posts, the call made to the Twitter API never returned the maximum 3200 for any respondent; rather 3199

Table 2 Data from Survey Respondents' Twitter Accounts

	Median	Mean	SD	Min	Max
IP					
Collected	304	933.63	1157.60	1	3199
Total Posts	306	2255.32	6057.36	1	36,451
Followers	71	260.25	568.95	1	3734
Friends	182	350.95	567.54	0	3912
NCP					
Collected	485	1174.11	1269.28	1	3200
Total Posts	519	3655.01	14020.68	1	139,754
Followers	70	442.19	1750.14	0	17,941
Friends	235	455.30	557.02	3	3318

was the maximum number of posts extracted. Due to the high skew to the distribution, the mean is substantially larger than the median across all measures, with the standard deviation larger than the mean in every case. This large variation in the amount of data available contrasts with the more planned nature of survey data. While respondents may provide different amounts of data because of routing or item-nonresponse, the variation in amounts is unlikely to reach levels seen in Twitter data, and unlike in survey data, this variation should be an expected feature of the data.

This leads to the first research question, whether this variation is related to individual characteristics or not. If this variance in amount of content produced is random, then we may not be too concerned about the range in activity between users. However, if it is not, and if the amount of content is a function of demographics or other factors, then there is the risk of introducing bias into analyses of linked data. To examine this, we specify a model in which the total number of tweets is the dependent variable, with predictors coming from both the survey and Twitter data. Due to the high variation in the total number of posts made seen in Table 2, the natural log of total posts is used as the outcome in a general linear model. The predictors include both the demographics and subjective economic well-being from survey data and the friends and followers on Twitter as outlined above. For the same reasons as with total posts, we use the natural log of the number of friends and followers. Table 3 presents the results for models of both IP and NCP data.[1]

The results suggest there are indeed characteristics related to variation in the volume of social media posts linked to survey data. There is a strong relationship between the number of followers and the total number of posts made in both surveys, consistent with previous findings (Kwak et al. [11]; Yang and Counts [27]). Unlike those studies, the linked survey data allow for additional controls in this analysis and show that this effect persists after accounting for individual characteristics. However, the number of friends shows no relation to posting behavior. The directionality between followers and posting is not entirely clear; more posts could attract followers, or having more followers may compel users to post more. In either case, it is important to control for this relationship in identifying the relationships between other respondent characteristics and total posts.

Few of the respondent characteristics included have a significant relationship to the number of posts. None are significant in the NCP. Women post significantly less (at $p < 0.10$) on Twitter than men in the IP. The only other significant impact found among IP respondents is that those with professional or A-Level degrees post more than those with

[1] Using the count of tweets collected as the dependent variable did not change substantive results.

Table 3 Prediction of Number of Total Twitter Posts (log)

	IP (S.E.)	NCP (S.E.)
#Followers (log)	1.004**	0.958**
	(0.114)	(0.137)
#Friends (log)	0.155	0.242
	(0.122)	(0.164)
Female	–0.424*	–0.420
	(0.250)	(0.317)
Employed	–0.539	–0.071
	(0.390)	(0.364)
Age	–0.004	0.005
	(0.011)	(0.013)
HH Income £2400+	–0.298	0.074
	(0.360)	(0.386)
Education (baseline other education)		
Higher Ed. Degree	0.593	–0.516
	(0.386)	(0.368)
Professional/A-levels	0.843**	–0.515
	(0.375)	(0.485)
Frequent Internet Use	1.106	0.507
	(0.844)	(0.552)
White Ethnicity	0.420	–0.026
	(0.342)	(0.727)
Married	0.003	–0.073
	(0.300)	(0.326)
Positive Econ. Well-Being	0.043	–0.150
	(0.280)	(0.332)
Constant	–0.386	0.983
	(1.063)	(1.025)
Chi-Sq/DF	*1.86*	*2.42*
n	*127*	*112*

*$p < 0.10$
**$p < 0.05$

less education. University educated respondents tend to post more in the IP as well, but this does not achieve a statistically significant effect. Overall, there is some uncertainty about how respondent characteristics relate to the amount of Twitter data available to use with linked survey responses; where it is significant, it appears to be greater amounts of data available for men and somewhat more educated respondents.

As well as whether there is variation related to respondent characteristics, it is also important whether differences in the number of tweets available impact the analysis of linked data. To address this second research question using an exemplar, we estimate the relationship between the quantity of positive, neutral and negative tweets collected and respondents' general happiness measured in the survey. For tweets with scores on both Afinn and Bing lexicons, the correlation of scores across the two lexicons is very high, $r = 0.88$. As the two lexicons are intended to indicate affect broadly in the same way, this finding is reassuring and not surprising. Table 4 presents additional descriptive statistics for the number of scored tweets using the Afinn and Bing lexicons. Overall, more tweets are positively scored than negatively scored using either lexicon, suggesting a generally more positive tone to content.

While the coded tweets are more positive overall, there are also a sizable number of negative tweets per respondents in either lexicon. There are relatively few neutrally coded tweets, however, although some respondents have more of these using the Bing scoring. As with the overall quantity of Twitter data available, the statistics in Table 4 show great

Table 4 Number of Tweets by Afinn and Bing Sentiment Score, Innovation Panel

	Median	Mean	SD	Min	Max
Afinn					
Positive	52	167.54	242.78	0	1233
Negative	24	98.69	160.64	0	788
Neutral	2	7.94	12.76	0	62
Bing					
Positive	64	188.58	265.84	0	1350
Negative	28	114.54	181.71	0	946
Neutral	8	35.26	54.52	0	250

variation in the number of sentiment-scored tweets collected. The high skew in distributions leads to means substantially larger than the medians across all sentiments, with large standard deviations and shows the substantial variability in data availability across survey respondents.

To identify if the quantity of sentiment-scored tweets is related to well-being measured in the survey, we estimate a logistic regression model predicting the "unhappy" category of response on the general happiness measure. We include the number of positively and negatively scored tweets as covariates in separate models for the Afinn and Bing lexicons. As with the above analyses, the natural log of the number of sentiment-scored tweets is used to account for the skewed distribution. The models do not include the number of neutral tweets, to account for the inclusion of respondents with no coded tweets at zero values on all counts, but also because these may be less relevant given the desire to relate affective content to well-being. The remainder of covariates are those included in the above analyses on overall tweet counts as possible controls on quantity and to maintain consistency. The exception is that frequent internet use had to be dropped due to lack of variation on this variable and the dependent variable. Results of the models using counts of Afinn and Bing sentiment-scored tweets are presented in Table 5.

The key result is the relationship between negatively scored tweets and general happiness. For the Bing lexicon negatively scored tweets, there is a statistically significant relationship ($p = 0.052$) with general happiness, and for the Afinn negatively scored tweets there is a near-significant relationship ($p = 0.108$). The difference in significance is likely to be attributable in part to the fewer words and hence coded tweets in the Afinn lexicon. The impact of coding differences is further suggested by the smaller coefficient estimated using the Afinn lexicon. However, there is no impact found in either model for the number of positively coded tweets, although these are more commonly found in the data. This finding suggests that the significant relationship between negatively scored tweets and general happiness is not only a function of the number of tweets overall, but that the sentiment of tweets is important.

The interpretation of the estimated effects of negative tweet counts on general happiness suggests a relationship that may be contrary to expectation. That is, a greater number of negatively coded tweets are related to a lower estimated probability of a respondent giving an "unhappy" response to the general happiness question. This result may be due to a number of factors which deserve greater exploration but are beyond the scope of this paper. In particular, there needs to be a greater examination of the actual content of these tweets. Regardless, the results suggest the differential relationship that the quantity of content extracted from Twitter may have on outcomes studied in linked data sources. In this

Table 5 Prediction of Unhappy Response from Sentiment Scored Tweet Quantity

	Afinn (S.E.)	Bing (S.E.)
#Positive Tweets (log)	0.058	0.230
	(0.327)	(0.361)
#Negative Tweets (log)	–0.525	–0.733*
	(0.326)	(0.377)
#Followers (log)	0.543	0.582*
	(0.349)	(0.352)
#Friends (log)	0.283	0.295
	(0.300)	(0.298)
Female	1.704**	1.804**
	(0.609)	(0.628)
Employed	2.073*	2.265*
	(1.172)	(1.211)
Age	0.033	0.030
	(0.025)	(0.025)
HH Income £2400+	–0.651	–0.703
	(0.804)	(0.812)
Education (baseline other education)		
Higher Ed. Degree	0.666	0.627
	(0.960)	(0.956)
Professional/A-levels	1.125	1.144
	(0.938)	(0.935)
White Ethnicity	0.653	0.777
	(0.769)	(0.782)
Married	–0.662	–0.686
	(0.629)	(0.627)
Positive Econ. Well-Being	–0.417	–0.397
	(0.568)	(0.574)
Constant	–8.366**	–8.772**
	(2.427)	(2.510)
Chi-Sq/DF	*1.90*	*2.01*
n	*126*	*126*

*$p < 0.10$
**$p < 0.05$

instance, the relationship between general happiness and negative sentiment expressed in Twitter posts is different for individuals with different quantities of available data. This finding evidences the need for analyses identifying possible biases in data quantities such as those presented in Table 3.

While the key finding is the impact of the quantity on a substantive outcome, Table 5 also displays how linked social media and survey data can provide additional insights. In particular, respondents with a greater number of followers (in the Bing coded count model) are more likely to provide an "unhappy" response. This result deserves further exploration and is one that is only possible to identify using similarly linked data. Only two other variables show a significant relationship to the general happiness outcome. Controlling for both social media and other survey variables, women and those employed are also significantly more likely to give "unhappy" responses. The combination of results from both social media and survey data are suggestive of the varied analyses this type of data linkage can provide.

5 Discussion

Linked Twitter and survey data provide a novel data source that is greater than the sum of its parts. However, even after respondents have given their consent and provided Twit-

ter usernames, it is important to understand the analytical limitations. Recent research exploring Reddit social media data similarly points out the importance of understanding the data, and that differences in amounts of data can impact analyses (Amaya et al. [2]). We expand on these ideas by linking Twitter data from survey respondents directly to their survey responses and provide initial evidence on the impact the amount of data collected from Twitter can have on analyses of these linked data. We find some consenters may post nothing, others may post but make their data private, while others may have plenty of posts available for analysis. We note that this is an advantage of the linked approach: a sample directly from the Twitter API is a sample of tweets and therefore misses out those people who do not post. By linking to survey data, we also know who does not post, and can possibly account for it.

Our research focuses on the nature of linked survey and social media in regard to the differences in the amount of data created by respondents posting on Twitter. In particular, we explore whether the variation in data obtained from survey respondents' Twitter account is related to individual characteristics and whether such variation can impact analyses of linked data. On both counts, the results illustrate important outcomes that should be considered by researchers using this methodology. The amount posted varies greatly, as do the number of friends and followers on the site. A strong positive relationship exists between the number of followers and posting behavior, although the directionality of this relationship needs further exploration. Importantly, the analyses provide some evidence that differences in content production are related to gender and educational attainment, but not to other important covariates. Given the finding that women and less educated individuals may post less (at least in the IP), there is a need for further exploration about the impact this may have when using the linked data in substantive research. As much research using Twitter is concerned with volume of content (e.g. to see what is 'trending'), this paper has demonstrated that some demographic groups may be under-represented in such studies as they tweet less.

We provide an initial example of how quantity may affect substantive outcomes. Sentiment analysis of tweets demonstrates the possible impact of data quantity created by posting on Twitter on estimating general happiness measured in linked survey data. The number of tweets coded as having a negative affect is significantly (or nearly so) related to "unhappy" responses recorded in linked survey data. The results show that differential production of content leads to different relationships with an outcome of interest. If the data contained mostly respondents which consistently produced more content (negative-valence tweets in this example), then analyses combining these data to survey responses could lead to different interpretations than in instances with fewer tweets.

The mechanisms for what may be driving these differences need additional exploration as the use of these linked data sources develops. For example, the relationship is the opposite of what might be expected, with more negative tweets indicating a lower probability of "unhappy" responses. One possible explanation could be people more unhappy (more directly indicated in the survey) are less likely to post to social media as the present analysis shows the number of positive tweets is not significantly related to this outcome. These results necessitate further research but may be subject to certain limitations. Basic automated sentiment analyses like those used here do not capture phrases. For example, "I wish I was happy" would be counted as positive given the positive word "happy".

The lack of significant results in the above analyses, both exploring who is creating Twitter data and how this data can impact other outcomes, is not definitive that a relationship or bias does not exist, particularly given the limitation of small sample sizes in both the IP and NCP. Given that a subset of the population uses any given social media platform, and the relatively low consent rates found across studies, small numbers are a potential issue researchers using this methodology may continue to face. However, the significant results found in this study warrant further research and future research will endeavor to, among other goals, generate these linkages on larger data sets. As well as linkages to other studies, current research is also focusing on ways to improve consent rates, which will increase sample sizes in smaller studies as well.

Appendix: Questions used

IP

Twitter handle

What is your Twitter username?

SOFTCHECK: "Twitter usernames must begin with an @ character, followed a maximum of 15 characters (A-Z, a-z, 0-9, underscore), no word spaces. Please check and amend."

General happiness

Have you recently been feeling reasonably happy, all things considered?

1. More so than usual
2. About the same as usual
3. Less so than usual
4. Much less than usual

Sex

[Are you] / [Is NAME]…

1. Male
2. Female

Ethnicity

What is your ethnic group?

1. British/English/Scottish/Welsh/Northern Irish
2. Irish
3. Gypsy or Irish Traveller
4. Any other White background
5. White and Black Caribbean
6. White and Black African
7. White and Asian
8. Any other mixed background
9. Indian
10. Pakistani
11. Bangladeshi
12. Chinese

13. Any other Asian background
14. Caribbean
15. African
16. Any other Black background
17. Arab
97. Any other ethnic group

Education

Can you tell me the *highest* educational or school qualification you have obtained?

1. PhD or equivalent doctoral level qualification
2. Masters or equivalent higher degree level qualification
3. Postgraduate academic below-Masters level qualification (e.g. Certificate or Diploma)
4. Bachelors or equivalent first degree qualification
5. Post-secondary academic below-degree level qualification (up to 1 year)
6. Post-secondary academic below-degree level qualification (2 and more years)
7. Post-secondary vocational training (up to 1 year)
8. Post-secondary vocational training (2 and more years)
9. Completed secondary school
10. Completed primary school
96. None of the above

Subjective economic wellbeing

How well would you say you are managing financially these days? Would you say you are…?

1. Living comfortably
2. Doing alright
3. Just about getting by
4. Finding it quite difficult
5. Finding it very difficult

HH income

Derived variable from several questions, with imputation. See Knies [10]

Employment

Can I just check, did you do any paid work **last week** - that is in the seven days ending last Sunday - either as an employee or self-employed?

1. Yes
2. No

Internet

How often do you use the internet for your personal use?

1. Every day
2. Several times a week
3. Several times a month
4. Once a month

5. Less than once a month
6. Never use
7. No access at home, at work or elsewhere

Marital status

What is [NAME's / CNAME's / ff_forname's / your] legal marital status?

1. Single and never married or never in a legally recognised Civil Partnership
2. Married
3. A Civil Partner in a legally recognised Civil Partnership
4. Separated but legally married
5. Divorced
6. Widowed
7. SPONTANEOUS: Separated from Civil Partner
8. SPONTANEOUS: A former Civil Partner, the Civil Partnership legally dissolved
9. SPONTANEOUS: A surviving Civil Partner (partner having died)

Age

What is [your]/[NAME]'s/[forname]'s date of birth?

NCP

Collected in BSA interview

Sex

INTERVIEWER: PLEASE CODE SEX OF RESPONDENT

1. Male
2. Female

Ethnicity

To which of these groups do you consider you belong?

1. BLACK: of African origin
2. BLACK: of Caribbean origin
3. BLACK: of other origin (WRITE IN)
4. ASIAN: of Indian origin
5. ASIAN: of Pakistani origin
6. ASIAN: of Bangladeshi origin
7. ASIAN: of Chinese origin
8. ASIAN: of other origin (WRITE IN)
9. WHITE: of any origin
10. MIXED ORIGIN (WRITE IN)
11. OTHER (WRITE IN)
98. (Don't know)
99. (Refusal)

Collected in July 2017 or earlier NCP wave

Twitter handle

What is your Twitter username?

SOFTCHECK: "Twitter usernames must begin with an @ character, followed a maximum of 15 characters (A-Z, a-z, 0-9, underscore), no word spaces. Please check and amend."

Education

Starting from the top, please look down the list of qualifications and select the first one you come to that you have passed

1. Degree or equivalent, and above
2. A levels or vocational level 3 or equivalent, and above
3. Other qualifications below A levels or vocational level 3 or equivalent
4. Other qualification (Please describe)
5. No qualifications

Subjective economic wellbeing

How well would you say you are managing financially these days?

Would you say you are …?

1. Living comfortably
2. Doing alright
3. Just about getting by
4. Finding it quite difficult
5. Finding it very difficult

HHIncomeChk

You have previously told us that your total *monthly* household income from all sources *before tax* was {HHIncomeTxt}.

To help us with our analysis, could you please confirm if that is still correct?

1. Yes
2. No

HHIncomeUpd {*IF HHIncomeChk* = 2}

Could you please tell us which of the following represents the total monthly income of your household from all sources **before tax**?

If you are unsure, please give your best estimate.

1. Less than £590
2. £591–770
3. £771–910
4. £911–1000
5. £1001–1200
6. £1201–1300
7. £1301–1500
8. £1501–1700
9. £1701–1900
10. £1901–2200
11. £2201–2400
12. £2401–2700
13. £2701–3000

14. £3001–3300
15. £3301–3700
16. £3701–4200
17. £4201–4800
18. £4801–5600
19. £5601–7200
20. £7201 or more

Employment

Which of these descriptions applied to what you were doing last week, that is the seven days ending last Sunday?

1. In full-time education (not paid for by employer, including on vacation)
2. On government training/employment programme
3. In paid work (or away temporarily) for at least 10 hours in week
4. Waiting to take up paid work already accepted
5. Unemployed and registered at a JobCentre or JobCentre Plus
6. Unemployed, not registered, but actively looking for a job (of at least 10 hrs a week)
7. Unemployed, wanting a job (of at least 10 hrs per week) but not actively looking for a job
8. Permanently sick or disabled
9. Wholly retired from work
10. Looking after the home
11. Doing something else

Internet use

On average, how often would you say you access the internet for personal use?

Please include time spent on the internet on all devices you use, for example a computer, laptop, tablet or smartphone"}

1. Several times a day
2. Daily
3. Weekly
4. Monthly
5. Less often than once a month
6. {IF TEL: Do not have access to the internet}

Marital status

Which of these applies to you at present?

1. Married
2. In a registered same-sex civil partnership
3. Living with a partner
4. With a partner you do not live with
5. Separated (after being married or in a same-sex civil partnership)
6. Divorced/dissolved same-sex civil partnership
7. Widowed/surviving partner from a same-sex civil partnership
8. Single (never married/never in a civil partnership)

Age

So we don't have to check your age each time we conduct a new survey, could you please tell us your date of birth?

1. Yes
2. No

Please enter your date of birth

AgeCheck

What was your age at your last birthday?

Range 18…110

Acknowledgements
Not applicable.

Funding
The first, third, and fourth authors are funded by a research award from the UK Economic and Social Research Council (award no. add ES/S015175/1) for "Understanding [Offline/Online] Society: Linking Surveys with Twitter Data". The remaining author received no financial support for the research, authorship, and/or publication of this article.

Abbreviations
NCP, NatCen Panel; IP, Understanding Society Innovation Panel; AAPOR, American Association of Public Opinion Research; RR, Response rate; API, Application programming interface; BSA, British Social Attitudes Survey; REST, Representational State Transfer; JSON, JavaScript Object Notation; ONS, Office of National Statistics; Ofcom, Office of Communications; PAF, Postal Address File.

Availability of data and materials
The NatCen Panel 2017 can be requested from NatCen Social Research, with initial contact being sent to info@natcen.ac.uk. Requests are subject to NatCen's Data Release Panel.
The Understanding Society Innovation Panel waves 1-10 are available through the UK Data Service, study 6614, at https://discover.ukdataservice.ac.uk/catalogue/?sn=6614
The linked datasets generated and/or analysed during the current study are not currently publicly available due to identifiability of survey respondents, but de-identified data will be available at the end of the project from the corresponding author on reasonable request.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
TAB analysed the data, wrote the paper, reviewed drafts of the paper. AW collected the data, wrote the paper and reviewed drafts of the paper. LS and CJ wrote the paper and reviewed drafts of the paper. All authors read and approved the final manuscript.

Author details
[1]University of Essex, Colchester, UK. [2]University of Mannheim, Mannheim, Germany. [3]Cardiff University, Cardiff, UK. [4]NatCen Social Research, London, UK.

Publisher's Note

Received: 3 September 2020 Accepted: 25 May 2021 Published online: 09 June 2021

References

1. Al Baghal T, Sloan L, Jessop C, Williams M, Burnap P (2019) Linking Twitter and survey data: the impact of survey mode and demographics on consent rates across three UK studies. Online first at Soc Sci Comput Rev. https://doi.org/10.1177/0894439319828011
2. Amaya A, Bach R, Keusch F, Kreuter F (2019) New data sources in social science research: things to know before working with reddit data. Soc Sci Comput Rev. https://doi.org/10.1177/0894439319893305
3. Cox BD, Blaxter M, Buckle ALJ, Fenner NP, Golding JF, Gore M, Huppert FA, Nickson J, Roth M, Stark J, Wadsworth MEJ, Whichelow M (1987) The health and lifestyle survey preliminary report of a nationwide survey of the physical and mental health, attitudes and lifestyle of a random sample of 9,003 British adults. Health Promotion Research Trust, London
4. Dahal B, Kumar SAP, Li Z (2019) Topic modeling and sentiment analysis of global climate change tweets. Soc Netw Anal Min 9:24
5. Guess A, Munger K, Nagler J, Tucker JA (2018) How accurate are survey responses on social media and politics? Polit Commun 36:241–258

6. Henderson M, Jiang K, Johnson M, Porter L (2019) Measuring Twitter use: validating survey-based measures. Soc Sci Comput Rev. https://doi.org/10.1177/0894439319896244
7. Jessop C (2018) The NatCen panel: developing an open probability-based mixed-mode panel in Great Britain. Soc Res Pract 6:2–14
8. Karlsen R, Enjolras B (2016) Styles of social media campaigning and influence in a hybrid political communication system: linking candidate survey data with Twitter data. Int J Press/Politics 21:338–357
9. Kearney MW (2019) rtweet: collecting Twitter Data R package version 0.69 https://cran.r-project.org/web/packages/rtweet/readme/README.html Accessed 5 April 2021
10. Knies G (ed) (2018) Understanding society: the UK household longitudinal study waves 1-8, user manual institute for social and economic research University of Essex, Colchester
11. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter. In: A social network or a news media? Proceedings of the WWW conference 2010, pp 591–600
12. Liu B (2015) Sentiment analysis: mining opinions, sentiments, and emotions. Cambridge University Press, Cambridge
13. Murphy J, Landwehr J, Richards A (2013) Using Twitter to Predict Survey Responses. Paper presented at the Midwest Association of Public Opinion Research conference. Nov 2013
14. Murphy J, Link MW, Hunter-Childs J, Langer-Tesfaye C, Dean E, Stern M, Pasek J, Cohen J, Callegaro M, Harwood P (2014) Social media in public opinion research: report of the AAPOR task force on emerging technologies in public opinion research. American Association for Public Opinion Research, Deerfield
15. Nielsen FA (2011) A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of the ESWC2011 workshop on 'Making sense of microposts': big things come in small packages 718 in CEUR workshop proceedings, pp 93–98
16. Office of Communications (Ofcom) (2019) Adults' Media Use and Attitudes Report 2019. Research Document
17. Office of National Statistics (2019) Average household income, UK: Financial year ending 2019 (provisional) *Statistical Bulletin*
18. Pennebaker JW, Booth R, Boyd RL, Francis ME (2015) Linguistic inquiry and word count: LIWC2015. Pennebaker Conglomerate, Austin
19. R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.r-project.org/. Accessed 5 April 2021
20. Sakshaug JW, Couper MP, Ofstedal MB, Weir D (2012) Linking survey and administrative records: mechanisms of consent. Sociol Methods Res 41:535–569
21. Sakshaug JW, Kreuter F (2012) Assessing the magnitude of non-consent biases in linked survey and administrative data. Surv Res Methods 6:113–122
22. Silge J, Robinson D (2017) Text mining with R: a tidy approach. O'Reilly Media, Inc, Sebastopol
23. Sloan L (2017) Social science 'Lite'? Deriving demographic proxies from Twitter. In: Sloan L, Quan-Haase A (eds) The SAGE handbook of social media research methods. Sage, Thousand Oaks, pp 90–104
24. Sloan L, Jessop C, Al Baghal T, Williams M (2019) Linking survey and Twitter data: informed consent, disclosure, security, and archiving. Journal of Empirical Research on Human Research Ethics. https://doi.org/10.1177/1556264619853447
25. University of Essex Institute for Social and Economic Research (2018) Understanding Society: innovation Panel, Waves 1-10, 2008-2017 [data collection] 9th Edition UK Data Service SN: 6849. https://doi.org/10.5255/UKDA-SN-6849-10
26. Wojcik S, Hughes A (2019) Sizing Up Twitter Users Pew. Research Center Report. April 2019
27. Yang J, Counts S (2010) Predicting the speed, scale, and range of information diffusion in Twitter. In: Proceedings of the fourth international AAAI conference on weblogs and social media