The Survey Statistician, 2023, Vol. 87, 27–34

New and Emerging Methods

Using Social Media to Enhance Survey Data

Annamaria Bianchi[1], Camilla Salvatore[2] and Silvia Biffignandi[3]

[1] University of Bergamo, annamaria.bianchi@unibg.it
[2] University of Milano-Bicocca, c.salvatore4@campus.unimib.it
[3] Consultancy in Economic Statistics Studies (CESS), biffisil@teletu.it

Abstract

This article provides an overview on the roles of social media (SM) in survey research. After examining the characteristics and challenges of using social media data in statistical research, we discuss recent approaches on ways SM have been used to enhance survey research. We then introduce a general modular framework for producing statistics taking advantage of the two data sources. Finally, we highlight important questions for future research.

Keywords: Augmenting information, Smart surveys, Smart statistics, Data integration.

1 Introduction

Probability sample surveys have been considered the gold standard for inference for many years, but they are facing difficulties related mainly to declining response rates and related increasing costs (Luiten et al., 2020; Brick and Williams, 2013). At the same time, an acceleration of technological advances has occurred, with the use of mobile phones and online social networks, specifically social media (SM), leading to the availability of vast amounts of new data. This is coupled with the development of new tools by computational social scientists to collect, process, and analyse digital trace data.

All this has led to an extensive use of SM data in research to better understand attitudes and behaviours with reference to socio-economic phenomena. SM data has been used, for instance, to examine political attitudes (Bail et al., 2020) and emerging political trends (Rill et al., 2014), active citizenship (Rosales Sánchez et al., 2017) and well-being (Luhmann, 2017; Iacus et al., 2022). A number of experimental statistics have been developed by Official Statistics using such textual data to study social tensions[2] and consumers' confidence in the economy, among other applications (Daas and Puts 2014; Istat's Social Mood on Economy Index[3]). For a complete overview on the use

[2] https://www.cbs.nl/en-gb/about-us/innovation/project/social-tensions-indicator-gauging-society
[3] https://www.istat.it/en/experimental-statistics/experiments-on-big-data

of SM and digital trace textual data in Official Statistics, please refer to Japec and Lyberg (2020). Over the last few years, an increasing amount of applied and methodological research has been conducted to understand how this new paradigm can leverage SM in different ways to advance survey research. In this respect, one important and promising direction consists in the combination of survey and SM data (Hill et al., 2019; Stier et al., 2020). This needs to take into account pitfalls inherent in SM data, including self-selection, limited demographic information about users, data accessibility, volatility, and coverage among others. Stier et al. (2020) advocate for the need to develop a conceptual and theoretical framework tailored toward the multidimensionality of such data, guiding researchers through the benefits and pitfalls of different approaches to data linking.

This paper first examines some of the challenges associated with the use of SM data. In light of these challenges, we present and provide a critical analysis of the potential roles of SM data to enhance survey research. Finally, we propose a novel modular framework for the construction of smart statistics integrating the two data sources, that could serve as a reference framework enabling to compare different applications and provide a common information basis.

2 Characteristics and challenges of using social media data for statistical research

There is a broad definition of SM which includes all websites and apps that allows users to share messages and digital contents (photos, videos, articles, etc.). Social networking, blogs/microblogs, content sharing, and virtual world applications and websites fall into this category. Their use is widespread among people, and they are also well integrated in the business strategy of small and big enterprises. However, SM coverage and usage differs worldwide.

According to the Global Digital Report 2022, released by We Are Social and Hootsuite (2022), the number of worldwide active users of SM, i.e., those who logged-in in the reference period of 30 days, follows an increasing trend and is equal to, on average, 58.4% of the world population. SM audience and rate of usage varies according to regions, age groups, gender and other socio-demographic characteristics. For example, Auxier and Anderson (2021) discuss the use of SM in the Unites States (U.S.). From this study it emerges that some SM are more common among adults under 30 (Instagram, Snapchat and TikTok), Pinterest is more popular among females and the proportion of Instagram users is higher among Hispanic and Black Americans rather than White Americans. Similarly, U.S. Twitter users are younger, more educated, and more likely to be Democrats than the general public (Wojik and Hughes, 2019). In the same report, the authors argue that the majority of tweets is posted by a small share of users.

Thus, a coverage issue in SM data is evident. In addition, users self-select themselves and the data generation process is out of the researcher’s control. SM users can be both real persons, organizations, or Internet robots (BOTs). BOTs are used to automatically publish content online (e.g. advertising) but their use can also be malicious (e.g. spam, fake news, or comments to influence public opinion). Also, the problem of multiplicity of accounts should be taken into consideration when analysing this type of data. For instance, individuals and organizations can have multiple accounts with different purposes.

When retrieving SM data, Application Programming Interfaces are generally employed. Together with data, a set of metadata is delivered with additional information about the content and the user. However, complete socio-demographic information is usually not provided. Data retrieval can be performed through a search query with relevant keywords related to the topic of interest, or the full set of data for a given user can be retrieved. It should be noted that SM posts can be modified or deleted over time, and related metadata can also change (e.g., likes, replies, and shares). Therefore, the results may differ based on the timing of retrieval. Similarly, different formulation of the search query in terms of the specified keyword can result in the delivery of different data.

Once data are retrieved, it is necessary to transform the unstructured data into structured data in order to obtain the information of interest. This transformation can be performed in different ways (e.g. sentiment analysis, topic modelling, supervised classification and clustering, among others). However, the final results might be influenced by the data cleaning, pre-processing and analyses choices (Denny and Spirling, 2018). Thus, it appears evident that also the analysis of SM is susceptible to errors.

Opposite to survey, where the Total Survey Error framework allows the identification and allocation of errors during the whole process, for SM data such a rich and comprehensive framework does not exist. SM sources have different characteristics, which require different quality frameworks. For example, Salvatore et al. (2021) discuss several quality issues related to SM and propose a quality framework for the analysis of Twitter data, and Amaya et al. (2021) describe specific features and issues related to the analysis of Reddit data.

When the objective of the analysis is the integration or augmentation of surveys with SM data, these aspects are even more important. Indeed, it is crucial to understand how errors arise, accumulate, and interact during the entire integration process (De Waal et al., 2019). Biemer and Amaya (2020) propose an error framework to evaluate the quality of integrated datasets generated using survey and non-survey data, and of the resulting hybrid estimates. In understanding the roles of SM in survey research, such characteristics and statistical challenges should be taken into account.

3 The roles of social media in survey research

The roles of SM in survey research have evolved through the years. In this respect, we could identify three main approaches on ways SM can enhance survey research to augment and improve the available information, namely, SM can be used as a replacement for surveys, as a supplement for surveys adding to the richness of the data, or to improve survey estimates.

First, as recently as ten years ago, when research on SM for social sciences began spreading, there was a lot of excitement about the possibility of replacing surveys with the study of SM. In many studies, correlations and alignment between indexes and statistics obtained from traditional surveys and from SM data have been demonstrated. For example, one of the most influential applications is the study by O’Connor et al. (2010), where the authors show that their novel and SM-based consumer confidence index was aligned with traditional indexes including the Gallup’s Economic Confidence Index and the University of Michigan’s Index of Consumer Sentiment. Similarly, Antenucci et al. (2014) construct an index of job loss showing that it was aligned with the Department of Labor’s Initial Claims for Unemployment Insurance. Ceron et al. (2014) demonstrated the ability of SM to forecast electoral results.

Despite the initial promising results, issues underlying SM data made it clear that using SM data as a replacement for surveys is very difficult. For instance, Conrad et al. (2015) replicated the analysis by O’Connor et al. (2010) until 2014 showing a degradation of the relationship between SM and traditional indexes after 2011. Similarly, the Social Media Job Loss index (Antenucci et al., 2014), starting from mid-2014 began to diverge to the actual claim for unemployment[4].

As a consequence, researchers started investigating conditions under which alignment is possible. Conrad et al. (2021), after several experiments on the original O’Connor et al. (2010) analysis, conclude that the relationship between the data was “more than a chance occurrence”. More

[4] http://econprediction.eecs.umich.edu/

importantly, they demonstrated that micro-decisions in the analysis can potentially strongly affect the results. In a similar direction, Pasek et al. (2018) argue that at the current time SM data may "only be fit for purpose in replacing survey data under very limited conditions".

Hence, if considering SM as a substitute for traditional surveys may be ambitious, a more plausible scenario is their use as a supplement. This is a quite recent and growing research area. In this respect, SM data can be collected passively and analyzed to investigate content shared by users on SM platforms and also uses of SM (connections, activities, etc.). This approach involves the inclusion of SM derived variables into statistical models based on traditional data (Bughin, 2015) or the linkage between survey units and SM-accounts (Al Baghal, 2020, Al Baghal et al., 2021). Referring to the case of linking survey respondents' SM profiles to their survey responses, Murphy et al. (2019) supplement survey data with respondents' Twitter postings, networks of Twitter friends and followers, and information to which they were exposed about e-cigarettes, finding the combined data to provide broader measures than either source alone. Recently, Salvatore et al. (2022) use Twitter data to augment traditional data on businesses to study Corporate Social Responsibility (CSR), by adding variables related to Twitter communication of CSR and building indicators based on the two data sources. A related concern is that such approaches could be used for a limited part of the sample, namely for respondents having SM, consenting to link these data, and providing correct SM handles to correctly link the data. In this direction, Al Baghal et al. (2020) explore the feasibility of linking Twitter SM data to survey responses in three British representative panels, with findings suggesting that consent rates for data linkage are relatively low and depend on mode. It is worth noticing that, in case of businesses, identification of accounts through the websites is much easier compared to individuals.

Another interesting and novel potential of SM as a supplement in survey research, is their use in generating qualitative insights. We could think at SM datasets as similar to data gathered from a huge focus group, in that there are comments generated by a broad range of stakeholders who have self-selected into discussing the topic and may display a broader range of opinions than a small focus group (Chen and Tomblin, 2021). In this respect, one also needs to consider that people posting on SM differ in many ways from a focus group that is moderated, topic-focused, and co-present.

Regarding the third approach, namely to use SM data to improve estimates, SM data can be used to combat nonresponse or to improve measurement. With respect to possible use for combating nonresponse, SM accounts represent a stable point of contact for an individual which tends not to change over time. Thus, linking survey respondents to their SM profiles makes them very attractive in longitudinal surveys. In such cases, when a respondent drops out of the study, some information can be retrieved from passive data collection from her/his SM profile. Again, innovative methods of participant engagement and tracing can take advantage of SM networking services, particularly Facebook (AAPOR, 2014; Calderwood et al., 2021). As for item nonresponse, adjustment methods can exploit data from linked SM accounts. Further, this additional data sources can be used as a way of completing otherwise missing items from surveys or to reduce response burden.

With respect to measurement error, Burnap et al. (2016) infer that SM very likely reduce the social desirability bias that affects respondents in a formal survey interview setting. In forecasting the outcome of 2015 UK General Election, the authors find significant support for right-wing parties on Twitter, contrasting the typical underestimation regarding the right-wing vote in the UK. This suggests that linkage of the two data sources can be used to improve measurement in surveys. In the longitudinal context, SM data can provide further information to the nature of change between panel waves, which is of particular interest and likely spurious. Further, SM data can be used to evaluate survey responses. In this direction, Henderson et al. (2021) use tweets to validate survey responses, comparing survey responses to observed behaviour in order to assess the validity of self-reported frequency of posting to Twitter, retweeting content, sharing photos, sharing videos, and sending direct messages. They find variation in the quality of self-reports across types of Twitter activity concluding that relying on self-reported SM behaviour distorts inferential results from what is found when relying on observed SM behaviour. Guess et al. (2019), linking survey data collected during U.S. 2016 election campaign with respondents' observed SM activity, validate self-reports of SM activity, finding that they are correlated with observed behaviour. However, they also find substantial discrepancies in reporting at the individual level.

Given the different possible uses of SM data in survey research, a common underlying conceptual and theoretical framework to guide enhancement of survey research through SM data could be extremely useful to provide a common and comparable basis of information.

4 A modular framework to produce smart statistics

In this section we present a modular framework that can be applied to produce smart statistics and indicators, i.e., generated by augmenting traditional data with SM data. The same approach is discussed in greater detail and applied in the specific field of smart business statistics in Salvatore et al. (2022).

We propose a modular methodological framework organized into three layers, each of which defines the tasks and the outputs. In this paper, we focus on the case of composite indicators as the basis for augmentation. Its structure has been inspired by the modular organization into three layers introduced by Ricciato et al. (2020).

The first layer involves the collection and transformation of data into structured data. Such data and their relative metadata represent the input for the second layer. This second block consists of extracting innovative statistical information and indicators. Using new elementary indicators based on textual unstructured data, the first and second layers enhance the statistical information. In the third layer, innovative statistics and indicators can be used to complement traditional source datasets through linkage and/or statistical integration or combined with existing indicators. As a result, Smart Statistics are produced. Figure 1 summarizes the framework.

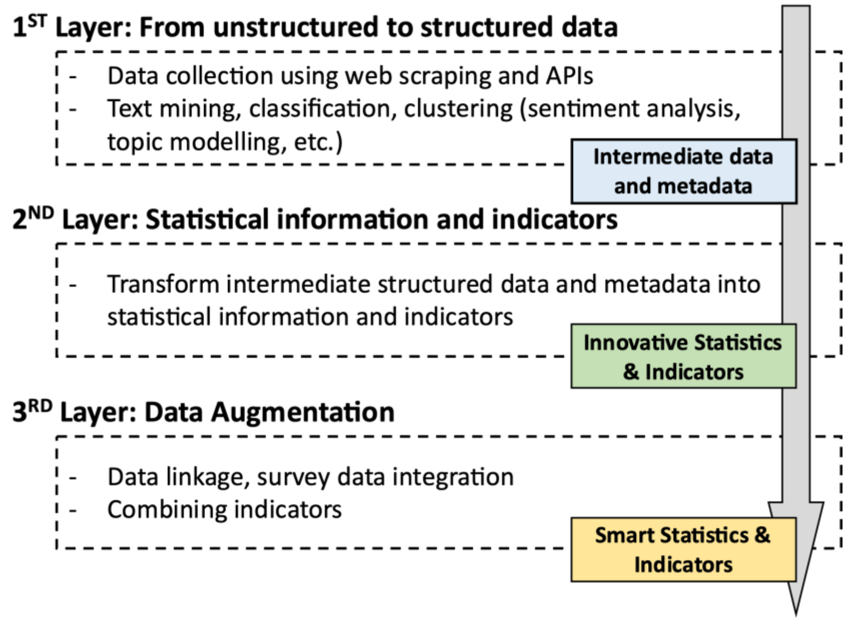
1ST Layer: From unstructured to structured data
- Data collection using web scraping and APIs
- Text mining, classification, clustering (sentiment analysis, topic modelling, etc.)
Intermediate data and metadata
2ND Layer: Statistical information and indicators
- Transform intermediate structured data and metadata into statistical information and indicators
Innovative Statistics & Indicators
3RD Layer: Data Augmentation
- Data linkage, survey data integration
- Combining indicators
Smart Statistics & Indicators

Figure 1. Modular methodological framework for producing smart statistics and indicators

As new and complex data sources are integrated with traditional ones, the modular approach can be followed. Modularity also facilitates exploration of other methodological variants (instances) within the same methodological architecture, plus possible improvements to specific modules or testing of the sensitivity of the results. As an added benefit, when enhancing survey data with SM data, the researcher may proceed across all three layers or may only refer to specific layers depending on information already available.

5 Discussion

The use of SM data to enhance survey research is a recent field of study with a wide range of prospective applications. We have provided considerations on the current state of research. Overall, the results are promising and the potential of these approaches is evident. In this respect, we envision a future survey world "that uses multiple data sources, multiple modes, and multiple frames" to enhance research (Lyberg and Stukel, 2017). SM data will clearly be part of this process, leveraging different ways to advance the survey research paradigm.

Many open problems remain, related to issues inherent in SM data and its data generation process, particularly selectivity, coverage, availability of information about users producing the data, data accessibility and volatility among others. Further, besides ethical and privacy considerations, linking survey and SM data about individuals requires identification of the accounts of interest and obtaining consent to collect and use this data. Also, understanding what questions may be more easily answered by passively collected SM data can help supplement traditional methods of survey data collection. The research agenda will have of course to deal with these issues, keeping quality considerations at pace of the developments.

A lot of experimentation is still needed to provide concrete results and to further explore the extent of the benefits of the use of such data. In this direction, the proposed modular framework will be very useful in structuring experimentations in a stepwise and common way so as to facilitate comparisons and have a broad common base of experimental reference results. Modularity will even allow experimentation focusing also only on a single module benefiting with already known results about other modules. We thus advocate the adoption of such modular framework in future experimentations.

References

Al Baghal, T. (2020). Linking survey and social media data*. Understanding Society Working Paper Series*, No. 2020-04.

Al Baghal, T., Sloan, L., Jessop, C., Williams, M. L., & Burnap, P. (2020). Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review*, 38(5), 517-532.

Al Baghal, T., Wenz, A., Sloan, L., & Jessop, C. (2021). Linking Twitter and survey data: asymmetry in quantity and its impact. *EPJ Data Science*, 10.

Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2021). New data sources in social science research: Things to know before working with Reddit data. *Social Science Computer Review*, 39(5), 943-960.

American Association for Public Opinion Research. (2014). Social media in public opinion research: *Report of the AAPOR Task Force on emerging technologies in public opinion research*. Retrieved December, 6, 2022, from https://www.aapor.org/AAPOR_Main/media/MainSiteFiles/AAPOR_Social_Media_Report_FNL.pdf

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., & Shapiro, M.D. (2014). *Using social media to measure labor market flows* (No. w20010). National Bureau of Economic Research.

Auxier, B., & Anderson, M. (2021). Social media use in 2021. *Pew Research Center*. https://pewresearch-org-preprod.go-vip.co/internet/2021/04/07/social-media-use-in-2021/

Bail, C.A., Guay, B., Maloney, E., Combs, A., Hillygus, D.S., Merhout, F., Freelon, D. & Volfovsky, A. (2020). Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. *Proceedings of the national academy of sciences*, 117(1), 243-250.

Biemer, P.P., & Amaya, A. (2020). Total error frameworks for found data. *In: Big data meets survey science: a collection of innovative methods,* (eds. C.A. Hill, P.P. Biemer, T.D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, L.E. Lyberg), Wiley, Hoboken,131-161.

Brick, J.M. and Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The Annals of the American Academy of Political and Social Science*, 645(1), 36–59.

Bughin J. 2015. "Google Searches and Twitter Mood: Nowcasting Telecom Sales Performance." *NETNOMICS: Economic Research and Electronic Networking* ,**16**(1–2): 87–105.

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

Calderwood, L., Brown, M., Gilbert, E., & Wong, E. (2021). Innovations in Participant Engagement and Tracking in Longitudinal Surveys. In: *Advances in Longitudinal Survey Methodology*, (ed. P. Lynn), Wiley.

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(2), 340-358.

Chen K., & Tomblin, D. (2021). Using data from Reddit, public deliberation, and surveys to measure public opinion about autonomous vehicles. *Public Opinion Quarterly*, 85, 289-322.

Conrad, F.G., Gagnon-Bartsch, J.A., Ferg, R.A., Schober, M.F., Pasek, J., & Hou, E. (2021). Social media as an alternative to surveys of opinions about the economy. *Social Science Computer Review*, 39(4), 489-508.

Conrad, F.G., Schober, M.F., Pasek, J., Guggenheim, L., Lampe, C., & Hou, E. (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.H.. & Puts, M.J.H. (2014). Social media sentiment and consumer confidence, ECB Statistics Paper, No. 5, ISBN 978-92-899-1403-1, European Central Bank (ECB), Frankfurt a. M., https://doi.org/10.2866/11606

De Waal, T., van Delden, A., & Scholtus, S. (2019). Quality measures for multisource statistics. *Statistical Journal of the IAOS*, 35(2), 179-192.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. Political Analysis, 26(2), 168-189.

Guess, A., Munger, K., Nagler, J., & Tucker, J. (2019). How Accurate Are Survey Responses on Social Media and Politics?, *Political Communication*, 36(2), 241-258.

Henderson M., Jiang K., Johnson M., & Porter L. (2021). Measuring Twitter use: validating survey-based measures. *Social Science. Computer Review,* 39, 1121–1141.

Hill C.A., Biemer P., Buskirk T., Callegaro M., Córdova Cazar A.L., Eck, A., Japec, L., Kirchner, A., Kolenikov, S., Lyberg, L., & Sturgis, P. (2019). Exploring new statistical Frontiers at the intersection of survey science and big data: convergence at “BigSurv18”. *Survey Research Methods,* 13, 123–135.

Iacus, S.M., Porro, G., Salini, S., & Siletti, E. (2022). An Italian composite subjective well-being index: The voice of Twitter users from 2012 to 2017. *Social Indicators Research*, 161, 471–489.

Japec, L., & Lyberg, L. (2020). Big data initiatives in official statistics. *In: Big data meets survey science: a collection of innovative methods,* (eds. C.A. Hill, P.P. Biemer, T.D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, L.E. Lyberg), Wiley, Hoboken, 273-302.

Luhmann, M. (2017). Using Big Data to study subjective well-being. *Current Opinion in Behavioral Sciences*, 18, 28–33.

Luiten, A., Hox, J., and de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, 36(3), 469–487.

Lyberg, L.E. and Stukel, D.M. (2017). The roots and evolution of the total survey error concept. In: *Total Survey Error in Practice* (eds. P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, B.T. West), 1–22. Hoboken, NJ.O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth international AAAI conference on weblogs and social media*.

Murphy, J., Hsieh, Y.P., Wenger, M., Kim, A.E., & Chew, R. (2019). Supplementing a survey with respondent Twitter data to measure e-cigarette information exposure. *Information, Communication & Society*, 22(5), 622-636.

Pasek, J., Yan, H.Y., Conrad, F.G., Newport, F. and 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(3), 470-492.

Ricciato, F., Wirthmann, A., & Hahn, M. (2020). Trusted Smart Statistics: How new data will change official statistics. *Data & Policy*, 2.

Rill, S., Reinel, D., Scheidt, J., & Zicari, R.V. (2014). PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis. *Knowledge-Based Systems*, 69, 24-33.

Rosales Sánchez, C., Craglia, M., & Bregt, A.K. (2017). New data sources for social indicators: the case study of contacting politicians by Twitter. *International journal of digital earth*, 10(8), 829-845.

Salvatore, C., Biffignandi, S., & Bianchi, A. (2021). Social media and twitter data quality for new social indicators. *Social Indicators Research*, 156(2), 601-630.

Salvatore, C., Biffignandi, S., & Bianchi, A. (*2022*). Augmenting Business Statistics Information by Combining Traditional Data with Textual Data: A Composite Indicator Approach, submitted.

Stier, S., Breuer, J., Siegers, P., & Thorson K. (2020). Integrating survey data and digital trace data: key issues in developing an emerging field. *Social Science Computer Review,* 38, 503–516.

We are social and Hootsuite. (2022). *Global digital report 2022*. https://www.hootsuite.com/resources/digital-trends.

Wojcik, S. and Hughes, A. (2019). Sizing up Twitter users. *Pew Research Center*. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/