Public Opinion and Social Media: Potential and Pitfalls*

Michael C. Smith[1][0000−0002−1200−0855] and David A. Broniatowski[1][0000−0002−3302−9497]

The George Washington University, Washington, DC, 20052
{mikesmith,broniatowski}@gwu.edu

Abstract. Social media are a promising new data source for surveying public opinion. Despite clear advantages, analyses of social media data face some challenges. We seek to elucidate these challenges and draw relevant lessons from more traditional survey techniques. Beyond machine learning approaches, we make the case that social media studies must carefully consider elements of study design, focusing on issues of research validity and providing examples throughout. Common pitfalls and techniques to avoid these are discussed.

Keywords: Social media · Validity · Study design · Public opinion

1 Introduction, Background

Social media are a promising source of data for surveillance of public opinion, with applications to disaster response, public health, and political views [e.g. 10, 7, 9]. Social media possess significant potential to enable large-scale analysis with their breadth and depth of data, lack of response bias, and access to minority viewpoints [e.g, 4, 8, 6, 2]. However, they also possess disadvantages that should be addressed if we are to leverage their potential. Because of their relative novelty, social media's weaknesses are not yet well-characterized [4]. In contrast, traditional public opinion surveys have well-studied limitations including cost, limited coverage of minority populations, and non-sampling biases [e.g., 8, 11]. High quality research takes these weaknesses into account when designing surveys and analyzing results. This proposal makes the case that social media studies must account for similar elements of study design.

[12] define four types of research validity relevant to drawing conclusions from data: statistical conclusion, construct, internal, and external. Whereas social media data are relatively insensitive to threats to external validity due to their being observed "in the wild", conclusions based on this organic data are more vulnerable to the other types [12]. We detail validity concerns and lessons from survey research as they relate to social media data with examples, including using Twitter to investigate vaccine hesitancy and refusal.

* Preparation of this article was supported in part by the National Institute of General Medical Sciences under grant number 5R01GM114771

2 Pilot Results: Internal Validity

For example, not all studies require representativeness, but designs should account for necessary differences as internal validity checks. Public opinion studies decide between representing a population of interest and understanding nuances of a specific group or idea. If appropriate, social media research need not claim representativeness, instead relying on nonprobability sampling [as suggested in e.g., 11]. If such a study aims to study a specific construct, selecting information based on that construct may provide another window into the conversation beyond surveys. For instance, [4] restricted their analysis to vaccine-relevant conversation using a pipeline of keyword and machine-learning classifiers. Building upon this data, we segmented tweets into topics using Latent Dirichlet Allocation [3] and conducted a Gi-* statistical hotspot analysis [5], to determine geospatial locations where Twitter users were most likely to discuss these topics. Here, we do not claim data representativeness – we simply aim to illustrate how vaccine conversations can vary spatially. It is interesting that discussions of the Californian vaccination bill appear in Ohio and the Appalachian area, suggesting a direction for future work; see Fig. 1. Such nonprobability sampling is powerful to understand nuances among ideas and subgroups [1], if appropriately validated.

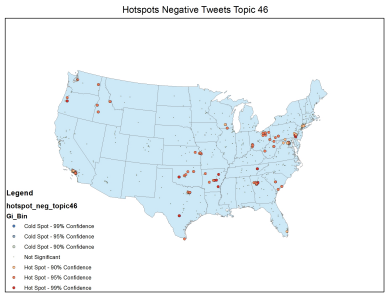

Fig. 1. Geographic hotspots of vaccine-negative discussion w.r.t. the California government's bill eliminating exemptions from vaccinations in schoolchildren. Redder indicates more statistically significant differences from surrounding discussions

3 Discussion

Despite potential challenges to research validity, studies of social media data present a golden opportunity to understand human behavior like vaccine rationales, with actionable policy implications. By leveraging advances in public opinion research for social media, we "are left with the type of natural observation that is valued by many to study behavior without influencing it" [8].

Bibliography

[1] Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J., Tourangeau, R.: Summary Report of the AAPOR Task Force on Non-probability Sampling. Journal of Survey Statistics and Methodology **1**(2), 90–143 (Nov 2013). https://doi.org/10.1093/jssam/smt008

[2] Betsch, C., Brewer, N.T., Brocard, P., Davies, P., Gaissmaier, W., Haase, N., Leask, J., Renkewitz, F., Renner, B., Reyna, V.F., Rossmann, C., Sachse, K., Schachinger, A., Siegrist, M., Stryk, M.: Opportunities and challenges of Web 2.0 for vaccination decisions. Vaccine **30**(25), 3727–3733 (May 2012). https://doi.org/10.1016/j.vaccine.2012.02.025

[3] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)

[4] Dredze, M., Broniatowski, D.A., Smith, M.C., Hilyard, K.M.: Understanding Vaccine Refusal: Why We Need Social Media Now. American Journal of Preventive Medicine **50**(4), 550 (2016)

[5] Getis, A., Ord, J.K.: The Analysis of Spatial Association by Use of Distance Statistics. Geographical Analysis **24**(3), 189–206 (Jul 1992). https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

[6] Groves, R.M.: Three Eras of Survey Research. Public Opinion Quarterly **75**(5), 861–871 (Dec 2011). https://doi.org/10.1093/poq/nfr057

[7] Huang, X., Smith, M.C., Paul, M.J., Ryzhkov, D., Quinn, S.C., Broniatowski, D.A., Dredze, M.: Examining Patterns of Influenza Vaccination in Social Media. In: AAAI Joint Workshop on Health Intelligence (W3PHIAI) (2017)

[8] Murphy, J., Link, M.W., Childs, J.H., Tesfaye, C.L., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M., Harwood, P.: Social Media in Public Opinion Research - AAPOR (May 2014)

[9] O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. ICWSM **11**(122-129), 1–2 (2010)

[10] Salath, M., Khandelwal, S.: Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. PLOS Computational Biology **7**(10), e1002199 (Oct 2011). https://doi.org/10.1371/journal.pcbi.1002199

[11] Schober, M.F., Pasek, J., Guggenheim, L., Lampe, C., Conrad, F.G.: Social Media Analyses for Social Measurement. Public Opinion Quarterly **80**(1), 180–211 (Jan 2016). https://doi.org/10.1093/poq/nfv048

[12] Shadish, W., Cook, T.D., Campbell, D.T.: Experimental and quasi-experimental designs for generalized causal inference. Wadsworth Cengage learning (2002)