Sadaf Zia
Mississauga Library System, Mississauga, ON, Canada

Celina De Lancey
Western University, London, ON, Canada

Priscilla Regan
George Mason University, Fairfax, VA, USA

Jacqueline Burkell
Western University, London, ON, Canada

THERE FOR THE REAPING: THE ETHICS OF HARVESTING ONLINE DATA FOR RESEARCH PURPOSES (Paper)

Abstract

Online social environments offer a rich source of data that researchers can harvest to gain insight into a wide range of social issues. This type of research is sometimes considered as observation of public behaviour, and therefore exempt from ethical review. This type of research, however, raises ethical issues with respect to the public/private nature of online spaces, consent, and anonymity in the online environment. This project examines research ethics guidelines for recommendations regarding the use of harvested online data, identifying best practices for researchers who engage in this type of research.

Introduction

Online social environments, including social media platforms and online discussion groups, provide important platforms for individual expression and interpersonal connection. These platforms also provide researchers with a source of data that can be mined for valuable insight into social issues. The advantages of these data for research are myriad, and include limited costs (time and money) for data collection, access to information on sensitive issues, and an absence of reactivity. At the same time, use of these data, which are generated by and about individuals, warrants careful ethical consideration. Research ethics guidelines, including the Tri-Council Policy Statement (TCPS2) in Canada, provide guidance to researchers and research ethics boards (REBs) tasked with making decisions and recommendations regarding the ethical conduct of research involving human participants. This guidance, however, is variable and often incomplete, with the result that ethical considerations and practices vary substantially between researchers and across institutions. This project examines research ethics guidelines, including the TCPS2 and others, with respect to their coverage of and recommendations concerning research involving the harvesting of social media data. The results summarize the treatment of these issues in guidelines, available in English, that apply to research and researchers in Canada and internationally, culminating with a set of best ethical practices for research that involves this type of data.

Research Ethics Issues and Guidelines

Online social media posts, including tweets, online discussion group interactions, Facebook updates, and listerv archives, provide a rich source of 'naturally occurring, everyday talk' (Sixsmith and Murray, 2001, p. 424) – precisely the kind of data that, according to Potter and Wetherell (1995), offer deep insight into social phenomena. Ethical issues arise, however, with any use of data from or about people, and harvested data from online social interactions is no exception. Researchers and those focused on the ethics of human subjects research have long recognized that online research presents new and specific ethical issues, potentially requiring new ethical approaches (see, e.g., Elgesem, 2002; Eysenbach & Till, 2001; Flicker et al., 2004; Neuhaus & Webmoor, 2012; King, 1996; Swirsky et al., 2014; Taylor & Pagliari, 2018, Vitak et al., 2016). The harvesting of online productions and interactions for research purposes presents particular challenges, revolving around the issues of participant autonomy (including questions of consent), and participant anonymity and confidentiality.

Under many research ethics guidelines harvesting of online social media data is a form of observational research that, if carried out in 'public' venues, may be exempt from ethical review. Questions arise, however, regarding the public/private nature of online spaces – and, if the spaces are deemed 'private' (or at least not so 'public' as to remove the requirement for ethical review), issues of consent and anonymity/confidentiality become paramount. Some research ethics guidelines (e.g., the TCPS2) provide limited specific guidance regarding this type of online research; others, such as the guidelines provided by the Association of Internet Researchers, (Franzke et al., 2020) provide much more focused and detailed recommendations. Researchers and REBs considering the ethics of this type of research would benefit from a comprehensive and integrated summary of the ethical considerations and approaches available in various guidelines: it is exactly such a summary that we present here.

Methodology

A total of 31 research ethics guidelines, written in English, were identified through a combination of methods including Google searches (e.g., for 'research ethics' and 'ethics guidelines), review of research using harvested online data for mention of research guidelines, and review of publications addressing online research methods for cited guidelines, and specific search of the websites of North American universities for ethics guidelines.

Results

Ethical guidelines (e.g., the TCPS2 in Canada) typically allow for the observation of public behaviour without full ethical review. Guidelines also acknowledge that it is critical to consider and respect expectations of privacy on the part of those being observed in deciding whether research requires ethical review. The public/private nature of the data collection site is therefore a key question in the ethics considerations regarding research that uses online harvested data. Some signs that a space might be private include:

- membership requirement for joining the discussion, particularly if membership requires creating a personal profile with detailed and particularly identifying information;
- a sensitive topic of discussion;

- terms of reference or privacy policies that limit research use of the data, or that specifically state the content will not be used for other purposes.

Factors that mitigate against an expectation of privacy include

- terms of reference or privacy policies that explicitly allow the use of the data for research purposes;
- open discussion groups that do not require membership to join;
- searchable archives of the discussions, particularly if these are available on the open web.

If the online space is not determined to be public, a more comprehensive ethical review is warranted, focused on two issues: consent and anonymity.

In general, research ethics guidelines require consent of participants in order to collect data from or about individuals. It can be very difficult to gain consent for the research use of harvested data, particularly because contact information may not be available, especially if the data are harvested some time after they were produced. In cases where data are drawn from a moderated forum, consent can be sought from the moderator as a proxy for participant consent. If data are collected in real time (i.e., not from archives), it is possible to seek participant consent prior to collection. This practice, though, presents its own ethical issues. In particular, when researchers seek to collect data in the context of an ongoing social interaction (e.g., a longstanding online support group), if only some members consent to collection, the result can be that the group is rendered inaccessible to those who do not consent. These and other relevant considerations must be weighed in determining whether explicit consent is required for data collection.

A second consideration is the protection of participant anonymity. When direct quotes from harvested material are used, participants should be assigned pseudonyms by the researcher – even when the original material is posted under a user-selected pseudonym. Enduring pseudonyms (i.e., pseudonyms used over time within or even across sites) are a form of identifier that is particularly important in the online context, and anonymity protection should extend to these chosen names. Pseudonymous anonymity can also be compromised if harvested content is indexed on the open web, potentially leading back to a specific site (e.g., a specific discussion group) and to an individual posting information on that site. The most stringent protection of participant anonymity would therefore require that researchers test anonymity by entering content they intend to quote in a publication or presentation into a search engine to determine if the search returns the source of the quote. In the event that the content is found, or simply to ensure protection against such identification, quotes can be combined or paraphrased to protect participant anonymity.

Conclusion

Researchers wishing to use harvested online data for research purposes must ensure that they respect the autonomy and anonymity of their (often unwitting) ‘participants.’ It cannot be assumed that online spaces are ‘public’, and therefore available for observational research without further ethical consideration. Instead, researchers must carefully consider whether

participants in the space have an expectation of privacy. If the context or the content suggests that the online space is private in nature, additional ethical considerations arise: consent, and anonymity. Where possible, unless disruptive to participation in the online space, researchers should seek participant consent for data collection; in some cases, moderator consent can serve as a proxy. Participant anonymity should be at minimum protected by the use of researcher-assigned pseudonyms. Additional anonymity protection, particularly important if harvested content is searchable on the open web, can be achieved through paraphrasing (rather than using direct quotes), and/or integrating statements from multiple participants.

Reference List:

Elgesem, D. (2002). What is special about the ethical issues of online research? Ethics and Information Technology, 4, 195-203.

Eysenbach, G., & Till, J. E. (2001). Ethical issues in qualitative research on internet communities. British Medical Journal, 323(7321), 1103-1105.

Flicker, S., Haans, D., & Skinner, H. (2004). Ethical dilemmas in research on Internet communities. Qualitative health research, 14(1), 124-134.

Franzke, A.S., Bechmann, Anja, Zimmer, Michael, Ess, Charles and the Association of Internet Researchers (2020). Internet Research: Ethical Guidelines 3.0. https://aoir.org/reports/ethics3.pdf

Neuhaus, F., and Webmoor, T. (2012). Agile ethics for massified research and visualization. Information, Communication & Society, 15(1), 43.65.

King SA. (1996). Researching internet communities: proposed ethical guidelines for the reporting of results. The Information Society, 12(2),119–128.

Potter, J.,&Wetherell, M. (1995). Discourse analysis. In J. A. Smith, R. Harre, & L.Van Langenhove (Eds.), Rethinking Methods in Psychology (pp. 80-92). London: Sage Ltd.

Sixsmith, J., and Murray, C. D. (2001). Ethical issues in the documentary data analysis of internet posts and archives. Qualitative Health Research, 11 (3), 423-432

Swirsky, E. S., Hoop, J. G., & Labott, S. (2014). Using social media in research: new ethics for a new meme?. The American Journal of Bioethics, 14(10), 60-61.

Taylor, J., & Pagliari, C. (2018). Mining social media data: How are research sponsors and researchers addressing the ethical challenges? Research Ethics, 14(2), 1-39.

Vitak, J., Shilton, K., & Ashktorab, Z. (2016, February). Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (pp. 941-953). ACM.