Investigating conditions for consent to analyze social media data

Ansgar Koene[1], Elvira Perez Vallejos[1], Christopher J. Carter[1], Ramona Statache[1], Svenja Adolphs[1], Claire O'Malley[1], Tom Rodden[1], Derek McAuley[1]

[1]HORIZON Digital Economy Research, University of Nottingham, UK

ansgar.koene@nottingham.ac.uk
elvira.perez@nottingham.ac.uk
christopher.carter@nottingham.ac.uk
ramona.statache@nottingham.ac.uk
svenja.adolphs@nottingham.ac.uk
claire.omalley@nottingham.ac.uk
tom.rodden@nottingham.ac.uk
derek.mcauley@nottingham.ac.uk

Abstract: The rapid growth in online social media and the associated internet mediated research have generated a range of new challenges and concerns related to research ethics. A key issue around the development of protocols for ethical approaches to social media analysis is the question if explicit informed consent should be required when conducting research based on messages that were posted on social media. One of the main arguments against such a requirement is the assumption that obtaining explicit informed consent would be unfeasible or at the very least would heavily skew the resulting data in favour of views that follow perceived 'socially acceptable norms'. In this study we propose a citizen centric approach to the question of identifying which types of social media research should require explicit informed consent from the social media users. We are preparing to run a survey to ask a broad range of citizens for which kinds of research, and under which conditions, they would consent to having their social media data analysed for research purposes. In order to understand if, and how, the process of obtaining consent is likely to bias the population sampling of social media studies the questionnaire responses regarding conditions for consent will be correlated with demographics information regarding the politico-socio-geographic-economic background of the respondents. In hopes of obtaining maximally representative sampling across the UK population the survey will be promoted through news items in the popular press.

The goal of this questionnaire study is to obtain practical evidence about the concerns of the social media users whose data is analysed in social media based research. It is our hope that this data will provide insights for the development of practical guidelines for the acquisition of social media data for research purposes. We are especially interested in clarifying the kind of transparency/information that prospective participants desire to know about, so that we might practically improve the success rate of ethical participant recruitment for research studies.

Keywords: Informed consent, Social Media, Privacy, Anonymity, Ethics, Internet Mediated Research, Survey

1. Introduction: Increasing numbers of research studies, both academic and non-academic, are looking to tap into the rich data about human behaviour, psychology and society that is generated by people's activity on social media. Some of the core features that make online social media platforms attractive for researchers include the comparatively low effort and cost associated with data acquisition and the unobtrusive nature of the acquisition process, which can often be performed 'behind the scenes' using application programme interfaces (APIs) or web scraping techniques, depending upon the affordances of the specific type of social media studies (e.g. Twitter, Blogs). In either case the researchers are making use of the fact that these internet based platforms, by their very nature, inherently record all communications. The acquisition of data thus becomes a matter of gaining access to an already existing data store, rather than a process of data capturing. While the unobtrusive nature of the data acquisition methods offers the advantage of avoiding the Hawthorne effect (McCarney et al 2007), ensuring that observed conversations are not influenced by an awareness of being observed, it raises ethical concerns around issues of privacy violation related to the use of insufficient informed consent procedures that do not match the requirements that social media users feel they are entitled to. The degree of entitlement for control of social media data by its users is however frequently not clear as it is often difficult to establish if a specific online communication should be considered to be in the private or the public domain (BAAL, 2006). Even when dealing with social media, such as Twitter, which most people recognize as being a 'public broadcast' platform the distinction between public and private communications can be challenging. Despite the acknowledged broadcast nature of the platform it is nevertheless often used as a means for communication within networks of friends, with little intention of broadcasting content to a wider audience. In such instances, social interaction upon Twitter might be viewed more like a private conversation in a public space rather than radio broadcasts.

Regardless of the public of (semi-)private nature of social media communications, most research that uses social media data is arguably extremely benign with effectively no impact on the people whose social media communications are used in the study. Academic research almost always focuses on establishing general trends in behaviour, or fundamental patterns of communication which are used to establish or test general theories, not criticize or manipulate individual social media users. In such cases it might be argued that it doesn't matter if the data was public or private, or if social media users gave explicit informed consent for having their data included in a study. Viewed from the perspective of the social media users, however, the requirement for informed consent is primarily an issue of respect for the autonomy and dignity of persons and as such does not depend on the impact that the research outcomes will have. In many respects, the requirement for informed consent represents a core value of any democratic society, and yet it is probably the principle that is most frequently violated on-line. An illustrative example of the conflict of perceptions about the requirement to seek consent was the controversy following the Kramer et al. (2014) publication on research using Facebook, for which they asserted that no explicit consent needed to be obtained from the Facebook users since "it [the study] was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research" (Kramer et al., 2014). The Data Use Policy however, did not provide any specific information about the nature of the 2014 Kramer et al study. The common view by the press, the public and various academics who commented on the controversy was therefore that the Data Use Policy was not acceptable as a means of gaining informed consent. The assertion by Kramer et al (2014) that explicit consent was not required because a general consent had already been given was, however, basically the same as the rationale for data from 'public' forums not requiring consent because users of the forum have in effect already consented when they accepted the 'terms and conditions' of the 'public' forum. Furthermore, the current reality of Internet usage is that the 'terms and conditions' policies of Internet sites are rarely read and are generally formulated in ways that are too vague and incomprehensible to constitute a means of gaining true informed consent (Luger, 2013). Faced with these dilemmas concerning the justification for using social media data without requesting explicit consent from the platform users, the best solution might be to require that all studies should obtain informed consent by contacting the users and providing (electronic) information and consent forms similar to the standard laboratory experiment procedure. There are however doubts concerning the practical feasibility of such an approach.

In order to gain a better understanding of the feasibility of obtaining informed consent from social media users, we are developing a questionnaire study to ask under which conditions, and for what kinds of research, citizens would be willing to grant consent for having their data included in a research study. As part of this study we also hope to estimate if, and how, the consenting population might be biased towards specific demographic/ideological subsections of the general population.

2. Method: The study we are running to investigate the feasibly of obtaining explicit informed consent for social media data access is based on a questionnaire consisting of three sections (currently under review for submission to the departmental ethics review board).

The first section focuses on demographic information, including basic factors such as gender, age, place and employment. We further also include some more ideological factors such as political affiliation and finally some intellectual factors such as level of education and computer literacy.

The second section focuses on identifying the specific conditions which participants would demand in order to give permission to access their data. These include information about the research topic, the research team, the way in which the data will be used and if the research is for academic, corporate, government or other purposes. In recognition of the potential differences in perceived value/sensitivity of data at various types of social media, we include separate questions for Blogs, micro-blogging (e.g. Twitter) and Social Network Sites (e.g. Facebook). We also explore how the conditions under which users would provide consent might differ depending on the type of institution that is doing the research, e.g. academic, government, corporate or third sector. Participants are asked to indicate the importance (highly important, mediam importance, low importance, irrelevant) of various levels of transparency concerning the research questions, methodology and the research team. The questions statements are:

- General willingness to share
- Know enough about the researchers (background and motivations) so I can trust them
- Clearly understand what the study is about
- Clearly know which data the study will analyze
- Have clear information about how the data will be analyzed
- Have clear information about how the results will be reported
- Am given pre-publication access to the results
- Am paid for my data

The third section focuses on the nature of the research questions, to map changes in participant's willingness to consent as function of the topic of study. The types of research questions are roughly split into pure-academic, applied-academic, socio-political and corporate-product development related. The hypothetical research questions are:

- A study about estimating psychological traits of social media users based on their postings.
- A study related to prevalence and types of cyber bullying.
- A Linguistics study about context dependent changes in word usage.
- A study to identify indicators of mental health problems through social media communication patterns.
- Comparison of work/school related communications on public and private social media platforms.
- A study on information propagation through friend networks.
- Analysis of advertising efficiency based on identifying how and when products are talked about on social media.
- Analysis of social media posts to improve targeted advertising.
- A study to track if and how a government initiated media campaign on healthy living is responded to on social media.
- Study to track how different political parties are commented on in social media, as part of election polling.

For each of these scenarios participants are asked to indicate if they would:

- refuse consent
- unlikely to consent
- possibly consent
- likely to consent
- definitely consent

In order to obtain a representative sample of the population, we are aiming to recruit participants by publicizing the questionnaire in a range of public news media.

3. Conclusion: In the current climate of research based on social media data, where stories of unethical behaviour for commercial or security related gain are flooding the mainstream media, academia has a responsibility to enter into the discussion of what constitutes good, ethical research conduct. This may be achieved by being transparent about the goals and methods of the research and gaining true informed consent from participants. The results from our questionnaire on the conditions under which people would be willing to consent to having their social media data used for research, will hopefully provide a citizen centric perspective on the debate about how and when explicit informed consent is required for accessing social media data. By better understanding the type of information prospective participants would want in order to satisfy their concerns about how their data will be used, it will be easier to establish the trust relationships that are required to allow internet mediated research to continue and grow as a research methods. Without such a trust based relationship internet mediated research, especially involving social media data, runs the risk of alienating the social media users and triggering a backlash that could result in highly restrictive regulations analogous to the controversy and subsequent restrictions on genetically modified crops in the EU.

References

McCarney R, Warner J, Iliffe S, van Haselen R, Griffin M, Fisher P, (2007). "The Hawthorne Effect: a randomised, controlled trial". BMC Med Res Methodology 7: 30. doi:10.1186/1471-2288-7-30.

BAAL (British Association for Applied Linguistics) (2006), “Recommendations on Good Practice in Applied Linguistics”. Available online at http://www.baal.org.uk/dox/goodpractice_full.pdf

Kramer, A.D.I., Guillory, J.E., Hanckock, J.T., (2014). “Experimental evidence of massive-scale emotional contagion through social networks”. PNAS 111, 24, 8788-8790.

Luger, E., 2013. Consent for all: Revealing the hidden complexity of terms and conditions. Proceedings of the SIGCHI conference on Human factors in computing systems, 2687-2696.