Social Media Research Ethics

The public is not expecting policymakers to be listening to conversations in publically owned channels, so when does the collection of social media insights become creepy?

How can we use existing evidence about attitudes to personal data provided online in order to create a set of guidelines for our project?

Final v1.0

May 2016

Fraser Henderson & Remmert Keijzer

Co-funded by the
Erasmus+ Programme
of the European Union

Contents

1.0 Purpose

The DEEP Linking Youth Project collects and analyses social media data to generate subsequent insights for policymakers. This requires ethical consideration. As such, this paper describes how the project will attempt to perform social listening with integrity and maintain a sense of responsibility when dealing with personal data in line with fair expectations of European citizens.

2.0 Context

Surveillance is almost universal through the internet, such as the use of Cookies to generate personalised choices. Most citizens are simply unaware of when they are being monitored, by who and for what purpose. Despite this, under the terms and conditions of the major social networks, data can currently be used by third parties for any number of reasons - including analysis for research.

There is mixed awareness of the fact that citizens agree to forego their content property rights and privacy as a term of membership. Yes, the conversation is public – but meant for an intended audience; so when others listen it may be considered impolite or feel intrusive.

Hiding surveillance so that people cannot avoid it constitutes removal of choice and diminution of freedom. Thus it is possible to argue from this perspective that the lack of choice to avoid surveillance constitutes coercion.

The use of third party digital data (e.g. loyalty card transactions) by government is arguably more intrusive than listening to conversations presented in the public domain. However, the idea that an official is listening can be perceived as 'creepy' and the project should not aim to alter citizens' behaviour as a result of being heard. *We therefore promote the idea that listening, because an organisation wants to be responsive and to engage better, is not the same as surveillance.*

Government needs to be open and clear that it is not listening to anything not already in the public domain. A clear and transparent methodology is vital; this can be achieved by making the process clear and sharing the outcomes. For example, stating whether or not a social media feed is monitored is one way of heightening transparency while alerting participants to the fact that their contributions may be analysed.

3.0 Public Attitudes

3.1 Variance in general attitudes

It is reasonable to assume that public attitudes towards privacy and data gathering online vary across the member states. A TNS survey commissioned for the EU in 2015 consisting of views from 27.980 respondents across 28 member states revealed the following headline findings relevant to our project[1]:-

- Half of Europeans have heard about revelations concerning mass data collection by governments. Awareness ranges from 76% in Germany to 22% in Bulgaria.

- A large majority of people (71%) still say that providing personal information is an increasing part of modern life and accept that there is no alternative other than to provide it if they want to obtain products of services.

- Roughly seven out of ten people are concerned about their information being used for a different purpose from the one it was collected for.

- Only a fifth of respondents fully read privacy statements (18%)

- Around 45% of respondents say they are concerned about the recording of everyday activities on the Internet.

In terms of concerns about not having complete control over the information citizens provide online, at least seven out of ten people express concern in 12 countries. These are citizens primarily situated in Western, Southern and parts of Eastern Europe. Specifically, the level of concern is greatest in Portugal, Ireland and the UK (all 79%).

The lowest levels of concern can be observed in the Nordic and Balkan countries, as well as in parts of Eastern Europe. Overall, the lowest proportions can be seen in Estonia (38%), Sweden (41%) and the Netherlands (47%).

The socio-demographic data show that people aged 55 and over are somewhat more likely than those aged 15-24 to feel concerned about not having complete control over the information they provide online (72% vs. 64%).

3.2 Views on using social media data for research purposes

Working on a 'worst case' basis in terms of the sensitivity of citizens to personal data concerns (as detailed above) we have considered some more in-depth research published in the United Kingdom.

Specifically, research from IPSOS Mori published in November 2015 reveals that a majority (60%) of those asked in the United Kingdom say they don't think social media companies should be sharing their data with third parties for research purposes[2].

[1] http://ec.europa.eu/public_opinion/archives/ebs/ebs_431_en.pdf
[2] https://www.ipsos-mori.com/researchpublications/publications/1771/Ipsos-MORI-and-DemosCASM-call-for-better-ethical-standards-in-social-media-research.aspx

The IPSOS report continues to suggest that public expectations about the privacy of their social media data aren't being met by current practice. For example:-

- Nearly three quarters (74%) would prefer to remain anonymous if a social media post was selected to be published in a research report.
- Over half (54%) agree that all social media accounts have the right to anonymity in social media research, even if the account is held by a public institution, private company or high profile individual.
- Nearly a third (32%) still thought that social media companies should not disclose high level data, such as volume of posts on a particular subject, even if this information is not attributed to individuals.

By contrast, a recent survey conducted by the Collaborative Online Social Media Observatory (COSMOS) into users' perceptions of the use of their social media posts found that 82% of those surveyed were 'not at all concerned' or only 'slightly concerned' about *university researchers* using their social media information[3].

The COSMOS survey went on to reveal that:-

- 94% were aware that social media companies had Terms of Service.
- 33% had read the Terms of Service in whole or in part.
- 73% knew that when accepting Terms of Service they were giving permission for some of their information to be accessed by third parties.
- 56% agreed that if their social media information is used for academic research they would expect to be asked for consent.
- 77% agreed that if their tweets were used without their consent they should be anonymised.

[3] https://www.cs.cf.ac.uk/cosmos/

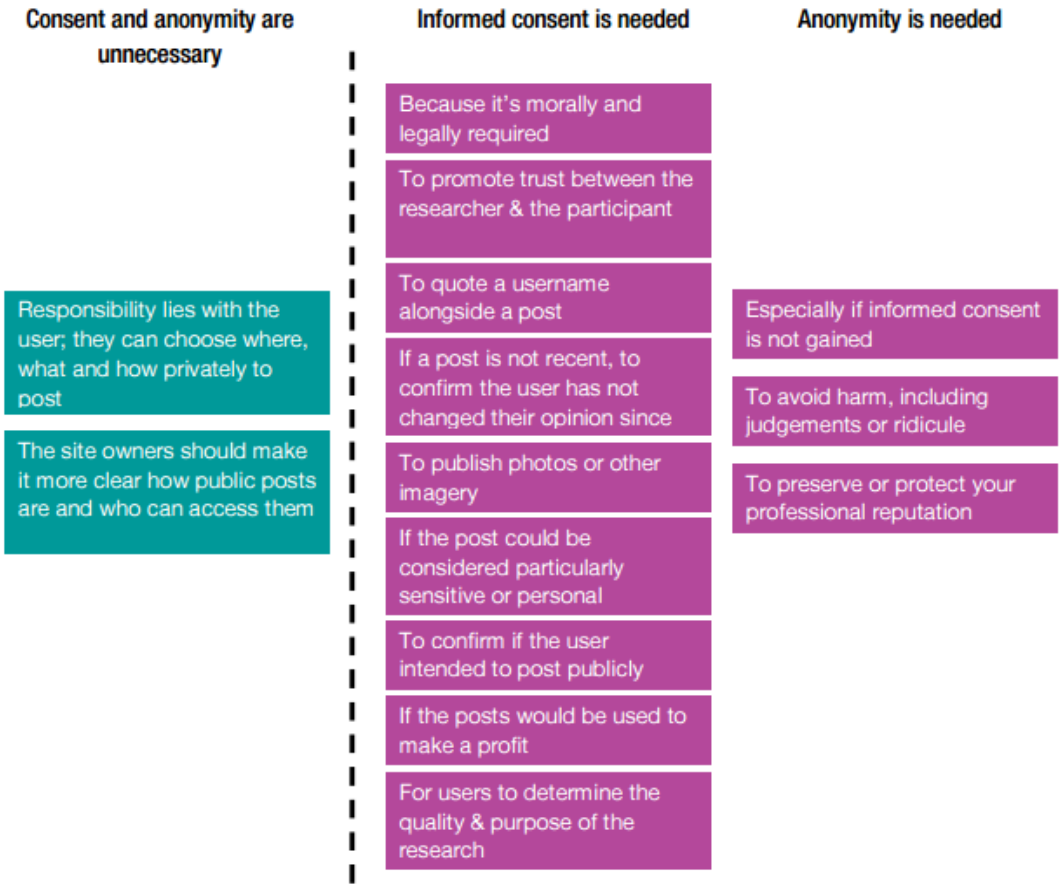
3.3 Consenting use of personal data

How social media research is perceived is likely to depend on a number of other factors such as the original intent and contextual nature of the research. For example, why the project has been commissioned and the topic of content to be explored. In other words, who is conducting the listening and for what purpose.

Research by the independent social research agency NatCen suggests that four factors of research context influenced participant's views and expectations of informed consent and anonymity[4]:-

1. Mode and content of a post, including written content, photos and the sensitivity of the content,
2. Social media website being used,
3. The expectations the user had when posting,
4. The nature of the research, including the organisation the researcher was affiliated with and the research purpose.

The arguments from this report for consent and anonymity are mapped below:-



[4] http://www.natcen.ac.uk/media/282288/p0639-research-using-social-media-report-final-190214.pdf

Further to this, a 'mini survey' was conducted by the Citizen-centric Approaches to Social Media Analysis (CaSMa) which investigated the relationship between consent and factors relating to approach. Although the sample size was small, 58% of respondents to this survey said that consent depends on the purpose of the research[5]. Of these, the priorities were:-

1. How well results will be published (6)
2. Research questions (6)
3. Research methods (6)
4. How well the data is anonymised (5)
5. Who is doing the research (3)

3.4 Synthesis

To conclude, there is only weak support for using publically available social media data for the purposes of research. However, context is important and of those citizens that are aware of the possibilities, most are concerned yet accepting of the situation on the basis that there are outweighed benefits. Anonymity is important to citizens for non-consensual use of social media data.

Subsequently, we interpret that citizens:-

1. Feel to have 'lost control' of the how their data is being used, and feel under-informed about the fact this social media research is even happening.
2. Would prefer to remain anonymous if a social media post was selected to be published in a research report.
3. Agree that all social media accounts have the right to anonymity in social media research, even if the account is held by a public institution, private company or high profile individual.

We are unlikely to be able to tackle prior notification of intent as we do not own the monitored spaces but there are a number of key questions which can be formulated on these principles. For example:-

- Are the sources public (open to all?)
- Can any harm be done by aggregating this information?
- When data is collected, are individuals identifiable?
- Does the process invade personal privacy or cause any obvious harm?

[5] http://casma.wp.horizon.ac.uk/2016/01/10/casma-going-forward-into-2016/

4.0 Public value of social media research

DEEP-Linking youth is clear about the shortcomings of this research method, such as the errors and tolerances of natural language processing. However, the public value of social media research is likely to correlate with the policymaker value of social media research.

Research from NatCen revealed that citizens' views about research using social media fell into three categories: scepticism, acceptance and ambiguity[6]. Views varied greatly depending on the research context, and on a participant's knowledge and awareness of social media sites. Participants expressed concerns about the quality of social media research associated with validity and representativeness.

For example, there were concerns that people behave differently online and offline and so online research could not reflect the 'real world'. Exaggerated views were a result of the anonymity the internet afforded and therefore research findings using views from online sources would lead to inaccurate conclusions about something or someone.

Impulsive comments posted online may result in data gatherers using a view that does not accurately reflect someone's 'normal' viewpoint but instead only something they held for a moment in time. Finally, inaccurate profiles taken without further context could lead to inaccurate information and findings.

Our view is that there is a genuine research question which is: is an online avatar human?

[6] http://www.natcen.ac.uk/media/282288/p0639-research-using-social-media-report-final-190214.pdf

5.0 Suggested code of conduct

Our code of conduct below stems from the public attitudes in section 3.0 and similar codes such as the 'big boulder initiative' and ESOMAR guidelines[7,8]. We have appointed an ethics officer to own these rules and use them accordingly to safeguard project data.

General Principles

1. No deception (research in the guise of marketing).
2. Use 'opt-out' where citizens can contact us to have their data withdrawn from future social media analysis.
3. Avoid harm to the data owner.

Data collection

1. Treat data collected from suspected under-16s as sensitive.
2. No use of 'wall garden' content (where the data gatherer mush join or register a network). This code applies to public social media channels only.
3. Inform participants when an 'owned' digital channel is being used for research purposes.
4. Minimise the collection and analysis of unnecessary 'meta-data', such as location data or the username or @ handle, where this information is not necessary for the project.

Data handling

1. Do not keep data for longer than it is needed. This is because there is an increased risk that the information will go out of date but also poses an increased risk that it will not be held securely. We suggest a maximum retention period of 6 months.
2. Keep all raw data securely, e.g. by using encryption or password protection.
3. No transfer of ownership or sale of collected data to third parties.

Attribution

1. Do not use quotations or material that could be traced back to individuals. This may mean broadcasting a post without attribution, or with a blurring of the name and preserving original context so as not to surprise the originator (see 'masking' in section 5.0).
2. Quotes from accounts maintained by *public organisations* (e.g. government departments, law enforcement, local authorities, national press and broadcasters) are allowed *without seeking prior informed consent.*

[7] http://blog.bigboulderinitiative.org/2014/11/14/draft-code-of-ethics-standards-for-social-data/
[8] https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-Guideline-on-Social-Media-Research.pdf

Reporting

1. Publish results (or a subset thereof) openly to demonstrate transparency around how the data is being used.

Engagement

1. Research participants will be protected from unnecessary and unwanted intrusions and/or any form of personal harassment.
2. Be clear about our aims and role as a research project.

5.1 Masking

The degree of masking required will depend on the nature of the comment and its author. Masking can be applied in varying degrees such as just changing the odd word through to altering key features of a comment. It is the responsibility of the data gatherer to decide the most appropriate degree of masking. Factors to take into account include:

- If the topic being discussed is sensitive or personal,
- If abusive or aggressive language is used,
- If it includes anything against the law,
- If it includes anything embarrassing or is likely to impact career opportunities,
- If it includes any personally identifiable information about the participant or others (except when it is about a well-known person in the public domain and it is not libellous),
- If it includes any data about others that is not already public. In the case of public pictures or videos, consideration should be given to techniques such as pixilation of faces, where masking is required.

6.0 Abridged code of conduct statement

We are only concerned with data posted to *public sources* and will respect the privacy of *individuals* for data which is later harvested for our research, avoiding attribution and third party use. We will only use collected data for our own, *non-commercial* purposes and seek to *actively* raise awareness of our activities in any particular digital channel where monitoring occurs.