Guidance document for research involving social media data

Introduction

Social media refers to any online platform that allow users to create and share content, or to participant in social networking, for example Facebook and Twitter, video sites such as youtube, online dating sites, online messaging services such as WhatsApp and blogging sites. Usage of these sites has grown rapidly over the last few years with an increasing number of people using social media for socializing, networking and expression of thought. These sites can therefore be a rich source of data and provides researchers with the opportunity to gather data sets that would otherwise take significant time and/or resources to obtain [(1)].

Types of data that can be collected from social media includes but is not limited to:

- Content created by users (blog posts, photographs, videos, comments, tweets etc.)
- Social network data (friend lists and followers etc.)
- Data on engagement with content (likes, shares, retweets etc.)
- Other data, such as location data, that the users may not actively post, but may still be collected by the site.

All research should be guided by the ethical principles of respect for persons, beneficence, justice and non-maleficence, and research involving social media is no exception [(2)]. However, using social media for research brings up different contextual problems than traditional research approaches, and therefore there are additional considerations to take into account.

Due to the changing nature of social media, it is not possible to provide a strict set of procedures for researchers to follow, instead the aim of this guidance is to highlight some of the main concerns surrounding the use of social media data for research while still encouraging researchers to use their professional judgement when considering these concerns in relation to their individual research study. If additional advice is required please contact Ethics@lshtm.ac.uk.

Any research involving human participants, their tissue and/or their data, must be referred to, and approved by, the relevant LSHTM Research Ethics Committee.

Social media users are considered human participants if they are being observed or having their data used for research purposes.

Data collected directly from social media sites is considered primary data and this should be reflected in the application to the ethics committee.

1. Public vs Private

One of the biggest areas of concern when using social media data is determining to what extent the data is considered public or private.

The main argument for considering social media data public is that users agree to a set of terms and conditions for each social media platform they use, and these terms and conditions often contain clauses on how user's data may be viewed and processed by third parties, including researchers. However, this argument is problematic as several studies have highlighted that while users have agreed to these terms and conditions, many do so either without reading or fully understanding the consequences of what they are agreeing to [3, 4].

Studies have also suggested that there can be a mismatch between user's expectations of privacy in their online interactions, and the reality of privacy [5]. In one study, every single Facebook user-participant confirmed at least one inconsistency between their sharing intentions and their actual privacy settings [3]. So it can be unclear to what extent the privacy intentions of users align with the actual privacy settings applied.

This creates a complication for researchers as the ethical principle of respect for persons emphasises the importance of recognising people as autonomous individuals with the right to make their own decisions. The principle requires that individuals be empowered to make free decisions and be given the information needed to make informed decisions. If, as the evidence suggests, a large number of social media users either do not read, or do not fully understand, the terms and conditions that they are agreeing to it would be difficult for a researcher to argue that an informed decision was made by the user to put their data into the public domain.

Researchers should therefore act with caution when deciding to whether data collected from social media platforms is fully in the public domain.

The British Psychological Association (2017)[6] argues that social media data should be considered public or private based upon the specific online context, and the likely perception of the social media user. Therefore when deciding whether to consider data public or private researchers should consider whether the social media user would have a reasonable expectation of privacy [6]. If the social media user has reasonable expectation of privacy, the data should be considered private.

The following should be considered:

- Is the data on an open forum or platform (no registration required and anyone can view), for example Twitter, or are they located in a private group or closed forum?
- Is the group or forum password protected or a closed group where access must be requested?

- Would the platform user expect other visitors to have similar interests or issues as themselves (for example mental health support groups)
- Does the group have a gatekeeper that could offer advice?
- How have the users set up their security settings? [1]

Data collected from open and public online spaces present fewer ethical issues than private spaces, as the users of these spaces can be assumed to not have a reasonable expectation of privacy and are therefore are unlikely to consider their information as either private or confidential. However, researching in public spaces still requires caution because:

- As mentioned above, some users may not fully understand that the information they post is publically visible
- Users post information to that space expecting a certain audience will see it, so they may feel upset or violated if it's taken out of context or presented to a larger audience
- Inadvertent exposure of the real-world identity of an online persona can have real-world repercussions [7].

Even when in the public domain, information linked to an individual can still be sensitive. If in any doubt about whether to use a particular post, seek the user's consent.

If the data will be collected from closed of private online spaces there may be further ethical issues to consider and researchers should consider obtaining consent from the social media users to collect data from these spaces.

For the purposes of ethical review, fully in the public domain means that the data is freely available without having to register, request, or ask any permissions. For example, a lot of Facebook data would not be considered fully in the public domain as registration is required.

2. Risks and Harms

The ethical principle non-maleficence emphasises the need to avoid doing harm [2], and this is also true of research involving social media. Researchers need to thoroughly consider the potential harm that may result as a consequence of their research and should put protections in place to reduce any foreseeable risks.

Researchers using social media are at a disadvantage as by engaging with people online they have no direct contact with participants and therefore it can be difficult, if not impossible, to verify age or to assess the vulnerability of individuals. If the data is suspected to come from a young or vulnerable individual, informed consent cannot reliably be given and the data should not be used. In the case of children, researchers may decide to seek parental consent if this is possible [8, 9].

Research involving data extracted from private online spaces has a higher potential for harm. Intrusions from researchers into spaces considered private by their users may be invasive and unwelcome. Risks can include the disruption of social dynamics, users altering their use of the platform and in some cases can lead to individuals feeling like they can no longer participate on the site [10].

Data that is considered sensitive in nature has a higher risk of causing harm regardless of whether the data is public or private. The potential for research to draw attention to posts that would have otherwise been lost in a crowd or hidden in a private group has the potential to result in harms to the user who posted the information, particularly if it is possible to identify the individual. Harms can

include the disruption of personal relationships, social stigmatisation, feelings of embarrassment, feelings of being manipulated if the user discovers they have been disclosing information to a researcher and not a peer, and in extreme cases can result in discrimination in access to benefits, services, employment or insurance, or the discovery and prosecution of criminal activity [7,10]. If a research projects involves the collection of sensitive data special consideration needs to be given to the anonymisation of the data to ensure that risks to participants are minimised as much as possible.

Example
A researcher wants to study pro-legalisation narratives on Marijuana use by collecting posts with relevant hashtags from Twitter.

The data in this case can be considered fully in the public domain as anyone can view the information without the need to register an account or ask any permissions. It is also reasonable to assume that by posting on a public site with the use of hashtags that the users do not have an expectation of privacy.

However the data is sensitive as it refers to an activity that is still illegal in the UK and could therefore potentially be incriminating. It is also possible that users under the age of 18 may be commenting in the debate.

The data is public, therefore the data can be used without consent, however due to the sensitive nature of the data special consideration needs to be given to the anonymisation of the data. Results should be presented in aggregate form and no direct quotes should be published as this may lead back to the user's profile. The researcher may consider publishing paraphrased quotes with the ID handles removed, but should be aware of the relevant terms and conditions of the site. Direct quotes can only be used if informed consent from the user has been obtained, and steps have be taken to ensure that the user is over the age of 18 [8].

The timing of research is also something to be considered when determining the potential for harm. Inflammatory or offensive content is not uncommon on social media, and it could result in harm to the poster if comments made in the heat of the moment, or views users have expressed in the past but no longer hold are re-surfaced in a research project. Researching 'live' social media activity poses the risks of altering the behaviour of users if they discover that they are being observed, and potentially there is a greater chance of individuals being identifiable from live data collection [11].

Researchers need to thoroughly consider the potential harm that may result as a consequence of their research. The higher the potential for harm, the more consideration needs to be given to issues of appropriate consent and anonymisation.

3. Consent

Informed consent is key to the ethical principle of respect for persons [2] and should be upheld wherever possible. This is particularly important when the data being collected is sensitive or private. However, as stated by the British Psychological Association (2017):

"Where it is reasonable to argue that there is likely no perception and/or expectation of privacy (or where scientific/social value and/ or research validity considerations are deemed to justify undisclosed observation), use of research data without gaining valid consent may be justifiable" [6]

Therefore when deciding whether it is appropriate to obtain consent researchers should consider the nature of their research and the associated risks, as well as whether the data is considered public or private.

If consent is being used at the legal basis for the processing of personal data according to GDPR then GDPR-compliant consent MUST be obtained.

3.1 Data that is fully in the public domain

If the data is fully in the public domain then consent from users is not usually required, however unless consent is obtained appropriate anonymisation should be used to ensure that no individuals can be identified either explicitly or by implication in any reporting of the research (see section 5 for more detail). The only exception to this is when the data is from public figures acting in their public capacity. In this case it may be acceptable to attribute the data to the individual.

Example:
A researcher wishes to look at the discourse used by public health officials on Twitter in relation to a new public health campaign.

The data in this case can be considered fully in the public domain as anyone can view the information without the need to register an account or ask any permissions. It is also reasonable to assume that by posting on a public site with the aim of raising awareness the users have no reasonable expectation of privacy.

The researcher can use the data without obtaining consent and direct quotes from public health officials acting in their public capacity can be directly quoted. Any data collected from individuals other than public figures acting in their public capacity should be adequately anonymised in any research outputs.

3.2 Data that is not fully in the public domain

If it cannot be reasonably argued that the data is fully in the public domain then it may be necessary to obtain consent. What is appropriate will depend on how private the data is, the sensitivity of the data, and whether data could be linked to individuals. For example, if the participants are unlikely to consider the data private, the research is low risk, and the data will be aggregated so that no individual users will be identified, it may be sufficient to check that the terms of conditions of the site state that users have agreed for their data to be used for research purposes [(11)].

If the data could be reasonably considered private, if the data is sensitive, or if the data will be analysed in such a way that could potentially lead to the identification of individual users, informed consent from individual users should be obtained.

Researchers may want to consider the following when making a decision about how to gain informed consent:

- The terms of conditions of the platform. Does the platform allow the use of data for research purposes?
- The presence of gatekeepers such as site administrators or forum moderators. If a study involves collecting data from a closed group or site, contact should first be made with the site or group administrator as they will have a better understanding of the social dynamics

of the group and can advise on how best to proceed. When approaching a site or group admin it is vital that researchers are transparent about their own identity and that data will be collected for research purposes [8]. It should be noted that obtaining consent from the gatekeeper alone is considered insufficient [7].

- Whether the nature of the research means that it is appropriate to undertake covert observation (see section 4.3)
- The potential for harm if the community become aware that a researcher has been observing their interactions.
- The practicalities of obtaining consent in an online setting. (e.g. can consent be obtained directly within the platform e.g. by message, could individuals be directed to a web page with information about the research, could a separate forum discussion be created so that only those who want to participate in the research are observed etc?)

Example:
A researcher wishes to study support mechanisms between members of a closed and password protected discussion forum which deals with mental health issues. Registration on the forum must be approved by a site administrator before access is granted.

The data in this case is both private and sensitive. Users of the forum will have a high expectation of privacy and are likely to view the forum as a safe space to talk to people in a similar situation. Before entering the forum the researcher should identify themselves to the site administrator and request advice on how to approach the forum users. The site administrator may then contact the group before responding to the researcher and offering suggestions. The researcher should remember that consent from the site administrator does not negate the need to obtain consent from individual users.

Depending on the advice from the site administrator the researcher may approach the forum users in a number of ways, for example by setting up a new discussion thread so that only users who consent to participating are observed, or the site administrator may set up a safe space for users who do not want to participant allowing them to opt out of being observed.

If the researcher wishes to republish posts informed consent from the user whose post will be republished should be sought. All other data should be fully anonymised in any research outputs.

3.3 Covert observation in private spaces

The principles of respect for persons and non-maleficence require that researchers maintain respect for and take steps to avoid disrupting social structures, as well as carefully considering any potential consequences or outcomes of a research study. Intrusions from researchers into spaces considered private by their users may be invasive, unwelcome and socially irresponsible. Risks to researchers include being insulted or 'trolled', while risks to social media users can include the disruption of social dynamics, the loss of enjoyment when using the site, feelings that their freedom of expression have been curtailed and in some cases can lead to individuals feeling like they can no longer participate on the site, or whole discussions/groups being closed down [10]. Where the scientific value of such research is considered very high, the researcher will need to make a decision about whether joining the group without disclosure as a researcher (i.e. covert observation) might be more appropriate than obtaining consent, in order to avoid disruption and potential harm.

3.4 Deleted posts and withdrawing from research

If a user deletes a post this suggests that they no longer want others to see it, and this could be interpreted as the equivalent of withdrawing consent for use of the data. This poses a problem for researchers since they may be unaware that the post has been deleted following collection, or the deletion may occur after the data has been analysed. It is therefore important that this is considered when planning the research to ensure that this eventuality is covered.

It is also worth being aware that should a person find out that their data have been accessed, stored and used as research data, they are likely to have rights under the GDPR to stop these data from being processed if they could be linked to them personally [6]. While it is unlikely that a person will ever find out that their data has been used for research purposes, this is still something that researchers should be conscious of.

3.5 Further guidance on informed consent for research

Further information on consent can be found in SOP-005 Informed consent for research which is available here: https://lshtm.sharepoint.com/Research/Research-Governance/Pages/standard-operating-procedures-(sops).aspx. Appendix 8 of this SOP includes guidance on GDPR requirements for consent for collection of personal data.

4. Confidentiality and anonymity

Anonymity is a fundamental right of research participants and the violation of this right can result in a number of harms, for example, the linking of a particular post with an individual could compromise their job prospects, their educational prospects, or their relationships [12]. Unlike with traditional research methods, identifying the poster of a quote published in a paper, or the identity of the person behind a username can be relatively easy for the motivated individual. Quotes can usually be traced back with a simple google search, while IP addresses can link a username with an individual. Therefore it is particularly important for researchers to protect research participant's identities.

Unless a researcher has obtained explicit consent from a social media user to publish identifiable information about them, data should always be anonymised in publications and other outputs regardless of whether the data is considered public or private. The only exception to this is when the social media user is a public figure and are acting in a public capacity.

Suggested ways to do this include:

- Not collecting more information than is needed as storing extensive amounts of identifiable information increases risk [13]
- Replacing identifiable information as soon as possible, for example by replacing the username with a unique ID number. (However it should be noted that such datasets are often re-identifiable so should be treated as potentially identifiable data in line with GDPR)
- Aggregating the data, so no individual is identifiable [12]
- Not using direct quotes, but paraphrasing instead. It should be noted that this can compromise the integrity of the data and introduce bias, so it is important to note when this has happened [10, 14].
- Obtain consent if there are plans to publish any potentially identifiable information e.g. direct quote [15].

Please be aware that anonymisation practices can go against the Terms and Conditions of certain social media platforms. In these cases additional consideration should be given as to what is ethically appropriate

If the researcher wishes to use photographs of people which have been shared on social media the researcher should consider whether the individual(s) shown in the picture have consented to their photograph being taken and shared. For example, if the person in the photograph is not the person who posted the photograph on the social media site, it cannot be assumed that the individual consented to their photograph being shared. Researcher should also check whether any photographs or images they wish to reproduce are protected by copyright.

For more information please see SOP-036 on Confidentiality and Anonymisation of Research Data which can be found here: https://lshtm.sharepoint.com/Research/Research-Governance/Pages/standard-operating-procedures-(sops).aspx

5. Legalities

5.1 Terms and conditions

Before beginning a research study, researchers should read the relevant terms and conditions of the platform(s) that will be used to obtain data. Researchers should regularly re-check the terms and conditions as these change regularly in accordance with changes made to the platform. Being familiar with the most current versions of the terms and conditions can offer protection from potential legal action should they be violated [8].

It is also worth being aware of any terms and conditions that would impact the researcher's ability to uphold their responsibilities to research participants. For example current Twitter terms and conditions state that any quotes must include the full twitter handle [9]. In these cases the research may want to consider contacting the social media platform asking for an exception.

5.2 Third Party tools

If using a third party tool to access/collect social media data, the researcher should ensure that the tool is compliant with the terms and conditions of the social media platform.

5.3 Data protection

Identifiable and potentially identifiable social media data is subject to regulations set out in the GDPR, and an appropriate legal basis for the processing of personal data must be identified. In addition, certain types of data are classified as 'special categories' of personal data and specific requirements apply when processing these types of data.

Examples of special categories of personal data include, but is not limited to, information about an individual's:

- Race
- Politics
- Religion
- Health
- Sex life or sexual orientation

The Schools Data Protection Policy can be found here and any queries with regards to compliance with GDPR should be sent to the data protection officer at DPO@lshtm.ac.uk.

Please note that social media data is still considered potentially identifiable even if the user-names have been removed and should be treated accordingly.

5.4 Intellectual property

If the researchers wishes to reproduce posts, images or photographs the researcher should check to ensure that the thing they wish to reproduce isn't protected by copyright.

6. Additional points to consider

6.1 Blurring boundaries

The nature of social media means that there is a chance that researchers can become searchable by participants. Therefore researchers should pay attention to their online identity and privacy settings, and consider keeping their research persona and personal persona separate [8, 10].

The LSHTM communications team provide guidance on staff social media accounts, including setting up School accounts and support for experiences of harassment. This guidance can be found on the Communications and Engagement Intranet pages here:
https://lshtm.sharepoint.com/Services/comms-eng/Pages/communications.aspx

6.2 Aggregate data

Aggregated data is the consolidation of data relating to multiple individuals, and therefore cannot be traced back to any one individual.

If data collected from social media sites is aggregated before the researcher has access to the dataset (e.g. a secondary dataset provided by a third party), this does not require ethical approval as the data is no longer at an individual level.

If data is extracted from the social media site, and then aggregated by the researcher, this would still require an application to be made to the ethics committee, as the researcher will have access to the individual level data at at least one point during the data collection process.

Please note that while security concerns for aggregated data are not as significant as for individual level data, it can still be misused and misinterpreted by others, so care should be taken with its dissemination.

7. Useful documents/links

All SOPs can be found on the Research Governance and Integrity Intranet pages here:
https://lshtm.sharepoint.com/Research/Research-Governance/Pages/standard-operatingprocedures-(sops).aspx

The standard operating procedures that may be particularly helpful when planning research involving social media data include:

- SOP-003 Ethics approval
- SOP-005 Informed consent for research
- SOP-036 on Confidentiality and Anonymisation of Research Data

The Association of Internet Researchers (AoIR) Ethics Working Committee has produced two reports to assist researchers in making ethical decisions in their research. They have also developed a chart for internet researchers to use as a starting point when considering the ethics of their research.

The guidance documents can be accessed here: https://aoir.org/ethics/

8. References

1. Townsend, L, Wallace, C, (2018) "The Ethics of Using Social Media Data in Research: A New Framework" in The Ethics of Online Research edited by Kady Woodfield. Emerald Publishing limited.
2. Belmont report (Last accessed Sept 2018): https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html
3. Fiesler, C, Michaelanne, D, Feuston, J,L, Hiruncharoenvate, C, Hutto, C.J, Morrison, S, Roshan, P.K, Pavalanathan, U, Bruckman, A.S, Choudhury, M, Gilbert, E, (2017) "What (or Who) Is Public?: Privacy Settings and Social Media Content Sharing", Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, February 25-March 01, 2017, Portland, Oregon, USA.
4. Fiesler, C and Proferes, N (2018) "'Participant' Perceptions of Twitter Research Ethics." Social Media + Society 4(1).
5. Hudson J.M., Bruckman A, (2005) "Using Empirical Data to Reason about Internet Research Ethics". In: Gellersen H., Schmidt K., Beaudouin-Lafon M., Mackay W. (eds) ECSCW 2005. Springer, Dordrecht.
6. The British Psychological Society, (2017) Ethics guidelines for Internet-mediated Research (Last accessed Sept 2018): https://www.bps.org.uk/sites/bps.org.uk/files/Policy%20-%20Files/Ethics%20Guidelines%20for%20Internet-mediated%20Research%20%282017%29.pdf
7. Office of Research Ethics and Integrity (OREI), "Internet and social media" (Last accessed Sept 2018): http://www.orei.qut.edu.au/human/guidance/internet.jsp
8. Townsend, L, Wallace, C, (2016) "Social media research: A guide to ethics" (Last accessed Sept 2018): http://www.gla.ac.uk/media/media_487729_en.pdf. Google Scholar
9. UK Research Integrity Office (2016) "Good practice in research: Internet-mediated research" (last accessed Sept 2018): http://ukrio.org/wp-content/uploads/UKRIO-Guidance-Note-Internet-Mediated-Research-v1.0.pdf
10. Golder S, Ahmed S, Norman G, et al. (2017) "Attitudes toward the ethics of research using social media: A systematic review" J Med Internet Res; 19: e195. (Last accessed Sept 2018): http://eprints.whiterose.ac.uk/117721/7/fc_xsltGalley_7082_129674_66_PB.pdf
11. The University of Sheffield Research Ethics Policy note no.14 "Research Involving Social Media Data" (Last accessed November 2018): https://www.sheffield.ac.uk/polopoly_fs/1.670954!/file/Research-Ethics-Policy-Note-14.pdf
12. Ayers, J.W, Caputi, T.L, Nebeker, C and Dredze, M, (2018) "Don't quote me: reverse identification of research participants in social media studies" npj Digital Medicin 1(30).
13. Zimmer, M, (2010) ""But the data is already public": on the ethics of research in Facebook" Ethics and Information Technology, 12(4), p.313-325
14. McKee R. (2013) "Ethical issues in using social media for health and health care research" Health Policy, 110(2), p298–301

15. Government Social Research (GSR). 2016. “Using social media for social research: An introduction” (Last accessed Sept 2018): https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/524750/GSR_Social_Media_Research_Guidance_-_Using_social_media_for_social_research.pdf