

RESEARCH USING SOCIAL MEDIA DATA - GUIDANCE FOR RESEARCHERS

Why Social Media Research Raises Ethical Concerns

Content posted to social media can be permanent, searchable, and visible to audiences far beyond those the content creator had in mind. Several well-documented controversies show the dangers of treating accessible online content as freely available for research: the repurposing of Facebook personality-test data by a political consulting firm (despite ethics approval being refused); a large-scale Facebook study of emotional contagion run without users' awareness; an attempt to predict sexual orientation from dating-profile photos without consent; and research that risked exposing the identities of political protesters. The lesson is that the ethical implications are real, and that accessibility alone is never a justification for proceeding without careful ethical consideration.

Methods of Collecting Social Media Data

Data can be collected from social media platforms in three main ways, each of which has different implications for platform compliance, data format, and risk.

Application Programming Interfaces (APIs) enable the structured transfer of data from a platform to a researcher's tools in a standardised format. This is generally the most reliable and complete method of data collection and is typically preferred where API access is available. That said, many platforms place restrictions on their APIs by limiting the volume of data that can be retrieved, the timeframe that can be queried, or the types of content accessible. Some platforms provide an elevated level of access specifically for academic researchers, which may remove or relax standard restrictions. Researchers should read and comply with the terms of service governing API use on any platform they intend to draw from.

Web scraping refers to the automated extraction of content from the pages a platform displays to its users, usually through scripts or crawling tools. It is commonly used where API access is unavailable or insufficient for the research purpose. However, many platforms explicitly prohibit scraping in their terms of service, and researchers who engage in it may face account suspension or other consequences. Some platforms offer their own data scraping tools, some of which must be used in place of independent scraping and often attract a fee. A thorough review of the relevant terms of service is essential before this method is used.

Manual collection involves gathering content by hand through screenshots, downloads, written notes, or other direct means. This approach is most suitable for smaller datasets or in situations where automated collection is restricted by the platform. The same ethical considerations apply as with other methods: collecting content from private spaces, closed groups, or personal accounts will generally require the consent of those involved and ethics approval before the research commences.

Ethics Considerations to Address in Your Application

The considerations below are best understood as a guide rather than fixed rules. Social media research varies enormously by platform, population, topic, methodology, and data type, and

there is no single approach that fits every project. Please address each of the following in your application, describing the steps you intend to take in response.

1. Have you reviewed the platform's terms and conditions and privacy policy?

- Which platform(s) will you collect data from and what do their current terms and conditions permit or prohibit?
- Have you checked both the terms directed at ordinary users and those governing third-party, developer or research access?
- Have you reviewed the platform's privacy policy to understand what users have been told about how their data may be collected, used and shared?

Why this matters

Platform terms and conditions change often so it is worth confirming you are working from the current version before you begin. The terms and conditions set out what you are permitted to do on the platform, while the privacy policy reflects what users have been told to expect about their data, which is relevant to their reasonable expectations of privacy. Be aware that a user accepting a platform's terms is not the same as that user consenting to take part in research.

2. What information will be collected (e.g. names, quotes, keywords)?

Please be as specific as possible here as this helps the Committee to understand how identifiable the data is.

Why this matters

Different types of data raise different issues. A text post, an image, and an audio or video clip each reveal different things about a person and carry different risks to their privacy and identifiability.

3. Will consent be obtained from users? If not, why?

Why this matters

Consent underpins the ethics of all research. Compared with more traditional research, social media raises particular difficulties as users are often unaware their data is being accessed and analysed. Accepting a platform's terms and conditions is not equivalent to research consent (as tempting as this is) especially since there is no guarantee participants have read them. Core aspects of consent, such as the right to withdraw, are also complicated: it is unclear whether deleting a post or account counts as withdrawal, or whether the researcher would even know it had occurred.

4. What are the reasonable expectations of social media users in this context?

And is there a reasonable expectation of privacy?

Why this matters

Privacy concerns arise when the proposed access to or use of data or information does not match the expectations of the individuals from whom this data was obtained or to whom it relates. Some forums or group may technically be open and accessible, yet members could feel violated if their contributions were used without consent. Therefore, such information

must be considered in the context of the need for consent or the waiver of the requirements for consent, and the risks associated with the use of this information (National Statement Section 3).

a. How public is the information?

Please explain,

- Is the data on an open forum or platform (e.g. X), or is it located within a closed or private group (e.g. within Facebook) or a closed discussion forum?
- Is the group or forum password protected?
- Would platform users expect other visitors to have similar interests or issues to themselves?

Why this matters

An open discussion on X, shared under a hashtag to reach the widest possible audience, sits at the public end of the spectrum; a password-protected Facebook group is closer to private. Where access depends on membership approval or a password, your first step should usually be to contact the group or site administrator. If you intend to ask users questions to generate new data, you must be transparent about who you are (i.e. university researcher) and that responses will be used as research data.

b. What is the nature of the space from which data will be collected?

Many online spaces are private or sensitive in nature and exist as a form of social support for a particular group of people. It is important to consider the invasiveness of the researcher's presence, even when no identifiable information is to be collected. Bear in mind, too, that a single post may capture data about people beyond the original author, such as those who comment on it.

Why this matters

The more private or sensitive a space is, the greater the expectation of privacy among its members. These spaces may also draw in young or vulnerable people who are difficult to identify because users often obscure their real identities online. Where consent cannot be reliably obtained and data is suspected to come from young or vulnerable individuals, consideration should be given to excluding it from the research. Where a real risk of harm exists, you should seek consent or adopt a more traditional method in which consent and confidentiality can be more reliably assured.

5. In what formats do you intend to publish or present your research findings?

- Do you intend to reproduce individual units of data (e.g. specific tweets, posts, images)?
- How will data be de-identified, or will consent be obtained from the individuals concerned?
- Could re-publishing the data compromise the anonymity of any individuals or groups?

Why this matters

A verbatim quotation can often be traced back to its author through a simple search, so pseudonymisation alone does not make it anonymous. Some platforms compound this by requiring that units of data be reproduced only in their original form and attributed to the original poster. A researcher's responsibility increases as the risk of harm to participants

increases: the stakes rise sharply where the subject matter is sensitive and exposure to new audiences could place users at risk of embarrassment, reputational damage or legal action.

6. Do you intend to share your data set?

- Confirm that the platform's terms and conditions permit data sharing.
- Does the dataset contain material that may cause harm if re-published? If so, how will you manage this (e.g. removing or paraphrasing sensitive information, or not sharing the dataset)?
- If you do not intend to share your dataset, can you explain your reasons (e.g. to a funding body)?

Why this matters

Because platforms tend to retain searchable data and metadata for long periods, de-identification in any secondary use warrants particular care.

Note: Sharing is generally considered safe where data is aggregated so individual postings are no longer discernible, where no sensitive data is present, or where individuals cannot be identified.