LITTLE QUICK FIX:

GET YOUR DATA FROM SOCIAL MEDIA

#LittleQuickFix

LITTLE QUICK FIX:

GET YOUR DATA FROM SOCIAL MEDIA

Nicola Thomas

SSAGE

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

Contents

Everything in this book!

Section 1 Social media data can serve many research needs. It is important to know the potential, and pitfalls, of social media data. Decide how social media data can provide a good fit for your research.

Section 2 Conduct your social media research to high ethical standards. Understand and address ethical dilemmas when obtaining and using social media data.

Section 3 Choose the type and scope of social media data. Decide what type of data you need, from where, when and from whom or what, to decide the best approach.

Section 4 Develop a robust sampling strategy to obtain social media data. Identify your population, sampling frame, sample and sampling technique(s) to best meet your research needs.

Section 5 There are three ways to extract social media data. Know how to extract social media data, either manually, computationally, or a hybrid approach, to best suit your research needs, capabilities and resources.

Section 6 Take strategic steps to plan your data collection. Thinking through, and planning out, your data collection will increase your chances of generating valuable insights from your social media data.

Social media data
can serve many
research needs

Section 1

Q

Can social media data meet my research needs?

A

Social media data can shed
light on a huge variety of social
phenomena. It can serve many
research needs, but it's important
to understand its limitations.

60 SEC summary

Social media research is diverse, dynamic and challenging

The enormous popularity of social media platforms and the digital means to generate, share and consume content, and socially interact, leave rich digital traces that can be used for research. Data enable us to describe, explain, predict and evaluate a huge range of online and offline social phenomena. But **as with all data types, social media data does have limitations.**

DEFINING SOCIAL MEDIA DATA

Social media data is an umbrella term for the different kinds of digital traces produced by or about social media users. A huge volume of data emerges from real-time social interactions on a variety of online and mobile social media platforms. For instance, some platforms offer opportunities for social networking (e.g. Facebook), questions and answers (e.g. Quora), reviewing (e.g. TripAdvisor), image and video sharing (e.g. Instagram), instant messaging (e.g. WhatsApp), collaboration (e.g. Wikipedia), gamifying mobile check-ins (e.g. FourSquare), expressing deeper thoughts (e.g. Medium) and sharing economies (e.g. AirBnb).

THE RESEARCH VALUE OF SOCIAL MEDIA DATA

Collectively, social media platforms create an interactive social web where billions of people come together to generate, share and consume content. The sheer volume, velocity and variety of social media data therefore presents a valuable opportunity to conduct research.

13
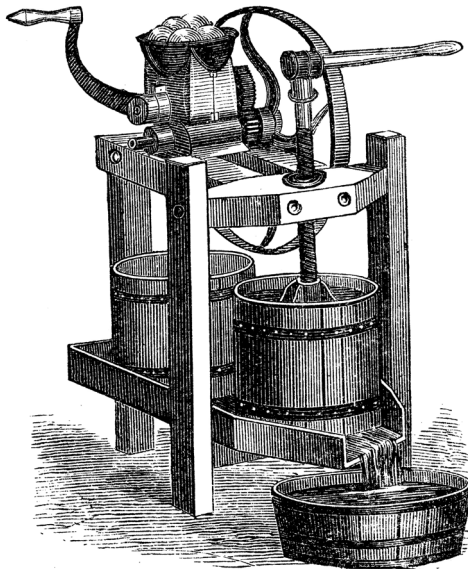
HUGE NUMBER OF RESEARCH TOPICS

There is no shortage of research topics. Given the massive amount of social media data generated by individuals, communities and organizations, we could choose from any number of topics – from consumer choices, disease outbreaks – to policy interventions. Social media data are now widely used in the social sciences. It doesn't matter whether you are studying psychology, anthropology or sociology, medical or political science, human geography or business, it's very likely your discipline has already found genuine research value in using social media data.

YOUR
RESEARCH NEEDS
WILL GUIDE YOU

Given the daunting range of social phenomena potentially available to research, it's best to start with your own research needs. Which topic ignites a passion in you? What problem do you want to solve? What will your theoretical contribution be? What is your research question? If you're currently unsure, the following sections will help you decide.

DATA FOR
DIFFERENT AIMS

We can use social media data for different research purposes. Do you want to describe, explain, predict or evaluate something? Below are some real-world examples of how social media data are being put to different uses:

Description: What is happening?

Understanding negative word-of-mouth dynamics in social media networks.

Explanation: Why is this happening?

The role of psychological traits in explaining Facebook use.

Prediction: What will happen in the future?

Forecasting stock market fluctuations.

Evaluation: Did it work?

Assessing social networking interventions to treat depression in young people.

THREE CROSS-CUTTING TYPES

Social media research tends to fall into three broad types.

1   **Online phenomena**   Researching social phenomena manifested in social media environments. For instance, how people construct online identities, or how they think, feel and behave.

2   **Offline phenomena**   To understand real-world events. For instance, by investigating the spread of diseases by analysing online reported symptoms.

3   **Platform phenomena**   Looking at how social media platform(s) shape and influence people, for instance, due to their design functionality.

A RICH VARIETY OF METHODS

Social media data are amenable to an increasing number of data collection and analytical methods. Below are some popular choices.

Quantitative methods

- Quantitative content analysis

- Data mining

- Social network analysis

- Natural Language Processing

- Visual analytics
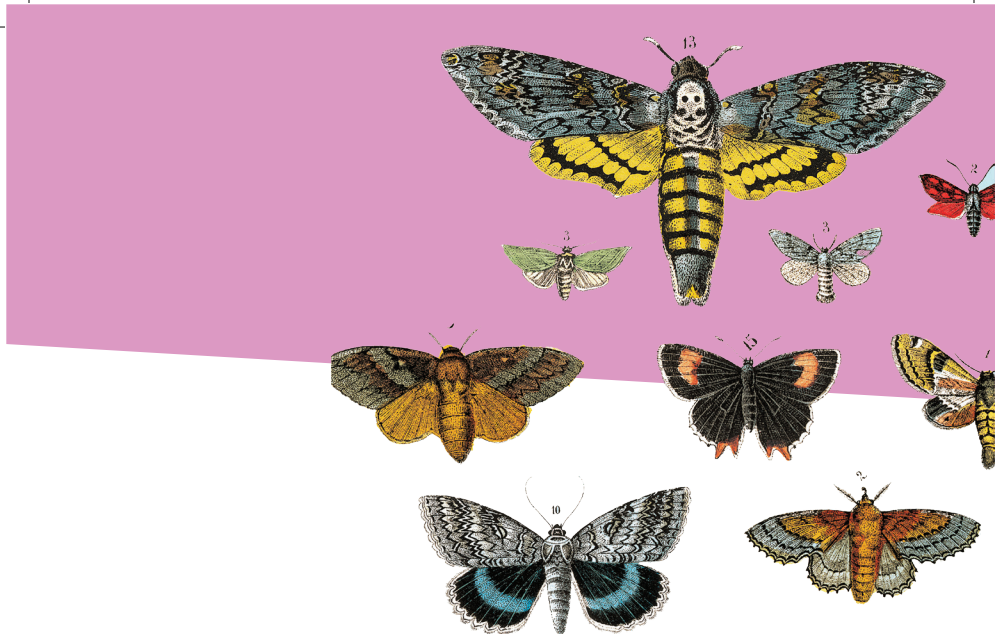
Qualitative

- Qualitative content analysis

- Discourse analysis

- Netnography

- Online ethnography

- Online focus groups

Tip: Consult contemporary sources as methods are increasingly being developed, advanced and adapted for social media contexts.

KNOW YOUR DATA
LIMITATIONS

Social media presents a significant research opportunity. However, we must be mindful of its pitfalls.

Shrinking data access. Many social media platforms allow data access. However, data accessed via application programming interfaces (APIs) are curated and data availability is shrinking. People often refer to the massive stream of social media data as the firehose. For most of us, the firehose is out of reach. The accessible portion resembles a garden hose (10%) or a sprinkler (1%).

Missing metadata. While profiles can be data rich, users are not required to publish demographic data. Without knowing who users are, attempts to understand how a phenomenon relates to age, gender, class, and so on, can be undermined.

Disproportionate representation. Social media populations do not represent human populations. For instance, the Twitter population overrepresents males, and underrepresents older cohorts. Neither should we assume the types of users on one platform are the same as another.

Identity play. People can present and express themselves differently in online and offline contexts. This can call the authenticity of our data into question.

Beware social bots. Non-human agents are prevalent on social media. Spot and remove them from your dataset … unless they are the topic of your research!

EVERY CLOUD HAS A SILVER LINING

Most researchers are in the same boat. Even fully established and well-funded researchers face the same pitfalls. You are not alone.

Honesty is a virtue. Demonstrating that you recognize and understand these types of research limitations and bias will be looked upon favourably.

Every problem is an opportunity. Research is increasingly attempting to understand and overcome these limitations. Perhaps you could too.

Answer the questions below

1   What is social media data?

....

....

....

2   What four research aims can be met using social media data?

1..

2..

3..

4..

3   It is difficult to access large representative datasets.   **True   /   False**

2

Section

Conduct your social media research to high ethical standards

Q

How do I gather and use social media data ethically?

A

10 SEC

summary

You need to address ethical guidelines and legalities, privacy and consent, as well as minimize potential risk of harm to participants.

Be ethical!

Before getting your social media data, it is essential to **think through, and address, a range of ethical issues**. Each stage of your social media research plan has its own distinct challenges. You need to consider your data source(s), target population, sampling strategy, and how you extract, store and use your data.

THE FOUR GOLDEN RULES OF ETHICAL RESEARCH

Ethical research adheres to four golden rules:

1   do good

2   do no harm

3   be respectful

4   treat people fairly

There are currently no 'hard and fast' rules on how to best apply these golden rules in social media research. The best advice is to err on the side of caution. To help you do this, this section provides guidance on how to address the ethical dilemmas raised by working with social media data. At the very least, and if in any doubt, always consult your institution's ethics guidelines and those of your discipline.

START WITH YOUR
DATA SOURCE

Check legal terms. Social media platforms provide terms for use and privacy. It is important you consult these terms to make sure the access and data you want is permitted. As terms change, make yourself familiar with the most current version to prevent you from violating any conditions.

BE AWARE OF SOCIAL MEDIA AGREEMENTS

Let's assume you are a regular social media user. How often have you read these terms? Like others, you may have skipped the reading bit and hastily clicked your agreement. In fact, you may have agreed to things that you didn't even realize! Due to this thorny issue, some people think that informed consent is a must, even if the data are publicly available. To help navigate this issue within ethical bounds, most researchers first consider whether the data they wish to collect is really public.
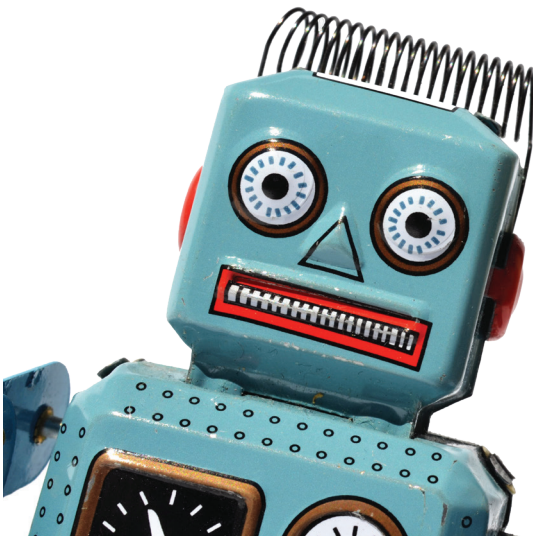
KNOW IF YOUR DATA IS REALLY PUBLIC

Is the data you want really public? The big question you need to answer is: Would the social media user(s) you wish to collect data from reasonably expect to be observed by you? To help answer this big question, let's break it down into three smaller questions.

1   Is the data you wish to gather from an open platform, such as Twitter?

2   Does the platform allow the user to control the amount of data that is publicly available, such as protected Tweets and direct messages?

3   Does the data you need already exist? In other words, has the data already been user generated, rather than elicited directly by you, such as by asking interview or survey questions?

If you answered yes to these three questions, then you could consider the data as public.

THE NEED FOR
INFORMED CONSENT

Alternatively, you may want to get data from private sources that require permission.

1   Is your data located within a private forum, for example a closed Facebook forum?

2   Is the data held within a password-protected space?

3   Is there a gatekeeper or administrator involved?

4   Will you be interacting with participants, for instance, by asking questions, to elicit new data?

If you answered yes to any of these questions, you must gain informed consent.

CONSIDER RISK OF VULNERABILITY

Regardless of whether your data are sourced from public or private spaces, user generated or elicited, it is essential you identify and respond to the level of risk your research could pose.

Are your research participants vulnerable? You have a responsibility to your participants. Potentially vulnerable people, for example children, older persons or adults with learning disabilities, need careful consideration. For instance, a vulnerable person may not have the ability to withdraw from an online interview, even if they feel uncomfortable. Nor can you necessarily rely on a participant's ability to provide informed consent. In the case of minors, you must always seek parental consent.

35
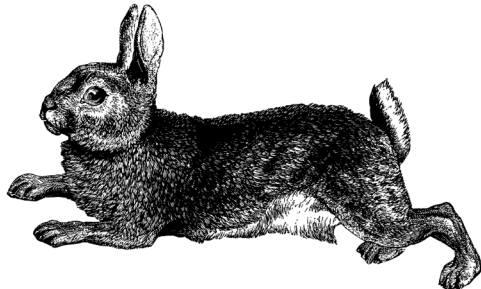
BE CAREFUL OF
THE UNKNOWN

Identity, age and capabilities can be unknown or hidden. For example, a young child may pretend to be older than they are. This means it is often difficult to infer vulnerability. Given this challenge, it's important to think and plan ahead. What could you do to minimize any potential harm?

IS YOUR DATA SENSITIVE?

It's ultimately up to you, and your institution, to decide whether your data could be sensitive. Here are some pointers to help you decide.

According to the European Union, 'sensitive' personal data covers:

- personal data revealing racial or ethnic origin

- political opinions, religious or philosophical beliefs

- trade union membership

- genetic data, biometric data processed solely to identify a human being

- health-related data; and data concerning a person's sex life or sexual orientation.

If your data falls into any of these categories, you can safely assume it's sensitive.

37

EVEN MUNDANE ACTIVITIES CAN CAUSE HARM

If the data you want concerns everyday stuff, like recipes, the weather or fashion, it's unlikely to contain sensitive information. Yet even mundane information can cause harm. For instance, if data are taken out of context and/or exposed to new audiences, it could lead to emotional distress and reputational damage. As such, it's important to consider how your data will be used, now and in the future.
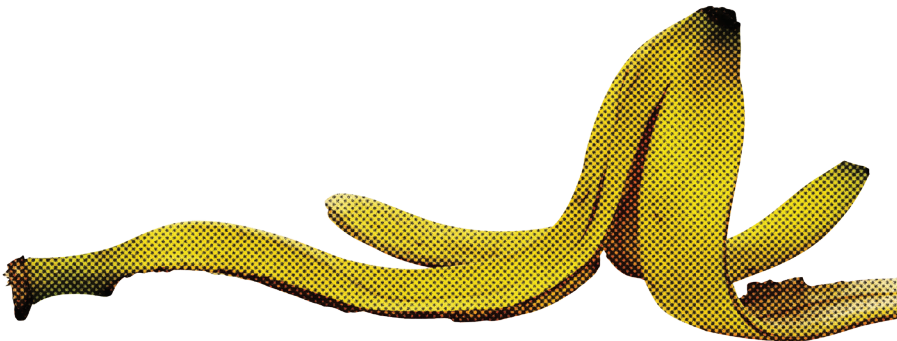
CONSENT
MATTERS

Informed consent is required if you want to directly elicit data, access data from private online spaces, if your topic is sensitive and/or participants are vulnerable. You will also need permission to reuse copyrighted materials, such as images, audio and video data. However, gaining informed consent when the desired dataset is large is problematic.

Realistically, gaining informed consent from hundreds or thousands of users may be too difficult, even impossible. Many researchers bypass the informed consent step, but only if their data are considered public. Regardless, there are several more steps that researchers should take to avoid harm to participants.

RESPECT
ANONYMITY

Remove all identifiable features, including user handles and profile pictures, from your dataset. Exceptions are made however for data generated by public figures or organizations, such as celebrities, campaigners or brands, as posts are intended to be shared widely.

In the case where you are mapping geotagged social media data, and individuals could be potentially identified, you should also apply a 'geomask' to protect a person's privacy.

Be aware of a caveat. Some social media users may prefer to waive anonymity, preferring to take credit for their content.

STORE DATA
SECURELY

All your data – raw and coded – must be kept safe, secure and protected. Password-protected devices, files and encryption are good options. Always consult your institution's guidelines as they may contain specific recommendations, including how long your data can be kept.

SHARE DATA WISELY

You may want to publish your data, for instance, at a conference or in a blogpost or article. Yet social media data extracts, such as Tweets, published verbatim, can be retraced back to their source. To prevent harm, consider three options:

1 aggregate data

2 gain consent from those users whose extracts you wish to publish, or

3 without losing context, paraphrase extracts to prevent identification.

Follow this checklist to make sure your social media research is ethical.

Legalities:

❑ Consult research ethics guidelines (e.g. institutional, disciplinary).

❑ Read platform terms for public data access and/or user contact.

❑ Seek permission to reuse copyrightable materials.

Privacy, risk and consent:

❑ Protect privacy and bring no harm to your research participants.

❑ Identify and address any risks, e.g. working with vulnerable people.

❑ Gain informed consent. Alternatively, explicitly state why informed consent is not possible and gain institutional authorization before data gathering.

Data storage, use and reuse:

❑ Secure all your data safely during and after your study.

❑ Anonymize your dataset to protect privacy and enable confidentiality.

❑ State how you will protect participants from harm if you publish data extracts.

43