3 Social Media Research

Rosanna E. Guadagno and Alberto F. Olivieri

Abstract

The purpose of this chapter is to review the contemporary methods used to collect and examine data on social media and to explore the common pitfalls of internet research. The discussion focuses on the importance of internet research while also reviewing common practices of data retrieval (e.g., crowdsourcing and snowball sampling). We will also explain a commonly used tool to analyze data collected using social media. Specifically, one section is dedicated to the Linguistic Inquiry and Word Count software (LIWC); another section focuses on a brief overview of machine learning (ML) techniques and data visualization. At the end of the chapter, we will also examine some common ethical concerns, focusing mainly on anonymity and privacy, while also giving a general overview on the European General Data Protection Regulation (GDPR). Future directions for social media will then be addressed.

Keywords: Social Media, Social Networking, Research Methods, Internet, Ethics in Research

The amount of data available on social media is vast and potentially overwhelming, especially for newer researchers. Nonetheless, the analysis of this type of data can provide real insights into human behavior. For instance, analyses of digital footprints have shown that gendered behavior persists on social media such that women's posts generally show more emotions associated with feminine gender stereotypes relative to men (Park et al., 2015). For men, their social media posts are also consistently masculine activities and match expectations for independence and agency. Other research has found that people's personality characteristics, sexual orientation, political leaning, and other individual differences can also be predicted by their digital footprints (e.g., Kosinski et al., 2013; Kern et al., 2019). The analysis of the text in people's social media posts can also predict their psychological states (Boyd & Pennebaker, 2017). Data collected from social media can also be used to predict and create models of the viral spread of information. However, social media are just a tool, and as such, can be improperly applied with unethical or malicious intent.

Love it or hate it, social media have become a staple of modern life. While this relatively new type of digital communication has existed in some form since the mid-1990s (Guadagno, in press), empirical work only began to accumulate later. Furthermore, there are still substantial gaps in the literature that need to be filled before scholars can fully understand the effect social media have on people's

thoughts, beliefs, behavior, and interpersonal relationships. For researchers, the stories recounted in this chapter serve as a cautionary tale of what not to do for scientists and scholars interested in using social media to conduct research. The present chapter provides a roadmap for ethical research on social media. We examine various methodologies for recruiting participants and collecting data and conclude with a detailed discussion of the ethical considerations for research on social media.

What Are Social Media?

"Social media" refers to a subset of Web 2.0 technologies that shifted people's internet use toward an emphasis on self-generated content (Oinas-Kukkonen & Oinas-Kukkonen, 2013). In the early days, social media largely consisted of *social networking sites* (SNS; e.g., MySpace). These days, "social media" is the more common phrase used to describe all related technology. Renowned internet researchers danah boyd and Nicole Ellison were the first to provide a concise operational definition of social media and their operationalization is still in use today. Specially, boyd and Ellison define social media as follows:

> Web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. (boyd & Ellison, 2007, p. 211)

While specific social media platforms have unique characteristics and purposes (e.g., LinkedIn is primarily used for career networking while Facebook is used primarily to keep in touch with friends and family), there are several features that are shared by social media applications. These are (1) each user has their own profile that they can then personalize and is visible to other users; (2) people connect with others – friends, family, strangers – and communicate with them through the application; (3) users read and interact with content (text, videos, photographs) created by other users they are typically (but not always) connected with. Over the years that social media have been widely used, this new technology has facilitated new friendships and romantic relationships, job opportunities, helped people find old friends, discover new family, and served as a source of news and information.

One other notable aspect of social media pertains to its global research. People all over the world use social media, although the different applications may vary by country. For instance, social media applications created by countries in the Western world are banned in China (Guadagno, in press). As a result, although people in China may use social media similarly to people elsewhere, the specific applications they use are different. Everything we do while using social media, whether it be simply browsing or actively posting photos or messages in public or private, is logged by the social media application. This information is used by social media companies for many purposes – some benign (e.g., improving the user experience and suggesting new connections and groups) and others potentially unethical (microtargeting ads, performing experiments on users without their explicit consent). This information is often used to create and refine algorithms – computer

programs that perform a specific pre-set task – that predict the type of social media content people are likely to engage with and subsequently present users with that content.

Just as social media applications record people's actions, so can researchers – either directly via webscraping digital footprints or indirectly by assessing people's perceptions of their social media use through experiments and surveys. In the sections below, we review the different options for finding and recruiting participants and conducting research on their social media use. The literature on social media is extremely broad both in terms of topic studied and methodology used. In light of this, the present chapter reviews the different options and considerations and also provides resources for further information.

Collecting Social Media Data

Social media have some unique affordances that make them a rich and almost endless resource for researchers (Bayer et al., 2020). One of the most prominent features of the "social" part in social media is the large amount of data on human behaviors and interactions now easily accessible. Researchers who want to focus on understanding social behavior and trends can access vast amounts of information regarding human communications, trending topics, and responses to events or stimuli. Other useful research avenues that could stem from diving deep into social data are sentiment analysis for specific topics or brands – an invaluable approach for fields like marketing, public relations, or political science.

Another crucial aspect to consider is the immediacy or real-time nature of the information gathered from social media. This enhances research on crisis and disaster management, and it holds potential for educational research. Moreover, controversial topics, events, and virality, in a continuously shifting landscape, must be constantly kept under scrutiny to properly track and study dis- and misinformation. Finally, a fundamental dimension of social media is the network structure itself. Techniques such as community detection can lead to the discovery of online groups and their interactions, while digital ethnography can offer novel views on the life cycle of these communities.

There are a variety of ways to recruit participants and collect data, and these are somewhat dependent on the population of interest, the specific research question(s) a scholar wishes to answer, and the type of data best suited to answer the question. In the sections below, we first explore the different ways in which researchers select samples and recruit participants, then transition to a discussion of the ways in which data on social media can be collected, and finish with an overview of data analyses that are well-suited for this type of data.

Finding Participants

In this section, we review options for recruiting participants, including recruiting participants straight from social media and recruiting crowdsourced

samples. We will explore the specific characteristics of each approach in subsequent sections, briefly explaining only the more superficial differences. Social media platform selection must also keep in mind the structure and focus of the platform and how those characteristics can influence the demographics and the types of interaction of its userbase.

Snowball Sampling

While this form of sampling is widely used for a variety of research projects (see Chandler, 2023 for an overview of the technique), the technique works especially well on social media. As the phrase suggests, snowball sampling refers to the selection of a small number of participants, then asking for their assistance in identifying and recruiting other potential participants. Researcher(s) can start the sampling by sharing a link to their study on their social media pages and ask their connections to share the link as well. If the research calls for a specialized sample, there are myriad specialty groups in many forms of social media, and the researcher(s) can post a request for participants in these groups. Once a participant has completed the study, researcher(s) can also ask participants themselves to share the link to the study with any additional contacts that may fit the recruitment criteria.

This technique is well-suited to sampling from atypical (e.g., people who enjoy ice swimming) or difficult to access samples (e.g., mothers of children with rare diseases). It is also very easy to implement. However, this technique does not provide a randomly sampled group of participants, so the results of any studies using snowball sampling have limited generalizability (Parker et al., 2019). Snowball sampling on social media has been shown to recruit such difficult to reach samples as mothers of children with developmental disabilities (Lee & Spratling, 2019) and nurses (Chambers et al., 2020). Researchers have also developed snowball sampling methods to obtain digital footprints such as tweets from difficult to identify online communities (Wang et al., 2017).

Crowdsourced Data Collection

The term "crowdsourcing" was coined in 2006 and describes "the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers" (Merriam-Webster, n.d.). While non-scholarly crowdsourced data appear all over social media, especially as reviews of films, television, video games, businesses, consumer goods, and software applications, crowdsourced data collection refers to the use of specific software services that provide a researcher with their sample of participants for a nominal fee (e.g., Amazon's Mechanical Turk, Prolific, Survey Monkey, YouGov). While some services offer the option to recruit directly from people who meet the recruitment criteria (e.g., Mechanical Turk), others offer access to pre-selected samples (e.g., Prolific Academic). Furthermore, some services are coupled with data collection

software (e.g., Survey Monkey). See Guadagno (2019) or Chapter 2 (this volume) for a detailed description of how Mechanical Turk works).

Recent empirical work has established the rising dominance of crowdsourced samples for social and behavioral research. For instance, Anderson et al. (2019) sampled studies published in the top-tier social psychology journals and found that for the year 2015, approximately half the articles published in these journals used such samples. Other research has demonstrated similar increases in the use of crowdsourced samples more broadly (Chandler & Shapiro, 2016; Zhou & Fishbach, 2016). Numerous disciplines in the social and behavioral sciences have largely benefited specifically from the use of MTurk. For instance, some disciplines (e.g., public administration and management research) found that using MTurk has proven useful to their fields (Wright & Goodman, 2019). Other disciplines such as market research provide another prime example where MTurk has successfully replaced traditional surveys.MTurk is often able to give results within hours and within an acceptable error rate compared to traditional approaches, making scientific investigations with it extremely cost-effective (Bentley et al., 2017). Thus, while this method is used more broadly than just research on social media, it represents a means of recruiting participants that is convenient, relatively inexpensive, and broadens the diversity of samples relative to undergraduate psychology samples (Casler et al., 2013; Berinsky et al., 2012). Furthermore, research suggests that data collected with a crowdsourced sample yield results that replicate studies conducted in a traditional laboratory setting. This has been demonstrated across a variety of disciplines, including psychology (Buhrmester et al., 2011) political science (Clifford et al., 2015), and market research (Bentley et al., 2017).

While this method of participant recruitment may seem like a panacea, crowdsourcing research in this manner is not without its drawbacks. First, the vast majority of potential participants are often from a limited number of geographic locations (e.g., MTurk research participants primarily hail from the United States and India), making it difficult to recruit representative samples outside of those nations (Antoun et al. 2016; Paolacci & Chandler, 2014; Ross et al., 2009). Furthermore, it is important for researchers to always keep in mind that they have limited control over the extent to which participants are completing anonymous online experiments. This characteristic of crowdsourcing platforms could potentially skew the results. Therefore, in the analysis phase, researchers should proceed carefully. It is fundamental to add quality controls to crowdsourced studies to ensure reliable data.

Attention checks that ask participants to select a certain response option have been shown to resolve this issue (Nichols & Edlund, 2020; Rouse, 2015). There also seems to be a relationship between rate of pay, length of study, and quality of data collected such that shorter, higher-paying studies tend to attract more attentive participants (Buhrmester et al., 2011). Other issues include differential attrition, with participants often dropping out of some experimental conditions at much higher rates than participants randomly assigned to other conditions (Zhou & Fishbach, 2016), and participant crosstalk, in which participants share the details of an experiment with other potential participants (Edlund et al., 2017). With respect to this latter issue, there are websites (e.g., https://turkopticon.net) created to allow crowdsourced

subjects to compare their experiences across the studies they participate in. On these websites, subjects compare details of the study, the length of time to complete the study, and assessments of the quality of pay. Fortunately, the problem of participant crosstalk on these third-party websites can be attenuated by asking participants in the study to refrain from sharing key details at the conclusion of debriefing (Edlund et al., 2017; see also Clark & Blackhart, 2023).

Obtaining Data

While the prior section addressed different ways to recruit participants into an experiment or survey on social media, this section reviews the different types of data that can be collected on people and the psychological properties surrounding their social media use. These include surveys, social media simulators, and web-scraping of digital footprints.

Surveys

Surveys are a series of questions in which researchers ask people about their thoughts, beliefs, feelings, and behaviors. This is also the method used to assess personality inventories (Burger, 2014). Some survey methods are intended to gather information about a target population; other times, surveys are used to assess the dependent variable(s) in an experiment. Surveys are inexpensive to administer and are often used to study psychological processes that are not easily observable because they are rare or occur as part of internal, cognitive processes. The generalizability of surveys is often limited, as it is difficult to recruit truly representative samples of a population. Furthermore, survey data is often prone to social desirability bias (Grimm, 2010) – the phenomenon in which people report thoughts, beliefs, feelings, or behaviors that are considered normative, appropriate, or socially acceptable. There are a wide variety of digital services that provide the tools to create and post a survey designed by members of the research team (e.g., Qualtrics, Google Forms, Survey Monkey), and most universities typically have a license for such software.

The utility of surveys goes beyond experimental research, as they are frequently used also for correlational and quasi-experimental research. For instance, a study where the authors evaluated how gender and personality traits influence the usage among undergraduate students of social networking sites, such as Facebook and MySpace, used surveys to assess this research question. Their findings revealed that men typically use these platforms with the goal of forming new relationships while women use them more for maintaining existing relationships. Additionally, it was observed that women low in agreeableness used instant messaging more often, and men low in openness played more games on these sites. These findings highlight the significance of individual differences in online behavior (Muscanell & Guadagno, 2012).

Another survey study compared the use of social network sites (SNS) between two age-matched groups of young adults. The first group was comprised of college students from a large introductory psychology subject pool; the second group was homeless young adults from metropolitan shelters. The findings indicated a similar use of technology across the two groups, regardless of socio-economic status and ethnicity. The results cast doubts on the concept of "digital divide" as still relevant, showing minimal differences between the subsamples and suggesting the need for a paradigm shift (Guadagno, Muscanell, & Pollio, 2013). See Chapters 16 and 17 in this volume for more details on these issues.

Social Media Simulations

While the researchers working for social media companies perform experiments on people regularly, scholars outside of these companies have more limited options. One option is to create a simulated version of social media for participants to use during an experiment. This is typically done by either running a simulation designed by other researchers (e.g., the many dis/mis-information games; Roozenbeek & Van der Linden, 2019) or programming your own content and pre-selecting responses to participants (Prevency, n.d.; Jagayat, 2022).

In the early days of social media research, we would create social media simulations (e.g., Guadagno, Muscanell, Rice, & Roberts, 2013) using a now defunct online survey and experiment creation tool that the first author helped create called RiddleMeThis (RiddleMeThis.net). Specifically, this software was created in 2003 for use in the first author's dissertation and then later commercialized. It was the first customizable web-based data collection tool with menu-based survey creation options. A user could create a randomized experiment and assess dependent variables all in one form posted on the internet. The software would also time participants' responses and had a menu for easy viewing and exporting of data once a study was completed. Unfortunately, this product was eventually made obsolete by Qualtrics and other, similar products.

Collecting Digital Footprints

Every time a person posts on social media, is tagged in a photo, sends a direct message, leaves a review, or purchases a product from an online store, they create a digital trace of their activity called a digital footprint. Digital footprints such as these get aggregated into a digital dossier – a file containing detailed records about a person. Generally, people inadvertently create such records with their internet use. Between our social media profiles, our mobile phone use (i.e., texting, emails, apps, GPS), our online shopping, banking, and other online activities, quite a bit of information about us is available for others to find online.

Social media companies use this information to sell to advertisers, to predict what content will keep people engaged on their site, and what products, services, and entertainment will be most appealing to individual users. The more digital footprints we accumulate, the more accurate these predictive models and algorithms are.

Although this may sound unethical, it is generally legal for social media companies to do this (although some parts of the world, such as the European Union, are changing their laws to provide people with more control over their digital dossiers), and consent to participate in social media companies' research is also currently embedded in the end user agreements that people are supposed to read as part of the account creation process.

Researchers working outside of social media companies generally do not have the same access to people's digital dossiers owing to corporate privacy policies, but there are methods to obtain them. An early solution to this issue was a process called webscraping in which researchers would download social media data and participant profile information to use in their studies. In some instances, participants consented to have their data collected in this manner; in others, the data was simply collected without the consent of the content creators. Zimmer (2018) detailed some of the more notable instances of these practices becoming public.

One example is the 2016 case of researchers from Denmark who released a data set from the online dating service OkCupid. The data included identifying information about the participants, including their user ids, age, gender, geographic location, type of relationship sought, and personality characteristics. Thus, it is important that, when using this technique, participants provide consent for their data to be used in a study and any publicly released data set does not include identifying information. In my own work, my research team created a Facebook web scraper, but it was only used on participants who consented and allowed us to briefly friend them to collect the data (Guadagno, Loewald, et al., 2013).

It is important to note that this type of webscraping is typically prohibited by social media companies today. Twitter/X policy regarding webscraping is clearly expressed in their TOS: "crawling the Services is permissible if done in accordance with the provisions of the robots.txt file, however, scraping the Services without our prior consent is expressly prohibited." Similarly, TikTok states in their TOS the prohibition that scraping can "not access any data or TikTok content other than through the TikTok Research API (including without limitation, no use of scraping or other technical or manual techniques for extraction of content)." Facebook takes a more aggressive approach as they actively combat scraping, as stated in an article from 2021: "We devote substantial resources to combating unauthorized scraping on Facebook products. We have a dedicated External Data Misuse (EDM) team made up of more than 100 people, including data scientists, analysts and engineers focused on our efforts to detect, block and deter scraping." This is common practice for most social media platforms. They restrict data access through the use of their official API to control more tightly any data retrieval process. This is just one of multiple layers of a more general user data protection system that should be integral to any social media platform that prioritizes the safekeeping of its userbase sensitive data.

In the past, social media companies collaborated directly with academic researchers, but this practice has become markedly less common as scandal after scandal (e.g., Cambridge Analytica's illicit accessing of Facebook data and the emotional contagion study described below) was revealed by the news media. Nonetheless, there are options for researchers interested in collecting digital

footprints from social media. For instance, many social media companies provide licenses for academic researchers to obtain an Application Programming Interface (API) to view the social media application from the perspective of software developers and access and download any publicly shared information on users such as their posts, likes, shares, the prevalence of search terms, URL (uniform resource locator, a clickable link to a specific website) archives, and performance data in either real time or historically. See Table 3.1 for a non-exhaustive list of social media companies currently offering this option. Furthermore, there are myriad published guides to using and analyzing this type of data (e.g., Acker & Kreisberg, 2020; Barrie & Ho, 2022).

In addition to the API option, in which researchers obtain data directly from a social media application, there are various academic and corporate archives that are available to researchers and sometimes to the general public. For instance, Google Trends allows anyone with the URL to view what people are searching for in real time or historically. This site also allows people to view this data globally or specific to a geographic location. This methodology and the fascinating insights that can be gleaned from such data are described in detail by Seth Stephens-Davidowitz (2017). See Table 3.2 for more details as well as a list of other similar options for collecting digital footprints.

Table 3.1 *A non-exhaustive list of social media companies offering academic APIs*

Application	Link	Data description	Restrictions
Facebook/ Meta	https://fort.fb.com/researcher-apis	Provides anonymized access to people's posts in real time or historically, keyword search, and data from FB groups	Limited to US researchers and some EU countries
Reddit	www.reddit.com/r/redditdev/comments/ 1twwz2/new_api_access_for_ researchers_ academic_students	Access to real-time and historical posts based on keywords, user id, subreddits	University researchers
TikTok	https://developers.tiktok.com/products/ research-api	Account information such as profile and activity on the app, keywords, and videos with corresponding meta-data	Access limited to some members of our Content and Safety Advisory Councils and US academic researchers
Twitter/X	https://developer.twitter.com/en	Real-time or historical data	Non-commercial use only

Table 3.2 *A non-exhaustive list of pre-existing social media data sets*

Description	Link	Data Description	Restrictions
Facebook data sets	https://fort.fb.com/researcher-datasets	Offers different data sets for academic researchers (e.g., ad targeting, civic engagement, URL shares)	Researchers must apply for access with a research proposal
Google Ad Trends	https://trends.google.com	Database of Google search terms across time and geographic location	None – tool is free for use
Social Science One	https://socialscience.one/facebook-dataverse	Provides over 57 million URLs shared on Facebook along with aggregated data on the type of user interaction with the content (e.g., liked, shared, flagged as hate speech, fact checked, or commented)	RFPs must be submitted by a researcher with PI status (i.e., can serve as a PI on a grant application)
Twitter/X data sets	https://transparency.twitter.com/en/reports/moderation-research.html	Provides data on information operations and other data sets related to content moderation	The information operation data set is available to anyone, other data sets limited to members of the Twitter Moderation Research Consortium

Analysis Approaches

In the sections below, various techniques are reviewed, regarding the modeling of the spread of viral content and the role of machine learning in assisting with the analysis of digital footprints.

Modeling Viral Spread

One of the ways researchers measure the viral spread of content on social media is through the number of likes and shares a particular message receives. From this approach, it is widely known that the timing of a social media message (Phing & Yazdanifard, 2014), the emotional response evoked by the message (Guadagno, Rempala, Murphy, & Okdie, 2013), and the number of followers predict the likelihood of content going viral (Jenders et al., 2013). More recent approaches use data visualization – a graphical representation of data – to model the viral spread of content. For instance, the work of Professor Kate Starbird and colleagues has used this method to model the viral spread of disinformation. Specifically, this work

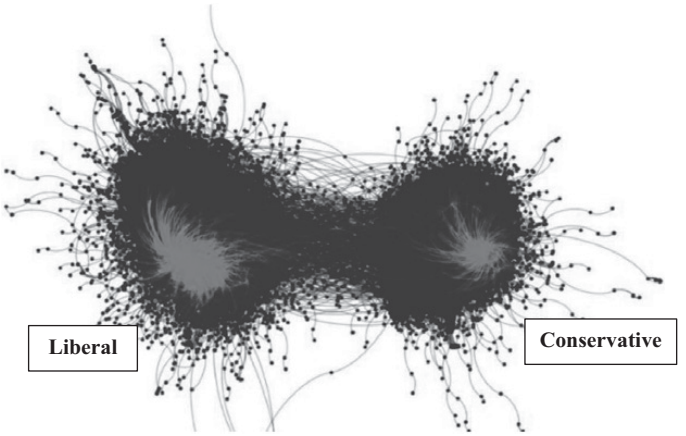

Figure 3.1 *Retweets of messages related to the Black Lives Matter social movement. Light gray indicates IRA bots and trolls, dark gray the other retweets. Reprinted from Arif et al. (2018)*

demonstrated how the Russian Internet Research Agency (IRA) used bots (i.e., computer programs pretending to be social media users) and trolls (i.e., employees of the IRA pretending to be Americans) to spread disinformation about the Black Lives Matter social movement (Arif et al., 2018). This work also illustrated how these messages differed as a function of whether the Russian bots and trolls were supporting this movement and pretending to be politically liberal Americans.

In the paper, the authors explain how they used a plugin of the Gephi software, Twitter Streaming Importer developed by Matthew Totet (https://github.com/totetmatt), to collect real-time data from Twitter, through the Twitter Streaming API. The team selected a series of hashtags as keywords, primarily associated with shootings and the Black Lives Matter movement ("gun shot," "gunman," "shooter," and "shooting," with the further addition of "BlackLivesMatter," "BlueLivesMatter," or "AllLivesMatter" – "*LM") to query Twitter and retrieved 58.8 million tweets of raw unfiltered data. After cleaning the initial data set, and applying filters to reduce interference from noise, the authors, using the community detection algorithms present in Gephi, were able to create a visualization of the retweet network among the accounts. This resulted in a clear grouping of the accounts into two, almost completely separated, communities. These communities remained consistently distinct even after tweaks and changes of the algorithm parameters.

The researchers, through this approach, were able to classify these two communities as divided by political lines (right and left leaning), basing their evaluation on the prevalent hashtags and most-followed accounts within each group. Moreover, they also discovered the most influential accounts in the two clusters. In the network, they were able to pinpoint the known RU-IRA accounts, giving a thorough description and visualization of this part of the online discourse revolving around #BLM and violent events involving shooting. Figure 3.1 reproduces a visual representation of

their findings, showcasing the descriptive power of a network visualization, and the value of tools like Gephi and the Twitter Streaming Importer.

Sentiment Analysis

Sentiment analysis refers to the analysis of text-based communications to infer psychological states (e.g., mood, opinions, motivations). Although there are myriad software products available to perform this task, the predominant software used in psychology and other social science disciplines is called the Language Inquiry Word Count (LIWC, pronounced "Luke"; Boyd et al., 2022) and has been used in over 20,000 studies. Sentiment analysis is also referred to as *opinion mining*, although the emphasis in this case is on extraction of psychological states.

The LIWC can analyze text written in many languages and, in addition to data on a person's psychological state, provides useful summary statistics such as the total number of words, mean words per sentence, and overall level of analytic thinking (Boyd et al., 2022). It also provides analyses of the focus of the text (e.g., social, perceptual, personal, and biological processes) and analyzes the cognitive processes (e.g., certainty, causation) expressed in the text. In addition, the software allows for the creation of custom dictionaries to focus on a particular topic. For instance, we once created a custom dictionary to assess prosocial behaviors in a study about helping in video games (Lee et al., 2017). Other studies have examined online civility (Ksiazek, 2015), task focus during a negotiation (Ireland & Henderson, 2014), and concerns about privacy (Vasalou et al., 2011) by creating custom dictionaries using the LIWC.

The LIWC currently is able to categorize any kind of text into more than 80 different language dimensions. It is also capable of providing a summary of some of the data dimensions of the text provided (e.g., the total number of words in a passage, the average number of words per sentence, the percent of words in the passages that were identified by the dictionary, and the percent of long words; Boyd & Pennebaker, 2015). Moreover, LIWC has integrated functionalities that can be used to classify text and then map it with psychological states, personal concerns, and punctuation. It is noteworthy to mention that the LIWC text analysis tools are able to uncover the various relationships between the linguistic characteristics of people's writing and their health, personalities, and thought processes (Boyd & Pennebaker, 2015). An interesting feature of this system is the ability to create ad hoc dictionaries. This ability paves the way for more focused research avenues; some examples of such ability is researching using terms related to privacy concerns (Vasalou et al., 2011), civility in online discussions (Ksiazek, 2015), and task engagement in negotiations (Ireland & Henderson, 2014).

One study that utilized the LIWC tools explored how people modified their online post habits, comparing a two-month period before the New York September 11 attacks and two months after them (Cohn et al., 2004). The author's approach focused on the analysis of more than 1,000 posts and discovered a sizable shift in the participants' psychology after the terrorist attacks. In the fortnight following these attacks, a spike in negative emotions was registered, and the participants

showed a heightened social and cognitive engagement. On the other hand, the social distance measured from their writing also showed signs of growth. The same group had all these changes reverted back to their original pre-attack levels after six weeks from the attack. The study clearly supports the efficacy of the LIWC. The tool also made it possible for the authors to delve deep into how people cope with and recover from traumatic events.

Another study, conducted by Ashokkumar and Pennebaker (2021), involved the analysis of Reddit language and large-scale survey data and also employed the LIWC tool to track the psychological effects of the COVID-19 pandemic. The authors found three distinct emotional phases: a "warning phase," an "isolation phase," and a "normalization phase." The LIWC scores were used to examine changes in a wide array of emotions such as anxiety, sadness, anger, and positive emotions. This revealed a surge in anxiety during the warning and isolation phases and a continued increase in sadness and decrease in positive emotions. LIWC analysis also highlighted a drop in analytic thinking during the isolation and normalization phases. This suggested a shift in the responses toward becoming more immediate, with a waning analytical thinking approach in response to the pandemic. The magnitude of these changes was seemingly directly correlated with the areas where COVID-19 was more virulent. The study showcases the utility of LIWC as a tool for understanding large-scale psychological responses to major societal events.

Further research has shown that the LIWC can be useful in predicting depression (De Choudhury et al., 2013), discovering people's personality characteristics (Park et al., 2015), and, presaging the #MeToo movement, detecting improvement in women's psychological well-being after tweeting about sexual harassment (Foster, 2015). Application software such as the LIWC has multiple times revealed its utility when properly employed by the researchers.

Machine Learning

It is recognized that machine learning (ML) techniques have had a huge influence in multiple research fields, but their reception in other traditional disciplines was somewhat colder. Jacobucci and Grimm (2020) propose that this difference might be explained by the unreliability of results, in turn exacerbated by measurement errors that can negatively influence the outcome. They demonstrated how measurement quality was the primary factor for electing to use ML or not, regardless of the sample size. This experiment validates a well-known but informal principle in computer science and information and communications technology known as "Garbage in, Garbage out" and highlights the fundamental importance of using the best practices in both data cleaning and wrangling. The general principle that the article proved still stands, but it is fundamental to acknowledge that the development of ML or deep learning (DL) tools is proceeding at breakneck pace. At the time of this writing (2023), results and error tolerance of newly developed algorithms could significantly differ from the ones presented in the article. Objects in pictures, text, and even context can now be recognized by new models and tools with more accuracy and speed than in the past. As an example, projects like ChatGPT,

Midjourney, and others from companies all over the world are creating new challenges and opportunities for researchers.

A brief analysis of the advantages and disadvantages of using ML techniques with internet research will follow. The huge amount of data that can be retrieved with internet research techniques is well suited for creating new ML models, as they need huge data sets to be able to generalize well and not overfit the data sample, as well as be effective as a predictive model (Zhou et al., 2017).

On the other hand, the researchers should be extremely aware of how biased data could create models that reinforce pre-existing biases. The need to correct unwanted bias unfortunately invites the possible introduction of personal biases in the attempt to clean the data set. The researcher should strike a difficult balance between the two issues – to remove as much bias as possible without falling into data manipulation to achieve the desired outcome. The situation is further complicated by the various definitions and interpretations of the word "bias" across different disciplines (Hellström et al., 2020). It is worth noting that the issue of bias is present both in ML approaches and in internet research, but the subject is particularly debated in the field of ML. This entire subject is vast and complex, and integrating ML techniques in an internet research project will come with its strengths and weaknesses that should always be kept in mind.

Ethical Concerns

"We spent $1 m harvesting millions of Facebook profiles"

Accounts vary on who came up with the idea of using social media for political microtargeting. Nonetheless, the role of Cambridge Analytica (CA) in using this methodology to influence political beliefs is incontrovertible. In the months leading up to the 2016 US presidential election, Cambridge Analytica was the primary proponent of a new form of internet advertising targeting people based on their personal characteristics. This approach is known as political microtargeting – "monitoring people's online behavior, and using the collected data, sometimes enriched with other data," with the final scope of influencing its targets, deploying "individually targeted political advertisements" (Dobber et al., 2019). CA was more than a data science and data analytics firm; it marketed itself as a "full-service propaganda machine" (Cadwalladr et al., 2021) and was led by Alexander Nix. CA set up a fake office in Cambridge to create an impression that they were based out of the University of Cambridge and sold their services to political candidates in the United States and elsewhere.

Their research methods were ethically and legally questionable. Specifically, in 2014, CA collected Facebook profile data, without consent, from millions of Facebook users. The operation involved the use of *thisisyourdigitallife* – an app where users knowingly and voluntarily shared their data as subjects of a personality test for a monetary reward. Unbeknownst to them, the app also collected data from

their social media connections, leading to the creation of an enormous database with tens of millions of datapoints from completely unaware users (Graham-Harrison & Cadwalladr, 2021). This data was used to create a system to microtarget voters with political advertisements tailored to them based on information such as their geographic location and personality profile.

While a smaller subsample of participants were paid to provide access to their Facebook profiles, CA collected more than participant data. CA's application collected profile data on everyone each participant was connected to on Facebook. These data were then used to create a microtargeting algorithm that would predict how to influence each person's political opinions by predicting the kind of advertising content that each person would find persuasive and the number of times these ads needed to be shown to their targets. As a result, every voter whose information had been collected was targeted by CA's algorithm and sent different advertisements leading them to different internet content.

It is worth noting that CA managed to retrieve data from millions of users without the consent of the majority of them. The scale of this endeavor would not have been possible without the consent and prolonged lack of concern regarding the data leak from Facebook itself. At least 87 million Facebook users had their data taken in this manner (Solon, 2018). Eventually, a data scientist within Cambridge Analytica, Christopher Wylie, went to the press and blew the whistle on the company's unethical conduct (Graham-Harrison & Cadwalladr, 2021; Cadwalladr et al., 2021). Once people knew what CA was doing, the company folded as governments and law enforcement started investigating the legality of their practices. CA's actions became one of the biggest scandals of the social media era. Not only did CA interfere in the 2016 US presidential election, and attempt to influence voters in favor of the Republican candidate, but it also has been widely accused of influencing the Brexit vote.

Another high-profile example comes from academia. A study on the virality of emotions online made headlines and stirred discussions both outside and inside various academic circles. During the summer of 2014, news regarding the manipulation of users' Facebook feeds and nonconsensual data collection by Facebook started to emerge. The headlines conveyed a sense of outrage, shock, resignation, and helplessness that properly reflected the emotional impact of the experiment Facebook conducted. In this experiment, data from 700,000 users was collected without their knowledge or consent.

This experiment took place in 2012 and lasted a week. For the duration of the experiment, Facebook collected data from several hundred randomly selected participants from their userbase. Their research question focused on whether positive and negative emotions were "contagious" and spread on social media. Moreover, the experiment comprised two different studies with a test and a control group for each. The user news feed of the first test group was altered with a reduced number of positive terms (e.g., love, nice, sweet; Pennebaker et al., 2007). The second test group was similarly manipulated but with negative terms instead (e.g., hurt, ugly, nasty). Both of the control groups had random posts removed from their news feed, in similar numbers to their paired test group.

The authors analyzed more than 3 million posts with a text analysis tool – the LIWC – described in detail earlier in this chapter. The objective was to determine whether strong positive or negative emotions could be viral and spread through the user's Facebook social network. The two dependent variables were the quantity of either positive or negative emotions expressed by the test subjects. The researchers found that their initial prediction that emotions would spread through social media was supported, such that users with a negatively manipulated news feed shared more negative emotions than the control group. Similar results emerged for the group where positive emotions were artificially enhanced; there was a major increase in positive emotions spread.

When the public was made aware of this experiment, the reactions were mostly negative, focusing especially on the lack of proper ethical considerations. The only reference to them was a statement where previous consent was mentioned: "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, p. 8789). In the academic community, this lackluster approach produced discussions regarding the procedures employed with human subjects on social media and the ethical treatment of these participants.

The social and personality psychology community and related social sciences were especially involved, as the paper directly connected to those disciplines. Even the editor of the journal that published the paper, Dr. Inder Verma, expressed editorial concerns regarding the lack of informed consent – both a standard practice in any research involving human test subjects and a requirement for any research using federally funded institutions (corporations such as Facebook are not currently held accountable to the same rules; Verma, 2014).

While conducting internet research, there are four primary pitfalls to always keep in mind, so as to avoid unethical behaviors. The first consideration is the anonymity and confidentiality of the participants. The second one focuses on privacy concerns. The third involves validating the veracity of the responses and the non-naivete of the participants. Finally, the fourth one addresses fair compensation. Academics have tackled these questions already, leading to the development of a comprehensive set of recommendations and guidelines to guide others in producing ethically conscious experiments. An example of a guideline comes from the Association of Internet Researchers (AoIR). In their publication they address ethics in internet research (Markham & Buchanan, 2012) in which examples and insights are given to readers approaching the field. Some interesting questions that are put forward are: Is an avatar a person? What are the long-term ethical considerations of directly quoting someone's online posts? and What are the risks for participants when a technology prevents the removal of identifying information (e.g., facial recognition data)?

Another perspective on ethical issues in internet research is provided by members of university IRBs. In one study by Buchanan and Hvizdak (2009), these IRB members raised attention on topics such as security of data, participant anonymity, data deletion/loss and recovery, access to electronic data, validity of data collected online, the impact of anonymity on participants' responses, age verification, consent,

and perception of the university. In the sections below, we discuss the first two issues, as they are relevant to all types of research on social media reviewed in this chapter.

Participant Anonymity, Confidentiality, and Consent

We outlined above four main ethical concerns regarding internet research. The first of these is related to obtaining informed consent from the participants while also being able to shield their identities and other sensitive information. Gosling and Mason (2015) highlighted some issues related to the anonymity of internet research. They wrote that "researchers have less control over and knowledge of the research environment and cannot monitor the experience of participants, or indeed their true identities" (Buchanan & Williams, 2010, p. 894). In contrast to a traditionally offline lab experiment, where the researchers have the possibility to debrief participants who decided to withdraw, in an internet-based experiment, where the participants have been recruited with the methods previously discussed in the chapter, they cannot be properly debriefed after an early dropout. Moreover, the researcher can't be sure that each and every participant has fully read and understood the debriefing (Miller et al., 2017). It is also worth noting how research conducted with anonymous participants is a double-edged sword. On the one hand it protects the identity of the participants; on the other hand, the researchers have new concerns regarding proper consent and debrief, double participation, and difficulties in case of needed direct intervention to help a distressed participant.

Furthermore, true anonymity is also difficult to achieve. In one experiment, Dawson (2014) was able to retrieve the source of the text from 10 out of 112 target articles. From the 10 papers he had source material from, only one paper reported receiving IRB approval to publish identifiable data. Of the remaining nine studies, five neither anonymized the text nor discussed ethical considerations, and one tried but failed to anonymize the data. The author summarized how it is possible, with the right expertise, motivation, or data, to deanonymize experiment participants. We, as researchers, should realize that, although it is important to safeguard the participants' anonymity and confidentiality, the anonymity itself can also limit certain research avenues in which it would be informative or useful to know the identities of the research participants. The advice of these authors is to use digital data with consent of the people providing the content if possible. If this is not possible, the data must be truly anonymized by a thorough check of the data to ensure that individual participants cannot be identified.

Privacy Concerns

Most people have an interest in knowing their past and their roots, and one of the most interesting startups that exploits this curiosity is the tech startup 23andme.com. The company extracts and analyzes DNA samples from saliva provided by their customers and sent to the company via mail. After the analysis, 23andme.com reaches out to the customer and gives them a variety of information about their genetic predispositions (e.g., likelihood of flushing after consuming alcohol, lactose

intolerance, being under- or overweight), DNA ancestry, and informs consumers if they are carriers for congenital diseases (e.g., Tay Sachs, Alzheimer's). This means that information regarding genetic makeup and other sensitive information is stored on the servers of this company. The implications of a data leak could be severe, potentially leading to violations of customer privacy and unexpected negative consequences such as discrimination based on genetic predispositions.

This small example should demonstrate how privacy is one of the main ethical concerns for internet research, and the more data-rich the internet becomes, the more prominent this issue will become. Researchers should also be mindful of how privacy standards can be vastly different from jurisdiction to jurisdiction. Some of the most well-known privacy laws are the European General Data Protection Regulation (GDPR), the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA), and the United States California Consumer Privacy Act (CCPA). Even with this variability between jurisdictions, it is worth remembering the preeminent position of the GDPR on the world stage. This piece of legislation is seen as a benchmark for any other privacy legislation, and it has influenced global discourse on topics including personal data processing and storage. Article 5 – "Principles relating to processing of personal data" – explicitly states that personal data shall be treated lawfully, fairly, and transparently; they should be collected in line with the stated purpose and not processed outside the original scope. Data also must be accurate and up to date, their storage should not exceed the necessary time based on the purpose of collection, and the data should remain integral and confidential. For all of those processes, the entity who controls the data is accountable for the proper handling and storing and must demonstrate compliance with the GDPR legislation.

This fundamental point about privacy had already been noticed by Gosling and Mason (2015). They argued how the many ethical concerns found in internet research are born from outdated guidelines and rules that were defined before the widespread diffusion of internet research itself. In addition to normal procedures about informed consent and debriefing, they argued how the definition of "public behavior" should be a factor in determining whether the data collected is to be considered public or private. This is especially true when researchers are planning to use webscraping techniques or APIs collecting data from social media, forums, message boards, and other online discussions.

Most internet users, or "netizens," do not realize how their continuous presence on the internet will inevitably lead to the creation of a digital footprint. The more time they spend online in almost any capacity, the more data about their personal life, thoughts, and activities will seep through from the offline to the online world. Though almost all online activities leave traces, social media applications, such as Facebook, LinkedIn, and Google, often contribute the most to this digital footprint. To make the situation worse, these companies provide the option for people to log in to third-party websites, allowing an easier than ever tracking of online activities of specific users.

Ease of access and use is always a trade-off with people's privacy, and companies are not always able to strike the correct balance between confidentiality, integrity,

and availability of data. Switching to a researcher's perspective, these online breadcrumb trails are invaluable, as they provide easy means for tracking people's online activities on different websites, and they give insights on technology use and perception. On the other hand, the privacy of the people under investigation will be in danger, both offline and online. This is still an open issue, as there is still no academic consensus among internet researchers. However, there are interesting considerations about determining whether a specific internet behavior is public behavior or not and whether it can be observed and/or recorded without consent (Buchanan and Williams, 2010).

Conclusion

Topics, narratives, and modes of interaction are always shifting in new, unique, and interesting ways. In this chapter, we tried to give an extensive review of contemporary research methods while also talking about common pitfalls and malpractices. Gosling and Mason (2015) make clear that internet research is giving scientists both more new avenues of study but also new challenges and ethical issues. Despite those challenges, the positives that these new technologies put forward by far outweigh any drawbacks. In their work, they also stress the point that researchers must be careful while studying events and people on the internet, as the targets of our research are constantly moving, and it is difficult to follow the continuous new iterations and developments. As the field of internet research continues to evolve, it is essential that researchers remain informed about new tools, techniques, and ethical considerations. Looking ahead, the field will undoubtedly present new challenges and opportunities, and it is our responsibility as researchers to navigate these with integrity and a commitment to ethical research practices.

References

Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*, *20*, 105–123.

Albertson, B., & Gadarian, S. (2014, July 1). Was the Facebook emotion experiment unethical? *Washington Post*. www.washingtonpost.com/news/monkey-cage/wp/2014/07/01/was-the-facebook-emotion-experiment-unethical/?utm_term=.15088275b53c

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, *45*(6), 842–850.

Antoun, C., Zhang, C., Conrad, F. G., & Schober, M. F. (2016). Comparisons of online recruitment strategies for convenience samples: Craigslist, Google AdWords, Facebook, and Amazon Mechanical Turk. *Field methods*, *28*(3), 231–246.

Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the part: Examining information operations within #BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–27.

Ashokkumar, A., & Pennebaker, J. W. (2021). Social media conversations reveal large psychological shifts caused by COVID-19's onset across US cities. *Science Advances*, *7*(39), eabg7843.

Barrie, C., & Ho, J. C. (2022). *Using the Twitter Academic API with R for Social Science Research*. SAGE Publications. https://dx.doi.org/10.4135/9781529609233

Bayer, J. B., Triệu, P., & Ellison, N. B. (2020). Social media elements, ecologies, and effects. *Annual Review of Psychology*, *71*, 471–497.

Bentley, F. R., Daskalova, N., & White, B. (2017, May). Comparing the reliability of Amazon Mechanical Turk and Survey Monkey to traditional market research surveys. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems* (pp. 1092–1099). ACM.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political analysis*, *20*(3), 351–368.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The Development and Psychometric Properties of LIWC-22*. University of Texas at Austin.

Boyd, R. L., & Pennebaker, J. W. (2015). A way with words: Using language for psychological science in the modern era. In C. V. Dimofte, C. P. Haugtvedt, & R. F. Yalch (eds.), *Consumer Psychology in a Social Media World* (pp. 222–236). Routledge.

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68.

Buchanan, E. A., & Hvizdak, E. E. (2009). Online survey tools: Ethical and methodological concerns of human research ethics committees. *Journal of Empirical Research on Human Research Ethics*, *4*(2), 37–48.

Buchanan, T., & Williams, J. E. (2010). Ethical issues in psychological research on the internet. In S. D. Gosling & J. A. Johnson (eds.), *Advanced Methods for Conducting Online Behavioral Research* (pp. 255–271). American Psychological Association.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Burger, J. M. (2014). *Personality*. Cengage Learning.

Butler, L., Lamont, P., Wan, D. L. Y., Prike, T., Nasim, M., Walker, B., et al. (2022). The (mis) information game: A social media simulator. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02153-x

boyd, D. M., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Cadwalladr, C., Khalili, M., Phillips, C., Silver, M., Jenkins, A., Search, J., et al. (2021). Cambridge Analytica whistleblower: "We spent $1m harvesting millions of Facebook profiles" [video]. *Guardian*, March 17. www.theguardian.com/uk-news/video/2018/mar/17/cambridge-analytica-whistleblower-we-spent-1m-harvesting-millions-of-facebook-profiles-video

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160.

Chambers, C. (2014). Facebook fiasco: Was Cornell's study of "emotional contagion" an ethics breach? *Guardian*, July 1. www.theguardian.com/science/head-quarters/2014/jul/01/facebook-cornell-study-emotional-contagion-ethics-breach

Chambers, M., Bliss, K., & Rambur, B. (2020). Recruiting research participants via traditional snowball vs Facebook advertisements and a website. *Western Journal of Nursing Research*, *42*(10), 846–851.

Chandler, J. (2023). Participant recruitment. In A. Nichols & J. E. Edlund (eds.), *Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences* (vol. 1, pp. 179–201). Cambridge University Press.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, 12, 53–81.

Clark, T., & Blackhart, G. (2023). Debriefing and post-experimental procedures. In A. Nichols & J. E. Edlund (eds.), *Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences* (vol. 1, pp. 244–265). Cambridge University Press.

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics*, *2*(4), 2053168015622072.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, *15*(10), 687–693.

Dawson, P. (2014). Our anonymous online research participants are not always anonymous: Is this a problem? *British Journal of Educational Technology*, *45*(3), 428–437.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *ICWSM*, *13*, 1–10.

Dobber, T., Ó Fathaigh, R., & Zuiderveen Borgesius, F. J. (2019). The regulation of online political micro-targeting in Europe. *Internet Policy Review*, *8*(4). https://policyreview.info/articles/analysis/regulation-online-political-micro-targeting-europe

Edlund, J. E., Lange, K. M., Sevene, A. M., Umansky, J., Beck, C. D., & Bell, D. J. (2017). Participant crosstalk: Issues when using the Mechanical Turk. *Tutorials in Quantitative Methods for Psychology*, *13*(3), 174–182.

Foster, M. D. (2015). Tweeting about sexism: The well-being benefits of a social media collective action. *British Journal of Social Psychology*, *54*(4), 629–647.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, 66, 877–902.

Graham-Harrison, E., & Cadwalladr, C. (2021). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *Guardian*, September 29. www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Grimm, P. (2010). Social desirability bias. In C. L. Cooper (ed.), *Wiley International Encyclopedia of Marketing*. Wiley.

Guadagno, R. E. (2019). Using the internet for research. In J. E. Edlund & A. L. Nichols (eds.), *Advanced Research Methods for the Social and Behavioral Sciences* (pp. 68–82). Cambridge University Press.

Guadagno, R. E. (in press). *Psychological Processes in Social Media: Why We Click*. Academic Press.

Guadagno, R. E., Loewald, T. A., Muscanell, N. L., Barth, J. M., Goodwin, M. K., & Yang, Y. (2013). Facebook history collector: A new method for directly collecting data from Facebook. *International Journal of Interactive Communication Systems and Technologies (IJICST)*, *3*(1), 57–67.

Guadagno, R. E., Muscanell, N. L., & Pollio, D. E. (2013). The homeless use Facebook?! Similarities of social network use between college students and homeless young adults. *Computers in Human Behavior*, *29*(1), 86–89. https://doi.org/10.1016/j.chb.2012.07.019

Guadagno, R. E., Muscanell, N. L., Rice, L. M., & Roberts, N. (2013). Social influence online: The impact of social validation and likability on compliance. *Psychology of Popular Media Culture*, *2*(1), 51–60.

Guadagno, R. E., Rempala, D. M., Murphy, S., & Okdie, B. M. (2013). What makes a video go viral? An analysis of emotional contagion and internet memes. *Computers in Human Behavior*, *29*(6), 2312–2319.

Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in machine learning: What is it good for? [preprint]. *Arxiv*. https://doi.org/10.48550/arXiv.2004.00686

Hill, K. (2014). Facebook manipulated 689,003 users' emotions for science. *Forbes*, June 28. www.forbes.com/sites/kashmirhill/2014/06/28/facebook-manipulated-689003-users-emotions-for-science/#593f5cbf197c

Ireland, M. E., & Henderson, M. D. (2014). Language style matching, engagement, and impasse in negotiations. *Negotiation and Conflict Management Research*, *7*(1), 1–16.

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

Jagayat, A. (2022, October 4). Mock social media website tool. *OSF*. https://osf.io/m2xd8

Jenders, M., Kasneci, G., & Naumann, F. (2013, May). Analyzing and predicting viral tweets. In Daniel Schwabe (ed.), *Proceedings of the 22nd International Conference on World Wide Web* (pp. 657–664). ACM.

Kendall, C., Kerr, L. R., Gondim, R. C., Werneck, G. L., Macena, R. H. M., Pontes, M. K., et al. (2008). An empirical comparison of respondent-driven sampling, time location sampling, and snowball sampling for behavioral surveillance in men who have sex with men, Fortaleza, Brazil. *AIDS and Behavior*, *12*, 97–104.

Kern, M. L., McCarthy, P. X., Chakrabarty, D., & Rizoiu, M. A. (2019). Social media-predicted personality traits and values can help match people to their ideal jobs. *Proceedings of the National Academy of Sciences*, *116*(52), 26459–26464.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790.

Ksiazek, T. B. (2015). Civil interactivity: How news organizations' commenting policies explain civility and hostility in user comments. *Journal of Broadcasting & Electronic Media*, *59*(4), 556–573.

Lee, J., Gillath, O., Kimbrough, A. M., & Guadagno, R. E. (2017, January). Development and validation of helping in gaming scales. Poster presented at the Media Psychology Preconference, San Antonio, TX.

Lee, J., & Spratling, R. (2019). Recruiting mothers of children with developmental disabilities: Adaptations of the snowball sampling technique using social media. *Journal of Pediatric Health Care*, *33*(1), 107–110.

Markham, A., & Buchanan, E. (2012). *Recommendations from the AoIR Ethics Working Committee* (Version 2.0).

Merriam-Webster. (n.d.). Crowdsourcing. In Merriam-Webster.com [dictionary]. www.merriam-webster.com/dictionary/crowdsourcing (retrieved March 6, 2024).

Miller, J. D., Crowe, M., Weiss, B., Maples-Keller, J. L., & Lynam, D. R. (2017). Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's Mechanical Turk. *Personality Disorders: Theory, Research, and Treatment*, *8*(1), 26–34.

Murray, M. (2014). Users angered at Facebook emotion-manipulation study. *Today*, June 30. www.today.com/health/users-angered-facebook-emotion-manipulation-study-1D79863049

Muscanell, N. L., & Guadagno, R. E. (2012). Make new friends or keep the old: Gender and personality differences in social networking use. *Computers in Human Behavior*, *28* (1), 107–112. https://doi.org/10.1016/j.chb.2011.08.016

Nichols, A., & Edlund, J. E. (2020). Why don't we care more about carelessness? Understanding the causes and consequences of careless participants. *International Journal of Social Research Methodology*, *23*(6), 625–638. https://doi.org/10.1080/13645579.2020.1719618

Oinas-Kukkonen, H., & Oinas-Kukkonen, H. (2013). *Humanizing the Web: Change and Social Innovation*. Palgrave Macmillan.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., et al. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952.

Parker, C., Scott, S., & Geddes, A. (2019). *Snowball Sampling*. SAGE Publications. https://dx.doi.org/10.4135/9781526421036831710

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. LIWC. www.liwc.net/LIWC2007LanguageManual.pdf

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. LIWC. www.researchgate.net/publication/246699633_Linguistic_inquiry_and_word_count_LIWC

Phing, A. N. M., & Yazdanifard, R. (2014). How does ALS ice bucket challenge achieve its viral outcome through marketing via social media? *Global Journal of Management and Business Research*, *14*(E7), 57–64.

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(12). https://doi.org/10.1057/s41599-019-0279-9

Ross, J., Zaldivar, A., Irani, L., & Tomlinson, B. (2009). Who are the Turkers? Worker demographics in Amazon Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (pp. 2863–2872). ACM.

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, *43*, 304–307.

Prevency. (n.d.). The solution for a realistic social media simulation. https://socialmediasimulator.com

Solon, O. (2018). Facebook says Cambridge Analytica may have gained 37m more users' data. *Guardian*, April 4. www.theguardian.com/technology/2018/apr/04/facebook-cambridge-analytica-user-data-latest-more-than-thought

Stephens-Davidowitz, S. (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us about Who We Really Are*. HarperLuxe.

Vasalou, A., Gill, A. J., Mazanderani, F., Papoutsi, C., & Joinson, A. (2011). Privacy dictionary: A new resource for the automated content analysis of privacy. *Journal of the Association for Information Science and Technology*, *62*(11), 2095–2105.

Vaughn-Nichols, S. J. (2014). We're all just lab rats in Facebook's laboratory. *ZDNet*, June 30. www.zdnet.com/article/were-all-just-lab-rats-in-facebooks-laboratory/

Verma, I. M. (2014). Editorial expression of concern: Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(29), 10779–10779.

Wang, T., Brede, M., Ianni, A., & Mentzakis, E. (2017, February). Detecting and characterizing eating-disorder communities on social media. In M. de Rijke (ed.), *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp. 91–100). ACM.

Wright, S. A., & Goodman, J. K. (2019). Mechanical Turk in consumer research: Perceptions and usage in marketing academia. In F. R. Kardes, P. M. Herr, & N. Schwarz (eds.), *Handbook of Research Methods in Consumer Psychology* (pp. 338–357). Routledge.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350–361.

Zimmer, M. (2018). Addressing conceptual gaps in big data research ethics: An application of contextual integrity. *Social Media + Society*, *4*(2), 2056305118768300.