Article

SM+S
social media + society

Social Media + Society
October-December 2023: 1–12

DOI: 10.1177/20563051231216947
journals.sagepub.com/home/sms
Sage

Disclosure Standards for Social Media and Generative Artificial Intelligence Research: Toward Transparency and Replicability

Ganna Kostygina, Yoonsang Kim, Zachary Seeskin, Felicia LeClere, and Sherry Emery

Abstract

Social media dominate today's information ecosystem and provide valuable information for social research. Market researchers, social scientists, policymakers, government entities, public health researchers, and practitioners recognize the potential for social data to inspire innovation, support products and services, characterize public opinion, and guide decisions. The appeal of mining these rich datasets is clear. However, there is potential risk of data misuse, underscoring an equally huge and fundamental flaw in the research: there are no procedural standards and little transparency. Transparency across the processes of collecting and analyzing social media data is often limited due to proprietary algorithms. Spurious findings and biases introduced by artificial intelligence (AI) demonstrate the challenges this lack of transparency poses for research. Social media research remains a virtual "wild west," with no clear standards for reporting regarding data retrieval, preprocessing steps, analytic methods, or interpretation. Use of emerging generative AI technologies to augment social media analytics can undermine validity and replicability of findings, potentially turning this research into a "black box" enterprise. Clear guidance for social media analyses and reporting is needed to assure the quality of the resulting research. In this article, we propose criteria for evaluating the quality of studies using social media data, grounded in established scientific practice. We offer clear documentation guidelines to ensure that social data are used properly and transparently in research and applications. A checklist of disclosure elements to meet minimal reporting standards is proposed. These criteria will make it possible for scholars and practitioners to assess the quality, credibility, and comparability of research findings using digital data.

Keywords

social data quality, reproducibility, reporting standards, scientific transparency, disclosure

Introduction

Social media are ubiquitous in today's communications environment. Once considered as recreational networks mainly used by youth and younger adults, social media now are used by corporations, news media, advocacy groups, and individuals of various ages and socioeconomic backgrounds. Since each post or upload leaves a digital footprint, social media generate an enormous quantity of data, creating unique opportunities for analyzing important questions about society, policy, and health (Schillinger et al., 2020). Corporations, academic researchers, government, and nonprofit organizations have begun to rely on these data to gauge people's attitudes toward products, marketing, and proposed policies; and to characterize public opinion and individual behavior (Bruns, 2013; Bruns & Stieglitz, 2014; Cohen & Ruths, 2013; Diakopoulos, 2016; Y. Kim et al., 2016; Kostygina et al., 2016; Tufekci, 2014; Yom-Tov, 2016).

The recent emergence of generative artificial intelligence (AI) tools (e.g., ChatGPT) represents similar opportunities and challenges (Salah et al., 2023). Leveraging the advanced capabilities of these technologies to analyze multiple streams and extensive volumes of data generated daily on social

NORC at the University of Chicago, USA

Corresponding Author:
Ganna Kostygina, Social Data Collaboratory, NORC at the University of Chicago, 55 East Monroe Street, 3165, Chicago, IL 60603, USA.
Email: kostygina-anna@norc.org

media with greater efficiency and speed can lead to an unprecedented depth and breadth of understanding of social phenomena by identifying patterns of information flow on previously unattainable scale, and model social dynamics and social contagion across platforms (Elmas & Gül, 2023; Haluza & Jungwirth, 2023). This can inform and enable significant advancements in social science and public opinion research at every step from problem definition, to data collection, analysis, and interpretation. However, there are no clear guidelines for conducting research with the help of generative AI tools or standards for assessing the quality of this research. It remains unclear whether such analyses can be reproducible or replicable due to the lack of transparency of generative AI models and potential innate undetected algorithm biases that can compromise the impartiality and validity of research findings, leading to skewed interpretations and inaccurate conclusions (Dwivedi et al., 2023; Mehrabi & Pashaei, 2021). Social media and generative AI are revolutionizing social science and public opinion research, which highlights the need to translate the social science transparency and replicability standards for this new media and technological landscape and update the social science data quality assessment guidelines, as well as disclosure standards and requirements.

The rush to take advantage of the bounty the rich social data offer occurs at a time of substantial public distrust of science and technology in general (Desmond, 2022; Kabat, 2017; Winter et al., 2022). This trend follows waves of controversy over suspect or failed experiments using digital data to gauge public opinion formation (Albergotti, 2014; Booth, 2014) and assess health trends (Lazer & Kennedy, 2015), and the harvest of Facebook profile data without user permission during the 2016 US presidential campaign (Rosenberg et al., 2018). According to the 2022 Pew Research Center, public trust in science also decreased following the COVID-19 pandemic, with only 29% of US adults reporting a great deal of confidence in scientists to act in the public's best interests in December 2021 (Kennedy et al., 2022). Cynicism or disbelief in science has increased to an extent that the research, government, and business communities interested in promoting scientific and technological progress cannot ignore (Kabat, 2017).

The emergence of new generative AI technologies introduces new problems for social data research. For instance, competition between such social media platforms and generative AI systems resulted in growing restrictions of social media data access and use (e.g., for X—formerly Twitter—and Reddit) for academic, organic, and commercial users due to unlicensed or unauthorized use of copyrighted proprietary digital data by these systems to train their generative AI models or build algorithms (Vincent, 2023). The capacity of ChatGPT and other generative AI to produce simulated social media posts and images can further undermine trust in what constitutes valid social data.

To help regain public confidence, prominent communication scholars have called for efforts to build transparency by establishing a climate of critique and self-correction; fully acknowledging the limitations in data, tools, and methods; accounting for seemingly anomalous data; and clearly, precisely specifying key terms (Hall Jamieson 2015). Researchers have to consider privacy and data provenance when using emerging AI technologies for social data analysis and processing.

We believe that the broad principles of transparency articulated previously to enhance credibility of science (Aczel et al., 2020; Hall Jamieson, 2015) can be applied to establish common disclosure requirements for social media and generative AI research. If we set clear reporting guidelines for social data acquisition, management, quality assessment, and analysis, public trust in the scientific findings and integrity of such research may increase, or at the minimum, research findings can be replicated or refuted, increasing scientific integrity.

Even as the number of research studies using digital data rapidly grows, relatively few have transparently outlined their data collection and analysis methods. Gradually, researchers have begun to critically examine the assumptions behind social media data findings, reproducibility, generalizability, and representativeness and call for higher transparency in documenting methods for such studies (Assenmacher et al., 2022; boyd & Crawford, 2012; Bruns, 2013; Center for Democracy & Technology n.d.; Cockburn et al., 2020; Council for Big Data, Ethics, and Society, n.d.; Fairness, Accountability, and Transparency in Machine Learning, n.d.; Fineberg et al., 2020; González-Bailón et al., 2014; Goroff, 2015; Graham et al., 2013; Jurgens et al., 2015; Y. Kim et al., 2016; Reed & boyd, 2016; Tufekci, 2014).

Challenges and Limitations of Social Data Research

As with any data source, the way in which social data are collected for research influences the conclusions that can be drawn (Japec et al., 2015). Although each social media platform has different technical constraints and poses unique methodological and programming challenges, there are common decisions that any project must address. Biases and other data quality issues arise from decisions researchers make about the platform selected and how the data are accessed, retrieved, processed, or filtered (or cleaned). In turn, each decision affects data quality and the validity of inferences based on the data analytics.

A number of specific limitations and challenges to conducting social data research have been described in the literature over 15 years since social media gained popularity. The challenges and limitations may be categorized as related to data collection, processing, analysis, and interpretation stages of inquiry. At the data collection stage, data-gathering

approaches may be opportunistic; for example, studies based on retrieving information using specific hashtags often abstract conversations from a much more complex communications universe; such analyses risk omitting context and creating and describing new realities which may not reflect lived experience (Bruns, 2013). Furthermore, infrastructure may be unreliable, subject to outages and losses during data collection; and the choice of methods to combine multiple data sources may result in potential bias and errors. In addition, platform terms of service restrict data sharing, preventing replication of research using the same dataset. Therefore, data-gathering efforts are often duplicated and uncertainty exists regarding dataset comparability (Bruns, 2013).

During the data preprocessing and analysis stages of inquiry, design decisions for cleaning and interpreting social data—that is, selecting which attributes and variables to count and which to ignore—are inherently subjective (boyd & Crawford, 2012), and there is no known best practice or standard. Tools and methodologies for processing digital data are continuously evolving, and sometimes pieced together from various platforms and technologies, making documentation and replication problematic. Some researchers alternatively turn to commercial analytics services or standardized tools which may operate as black box enterprises, or contain processing steps that lie outside the researcher's expertise to clarify (Bruns, 2013). Cross-platform analyses pose challenges because the data often appear in different formats that are difficult to combine, for example, text, images, and hyperlinks (Voytek, 2017).

Decision-making during the data collection and analyses stages impacts validity of research findings, interpretations, and conclusions as managing and interpreting the context in which conversations occur as well as implementing rigorous evaluation of the generated outputs to prevent the inadvertent propagation of biases or inaccuracies represent ongoing challenges for social data analysis.

Although these challenges and limitations are widely recognized as important, they are often neglected or dismissed in practice (e.g., Bruns & Stieglitz, 2014; Y. Kim et al., 2016). Disclosure of the decisions made during the conduct of social data research, and the reasons behind them, could dramatically enhance transparency and replicability. Without such reporting, evaluating the validity of findings and comparing methods and results across studies become impossible.

Validity Threats in the Social Media and AI Research Pipeline

Like traditional public opinion research, social data research methods—such as choice of platform, sampling strategy, and search filters for data collection—may affect the results and conclusions and have implications for a study's external, internal, and construct *validity* (Cook & Campbell, 1979).

Construct validity is the degree to which a study measures what it purports to measure. Reporting procedures for search filter construction and search filter assessment are critical for ensuring construct validity and reliability of social data measurement. Face validity (i.e., the extent to which a study or test appears to measure what it claims to measure based on face value) is often subjective and insufficient to support construct validity; we need objective criteria to assess search filter quality (Bagby et al., 2006). For example, poor construct validity for surveillance tools using social media data will lead to false discovery or false non-discovery. Objective measures that provide insight toward inferring false positive rates and false negative rates will help toward a proper interpretation.

Internal validity can be defined as a way to gauge whether the correct analyses are used to answer the research questions. Disclosure of analytic procedures (e.g., classifier type and training, performance measures, and quality assessment) is imperative to maintain internal validity in social data research.

External validity represents the validity of generalized inferences in scientific research. It is a criterion for assessing the level of generalizability of study findings in relation to the outside world or the larger population outside the study context. For social data research, platform selection is a critical step to ensure generalizability of findings to the larger population of interest as demographics of main users differ across platforms and platforms have different functions. Therefore, disclosure of the rationale for platform selection, including explaining whether the platform offers appropriate depth, format, mode of content, amount, timing, and representativeness of the target population, is essential to safeguard external validity. When these different types of validity are questionable, is it still worth using the social data? It would depend on the study purpose; therefore, it is important to evaluate the data in regard to these aspects and consider the implications.

Hsieh and Murphy (2017) proposed the Total Twitter Error (TTE) framework for social media data quality assessment, which recognizes that population coverage—or generalizing to the population as a whole—may not always be the goal of social media analysis and that topic coverage, that is, representing topics within a corpus of written material, may often be a more appropriate goal (Schober et al., 2016). The TTE approach identifies *coverage error* (pertaining to over- and under-coverage of topics), *query error* (resulting from inaccurate search queries used for data extraction), and *interpretation error* (variation between true value and interpretation) as potential threats to validity for inference from social media analyses.

Recognizing the value of the TTE framework, we identify connections between the proposed disclosure standards and insight provided for understanding coverage, query, and interpretation error. However, we also note that social media may be used to analyze research questions that are not related

to representing individuals within a population (population coverage) or topics within a corpus (topic coverage) and further social media may be used to support or supplement results from other traditional data sources. For example, online marketing efforts for emerging products like e-cigarettes and alternative tobacco products are difficult to fully monitor using traditional data sources because these products are not typically advertised widely at the point of sale or in print or broadcast media. They are typically first promoted on social media, which can provide critically important information to fully measure online marketing efforts (e.g., Huang et al., 2014).

The research standards for a given topic will depend upon the specific research question, and the three error components of the TTE framework may or may not be relevant. Thus, we emphasize that a flexible approach is needed to judge whether the standards of a specific social media analysis achieve the rigor needed for the research question, while noting that the proposed standards here encompass the needs for a broad array of research questions.

Methods

To guide rigorous analysis of social data and report findings using social science epistemology, we reviewed the literature related to data quality and methodological disclosure from biostatistics, computer science, and communications. We attempted to identify common constructs for qualitative and quantitative research methods and map these constructs to social data workflows and to the existing disclosure standards in the fields of opinion research and social sciences.

We drew upon the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) tool for the reporting of systematic review data, as a conceptual template as the data sources for reviews can be heterogeneous, very similar to the data obtained from social media, mapping the domains determining data quality in PRISMA to those needed for extraction and analysis from social media sources (Liberati et al., 2009; Page et al., 2021). We synthesized this approach with the American Association for Public Opinion Research (AAPOR) Transparency Initiative guidelines and the American Psychological Association Transparency and Openness Promotion (TOP) guidelines as a framework for social media data collection and quality assessment. Thus, the AAPOR Transparency Initiative Disclosure elements refer to the disclosure of information on data collection strategy; funding source/sponsor; measurement tools/instruments (e.g., questionnaires or coding schemes); population under study; method used to generate and recruit the sample; method(s) and mode(s) of data collection; dates of data collection; sample sizes; data weighting approach; data processing and validity checks; and acknowledgment of limitations of the design and data collection. The PRISMA reposting guidelines detail reporting recommendations pertaining to the study support sources, availability of data, code, and other materials, data collection process, and data items, among others. The TOP Guidelines cover eight general domains of research planning and reporting, including citation standards (citation for data and materials disclosures); data transparency (data sharing disclosures, such as posting to a repository); analytics methods transparency (e.g., disclosure of programming code); research materials transparency (materials sharing); design and analysis transparency (e.g., data preprocessing methods; reliability analyses); study design preregistration; analysis plan preregistration; and replication (disclosure of the publication of replication studies) (American Psychological Association, 2023). Thus, there is consensus regarding recommended transparency standards across social science domains which have to do with disclosures of research funding/sponsorship sources, data collection, processing and validation procedures, as well as analytic methods. These key concepts are also consistent with other literature detailing guidelines for evaluation of compliance with the scientific method, for example, Armstrong and Green (2022).

We synthesized and translated these practices and recommendations that are the standard for social science research to research using social media data and generative AI. While some disclosure elements were directly relevant across domains, including the social media data analyses (e.g., disclosure of the funding source), some items require translation or adaptation (e.g., description of the sample frame) or development of an analogous principle (e.g., data access point), or a novel disclosure element (e.g., amount of data decay in social media). Based on our findings, we propose a list of disclosure items as a reporting standard for social media research. We incorporate disclosure consideration regarding use of AI technologies (e.g., generative AI) and natural language processing tools. Our goal is not to direct researchers in their design choices, but to provide a framework and propose measures for evaluating the completeness of reporting and quality of data used in social media studies. Using data quality metrics, we show how selection of sampling and search filters affects the results and conclusions. We do not undertake to prescribe a short list of methods and tools to be used for social and digital media research, but rather to propose standards for how methodologies, procedures, and limitations are documented to increase transparency and replicability and allow consumers to evaluate research rigor.

Proposed Disclosure Items

Our proposed metrics for social data quality assessment and a list of minimal (or immediate) and optional (or preferred) disclosure items are detailed below and summarized in Table 1.

Table 1. Overview of Disclosure Items for Social Data Quality Reporting and Target Error or Bias Prevention.

Minimal/immediate disclosure elements	Target error/bias*	Optional/preferred disclosure elements	Target error/bias*
Funding source	T	Source code	T, R
Scope of study		Coding/labeling instructions manual	T, R
Platform	C, I	Ethical concerns/need for Institutional	T
Target population	C, I	Review Board (IRB) review (methods	
Point of data access (e.g., mode of data access and data providers)	C, Q, R, I, T	of protecting personally identifiable information of social media account users)	
Sampling approach (a description of the sample frame and its coverage of the target population or topic)	C, R	Data decay assessment (e.g., proportion of unavailable or deleted data at the time of analysis)	T, R, I, C
Number of units/data points	C, R	Spam index (e.g., method of detection,	T, R, I
Protocol and analytic tools (e.g., software used, programming language/ scripts)	R, T	proportion of robotic or "bot" accounts or messages)	
Data handling (preprocessing and cleaning procedures)	I		
Search query/filter construction (rationale for keyword or search rule selection)	I, Q, R, T		
Data quality assessment	Q, R, T		
Data Analysis	I, R, T		
Deductive: Classifier training and performance quality assessment			
Inductive: Qualitative interpretation (e.g., topic modeling)			

*Biases and errors: T = transparency; R = replicability; C = coverage error; Q = query error; I = interpretation error.

Minimal Disclosure

We propose that the following items should be included as minimal disclosure requirements in any and every report of research results, or made available immediately upon release of such a report.

Data Collection

Scope of the Study

The report should include the rationale for platform selection, description of the target population or topic, point of data access, sample frame coverage, data verification procedures, total participants, or data points (such as number of posts retrieved or number of social media accounts) on which data were collected, as outlined below. Method and dates of data collection (duration of the study, including when data were collected and for what time period) should also be disclosed. Description of the metadata used in the study, if applicable, is also critical to ensure replicability of the analyses. We propose reporting the following sub-items:

(a) *Target population/topic*: Research subject or topic should be defined with relevant specifics such as selected location, language, and user types (e.g., tobacco-related tweets posted in the English language in the United States, abortion-related X/Twitter content in Nevada)

(b) *Platform*: Platform selection is directly related to coverage error, that is, coverage of target population or topic (Table 1). The reasons for selection (justification for platform choice, given the context of the research questions; explanation for whether the platform offers appropriate depth, format, mode of content, amount, timing; and degree to which it matches the target population or topic.) should be described.

Rationale: Populations of different demographics are drawn to different platforms; thus users of one platform may be more or less representative of the population at large than another platform. Furthermore, communicative activities on a given platform may not represent the full breadth of the overall public debate because of different functionalities of platforms. In addition, social desirability and self-censorship may be more characteristic of some platforms (e.g., platforms offering less anonymity such as Facebook), compared with others (e.g., X/Twitter or Reddit). All of the above factors are related to coverage of target population or topic and thus may affect the results of the study and interpretation of findings. If social media accounts are analyzed, information on types of social media accounts (e.g., real people, verified accounts, bots, influencers) and whether certain categories are selected or removed should be described. Subgroups of platform users may behave differently on a given platform.

(c) *Data access*: Description of the methods of access and collection of the selected platform data should be provided, including the mode of data access and data providers (e.g. access to specific application programming interfaces [APIs], crawling [or scraping] strategy) as decisions made in choosing the approach to data access may result in coverage errors and query errors (Table 1).

Rationale: Different access points of data may produce data with different records. Data access also changes over time. Until early 2023, X/Twitter's streaming API provided access to 1% sample of all tweets, while PowerTrack API provided access to all public tweets, affecting coverage of target population and topic (Y. Kim, Nordgren, & Emery et al., 2020; Morstatter et al., 2013). Subsequent changes to X/Twitter restricted data access to third-party social listening service providers and scraping. Facebook data were fully available before access was restricted in 2016. Currently, CrowdTangle is the best source of Facebook and Instagram data from publicly available accounts. These different access points may produce data with different metadata, which may enhance or limit the scale of search queries (Y. Kim, Nordgren, & Emery, 2020), which applies to other platforms as well if multiple ways to access and pull data are available.

(d) *Sample frame*: A description of the sample frame and its coverage of the target population or topic for sample-based research (thus directly related to coverage error; see Table 1) should be included unless the census of social media posts/accounts matching a query is retrieved. The nature of any oversampling (e.g., of social media posts referencing top selling brands by product category) and definition of strata (e.g., stratification based on time increments or by geographic location) should be described.

Rationale: A sampling frame is carefully designed to represent a target population and derive representative estimates in survey research. While the universe/census of the target population *a priori* is not always known in social media research, researchers can describe parameters available to them that define a universe of interest and test these parameters. In other words, even if the universe of social media posts or accounts of interest is unknown, the available important parameters can be identified and used to set the sample frame. Such sampling frame should be carefully designed and executed to extract a representative data set for the target topic and/or make valid inference.

(e) *Number of units/data points*: The unit of analysis such as post, video/image, or account and number of units of analysis should be disclosed.

Rationale: The unit of analysis is closely tied to the target subject or topic, and replicability. Reporting number of analysis units enables comparability. It is worth noting that the total amount of posts, videos, or accounts related to a topic of interest may be relative (e.g., search volume on Google Trends).

Protocol and Analytic Tools. The software, programming language/scripts, any other analytic tools, and workflow for executing these tools should be described.

Rationale: There are a variety of tools available to analyze social media data, both open sources and commercial software, including emerging generative AI tools such as ChatGPT. Disclosure of computing tools is key to replicability of findings. For instance, social data are often analyzed or processed using Python, R, or other software geared to analyzing large corpuses of data among others. Same machine or statistical learning models are supported by more than one tools, and default settings for parameters and optimization may differ, resulting in different estimates. Certain software providers do not disclose module language and process of module validation. Use of generative AI tools for social media data analysis may augment the efficiency and speed of processing and analysis of large corpuses of social data, but may not be compliant with platform or provider terms of service and can have ethical implications (Elmas & Gül, 2023; Salah et al., 2023). Depending on the amount of contribution of AI systems to the analysis, description, and interpretation of findings, generative AI has been included as a co-author in the published literature, with some systems (e.g., ChatGPT) providing consent to be listed as a co-author (e.g., Haluza & Jungwirth, 2023).

Search Query Construction. The keywords selected to develop the search filter and the search rules for a more focused search should be provided. Outline your rationale for initial keyword selection (e.g., expert knowledge, resources/tools/skills used for systematic search, etc.) as well as for selecting or removing certain keywords. For example, report the relevance (precision) and frequency (number of posts retrieved) of the keywords, or the signal-to-noise (relevant to irrelevant data) ratio or the proper thresholds (by search term). Search filter construction is often an iterative process, alternating between keyword addition and removal based on relevance and frequency (Y. Kim et al., 2016). Generative AI technologies can also be used to identify terms relevant to a topic of interest, to generate search rules and convert them to regular expressions for search query construction. These tools can also translate or adapt search filters to other languages and cultural contexts to conduct multilingual analyses. Search filter is directly related to query error; a precise yet narrow search filter is likely to miss relevant content (i.e., false negative), while a comprehensive search filter is likely to contain false positive content; the balance between precision and completeness is important.

Rationale: Expressiveness of query languages and choice of keywords in combination with Boolean rules in queries define the resulting datasets. Thus, search term selection can

affect the study conclusions. For instance, using "smoking" as a search term for tobacco-related social media data collection could result in retrieval of non-relevant posts containing words like "smoking ribs," "smoking hot" (Emery et al., 2014).

Data Processing

Data Handling. Preprocessing and cleaning procedures, including de-duplication, aggregation, de-identification (if applicable), metadata (e.g., user profile, geographic location, time posted, etc.), and feature extraction, should be outlined. Use of software or tools, such as generative AI, for data preprocessing and text mining should also be disclosed.

Rationale: Converting data from a raw format to more manageable format, for instance, unpacking semi-structured data (e.g., JSON) to structured document-term matrix should be briefly described. Text mining techniques are often used in preprocessing of social media data (e.g., stop words removal, stemming, segmenting the language—factorization, speech-tagging), which can affect the subsequent procedures and analyses. In fact, data preprocessing and cleaning often influence the success of machine learning training and results, affecting interpretation error (as noted in Table 1).

Data Quality Assessment. The quality of retrieved data should be objectively assessed and quantified by inspecting a sample of data classified by search filter, for example, via cross-validation of automated coding based on a sample of data labeled by multiple human trained coders knowledgeable about the topic of interest to minimize potential error or bias, that is, the "gold standard" of filter quality assessment (Y. Kim et al., 2016). Reporting quality measures of the retrieved data, including retrieval recall (completeness of search filter; how much of the relevant data is retrieved by search filter) and retrieval precision (how much of retrieved data by search filter is relevant) helps comparability and transparency. The procedure to assess search filter quality—the selection of data sample (e.g., a subset of data based on random sampling stratified by keyword and account type may serve as a representative sample) and the evaluation strategy (e.g., agreement between coding based on human judgment vs. automated search filter selection, inspection of data that do not match search filter) must be disclosed. For example, several existing studies on the amount and content of tobacco-related tweets have included filter retrieval precision and retrieval recall assessments (e.g., Y. Kim, Nordgren, & Emery, 2020; Kostygina et al., 2016).

Thus, calculation of quality measures typically involves human judgment on a sample of data as a gold standard (Y. Kim et al., 2016). The human coding approach should be described as follows:

(a) *Sampling strategy*: If the quality assessment involves human coding of a sample dataset, a description of the sampling frame, sample size, and calculation of intercoder reliability should be reported. Results based on a sample that is too small may be less reliable, and coding a sample that is too large may be burdensome. Statistical consideration to obtain reliable results is required.

(b) *Human coding approach* and definition of each class should be described (Stryker et al., 2006). Whether human coding is assumed as the gold standard (no or negligible error and bias) is related to interpretation error. If human judgment is not considered as the gold standard for a study, the researchers should discuss how imperfect human coding may affect the search filter assessment. For example, could the filter lead to biased inferences? If biased, in which direction, and what are the consequences? Intercoder reliability and use of crowdsourcing for coding tasks should be reported as well.

Data Analysis

Analysis Methods and Measures. Detail the deductive or inductive methods used for data analysis, including statistical techniques, machine learning algorithms, or qualitative analysis (e.g., topic modeling). Explain how the data were categorized, classified, or clustered to answer to the study research questions. Specify the metrics and measures used in the analysis, such as engagement metrics, sentiment analysis scores, or content classification criteria.

(a) *Classifier training and performance quality assessment (deductive methods).* If machine learning is used for any part of data analysis, the process of building predictive models and their accuracy assessment should be described, including the process for training the classification model and its performance measures (e.g., Li et al., 2014). The classifier accuracy, precision, and recall (or *F*-score as a measure combining precision and recall; area under the curve (AUC) if logistic regression is used) should be reported. Numerous extant social media studies provide information on classifier training procedures, accuracy, precision, and recall measures (Czaplicki et al., 2020; Liu et al., 2019; K. Kim, Gibson, et al., 2020).

(b) *Qualitative analyses (inductive methods).* If topic modeling methods are used, clearly state the type of topic modeling algorithm used, whether it is latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), or another method or a generative AI tool (Chen et al., 2019). Include the hyperparameters and settings chosen for the model; provide details on the training process, such as the number of topics selected and the number of iterations. If relevant, describe how the model's performance was evaluated, such as using coherence scores or other metrics and report the results of this evaluation. If

topics were labeled, the methodology and criteria used for assigning labels to topics should be explained, and examples of topic labels should be provided. If visualizations were created, the tools or libraries used and/or parameters for creating the visualizations should be disclosed.

Researchers should disclose if generative AI tools are used for inductive or deductive analyses, for example, to create features for the classification model or to categorize social media data based on learned/ingested training data sample previously labeled by humans or a machine (e.g., to analyze social media posts to extract sentiment toward a particular topic). Since the predictive models built by generative AI are a "black box," additional methods for validation and accuracy/performance quality assessment should be described (see Supplemental Appendix 1 for an illustration of additional disclosure items that may need to be considered for studies using generative AI; the list was generated via ChatGPT 3.5 query).

Rationale: Data retrieved by comprehensive search filters are likely to include non-relevant content. To reduce the degree of the query error, we may train supervised learning classifier to further remove non-relevant data. However, since all predictive models make false positive and false negative errors, interpretation error is also likely. Reporting classifier training procedure and its performance metrics helps comparability and transparency of methods.

Funding Source. Disclose who sponsored the research study, who conducted it, and who funded it, including (to the extent known) all original funding sources.

Rationale: Disclosure of sponsor or sources of funding is the standard practice with any scientific research study (e.g., American Association for Public Opinion Research, 2021). This is a fundamental requirement as funder involvement in research question, study design, data analysis, and interpretation of results may bias study findings.

Optional (Preferred) Disclosure Items

Depending on the design and objective of the research study, additional information that can be disclosed to enhance transparency and reproducibility of social media research and minimize error includes as follows:

1. *Source code or scripts used*. Providing source code or scripts used to analyze social data enables reproducibility of the study findings.
2. *Coding or labeling instructions manual* (beyond simple definition) can help avoid potential interpretation error.
3. *Strategies to address ethical concerns* (if any). Researchers can outline measures taken to ensure the responsible use of social media data (e.g., Hunter et al., 2018; Taylor & Pagliari, 2018).
4. *Data decay assessment* (proportion of data that are unavailable, deleted by the platform or user, or made private at the time of analysis) can be provided to minimize coverage error.
5. *Spam index.* Researchers can describe their approach for defining or detecting spam content and report the proportion of robotic or "bot" accounts or messages retrieved.

Additional items discussed in the literature that are not shown in the above list of recommended disclosure elements—due to technical and possible contractual constraints—include disclosure of the raw data; procedure for acquiring consent to participate in the research study from social network users (e.g., whether consent was secured by the user checking a checkbox at the time of creating a social media profile vs. consent being obtained specifically for the research project); as well as procedures for participant debriefing upon study completion.

Discussion

Our approach aims to consolidate and map the concerns about lack of transparency, reporting, and documentation standards raised in the literature on social data analysis quality and replicability and take the process a step further to propose a list of specific disclosure elements grounded in social science epistemology. In fact, striking parallels exist between the current state of social data research and early public opinion research. For example, election polling in the early 1900s often relied on information provided by bookies (i.e., betting markets) or "man-on-the-street" interviews (Rhode & Strumpf, 2004). A classic example of poor results in early public opinion polling can be found in the 1936 prediction by *The Literary Digest* that Alfred Landon would be the next US president. Despite the *Digest's* correctly predicting several previous elections, Landon's landslide defeat in 1936 went against its prediction. This event is often cited as inspiring the onset of methodological reflection and development of a rigorous science of public opinion polling, which has yielded a widely accepted system of survey research reporting standards that ensure transparency of methods and replicability of findings. In the context of the current communication and research ecosystem, which includes vast amounts of data from digital sources, including social media, and the near-real time ability to analyze these data, the underlying need for disclosure and transparency is just as urgent as it was in the early years of public opinion research.

Thus, we proposed that the minimal disclosure standards should include description of funding source, platform, target population, point of data access, sampling strategy (if sampling is used), data verification procedures, protocol and workflow for executing software and analytic tools, data handling, search filter construction and assessment procedures, classifier training, and performance quality assessment, as detailed above. We believe this proposed framework

presents a viable and effective method for quality evaluation of social data research. These criteria go beyond the identification of potential limitations and biases related to the use of social data and generative AI in research, to offer documentation guidelines for auditing and mitigating these issues to ensure the maximum validity and replicability of findings.

While there are overlapping threats to validity and similarities in reporting requirements for empirical or survey research and social media data research, important distinctions exist, which warrants discussion and motivates the framework we proposed. For example, surveys are grounded in a statistical framework that accounts for inferential error (i.e., sampling error, coverage error, etc.), measurement error, assumptions that there are objective measures of the population itself, and that the survey items are knowable and measurable. With social media data, however, such assumptions do not hold because the tools used to measure the population and the "items" are generated by the group that is creating the population and messages; that is, the posts themselves comprise the population and items being measured, so there is no objective "ground truth" to compare with. In such a scenario, rather than throw up our hands in defeat, we are recommending an approach that entails extreme methodological transparency. While others have proposed quality standards for social media data (Hsieh & Murphy, 2017), we contend that these are an important first step, but insufficient because this approach does not address many of the decisions made in the data collection, preprocessing, and analysis, all of which can affect the study conclusions. Thus, disclosure standards for social media data research must be expansive and adaptive to change, as the platforms themselves change access policies rapidly and the public shifts their loyalty and attention as new social media platforms emerge.

Other scholars have cautioned against "too much transparency" in today's machine learning and statistical research due to intellectual property concerns, the fact that algorithmic logic may not be fully reflected in the source code, as well as the potential risk of backfiring and increasing distrust among members of the public whose research outcome expectations are violated (Hosanagar & Jair, 2018). These scholars have called for "explainable artificial intelligence (AI)" as a more palatable solution. Explainable AI approach does not open the "black box" of decision-making algorithms or machine learning-based analytics, but provides an explanation of the inputs that result in the greatest impact on the final decisions or outcomes of algorithm-based analyses. However, emerging AI tool transparency issues call this argument into question (Dwivedi et al., 2023). Explainable AI may lack efficiency as an approach of science communication if the goal is to establish replicability of social data research in the field of opinion research.

Our goal is not to direct researchers in their design choices, but to provide a framework and propose measures for evaluating the completeness of reporting and quality of data used in social media studies. We aim to translate and synthesize practices that are the standard for both computational research and conventional social science research, in an attempt to breach existing "silos" and make each domain more salient to the other. This translation can serve as a resource for manuscript and grant reviewers, journal editors, and funding organizations that enlist technical or subject matter experts to review studies that use social media data and/or AI to address social science or public health research questions. The proposed standards could be relevant to a range of studies that rely on data mining, natural language processing, and machine learning techniques to extract insights from the vast amount of textual and visual information available on social media, for example, from public opinion and sentiment analysis (analyzing the discourse and sentiment of social media posts to understand trends in public opinion and social norms); to social network analysis (examining the structure and dynamics of social networks to identify influencers, communities, and connections); and to language and linguistics research (studying language evolution, slang, and dialects through social media conversation) among others (e.g., Gallagher et al., 2021; Kozinets, 2020; Yadav & Vishwakarma, 2020). Detailed disclosure of parameters enables study quality evaluation, replication, and advancement across various domains of inquiry and methodologies. Our proposed standards apply whether the study aims to be generalizable to a broad population or focuses on a narrower community or topic, like a case study or netnographic research.

We do not presume that our proposed framework is the final word. Rather we propose the framework as a starting point, and urge the community of researchers and institutions that are involved in decisions about funding, conducting and disseminating social media research to open a larger dialogue. The goal of such a dialogue would be broad consensus and ongoing maintenance of a disclosure framework for social data research as a "moving target" in the evolving environment of rapidly changing media and technology use and access by organic, commercial, and academic users. Such a framework would enable funders, journal editors, research consumers, and those making decisions based upon social media research studies to evaluate the validity of a study, compare studies with conflicting results, and make decisions based on known parameters.

Disclosure

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Awards Nos. R01CA248871 and R01CA234082 and the National Institute on Drug Abuse of the National Institutes of Health under Award No. R01DA051000.

ORCID iDs

Ganna Kostygina https://orcid.org/0000-0002-8416-6168
Zachary Seeskin https://orcid.org/0000-0002-8005-4521

Supplemental Material

Supplemental material for this article is available online.

References

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, S., Benjamin, D., . . .Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, *4*, 4–6. https://doi.org/10.1038/s41562-019-0772-6

Albergotti, R. (2014, July 2). Facebook experiments had few limits: Data science lab conducted tests on users with little oversight. *The Wall Street Journal*. http://online.wsj.com/articles/facebook-experiments-had-few-limits-1404344378

American Association for Public Opinion Research. (2021). *AAPOR code of professional ethics and practices*. https://www-archive.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx

American Psychological Association. (2023). *Transparency and openness promotion guidelines: What are they?* https://www.apa.org/pubs/journals/resources/publishing-tips/transparency-openness-promotion-guidelines

Armstrong, J. S., & Green, K. C. (2022). *The scientific method: A guide to finding useful knowledge*. Cambridge University Press.

Assenmacher, D., Weber, D., Preuss, M., Calero Valdez, A., Bradshaw, A., Ross, B., Cresci, S., Trautmann, H., Neumann, F., & Grimme, C. (2022). Benchmarking crisis in social media analytics: A solution for the data-sharing problem. *Social Science Computer Review*, *40*(6), 1496–1522. https://doi.org/10.1177/08944393211012268

Bagby, R. M., Goldbloom, D. S., & Schulte, F. M. (2006). The use of standardized rating scales in clinical practice. In D. S. Goldbloom (Ed.), *Psychiatric clinical skills* (pp. 11–17). Mosby. https://doi.org/10.1016/B978-0-323-03123-3.50007-7

Booth, R. (2014, June 29). Facebook reveals news feed experiment to control emotions. *The Guardian*. https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Bruns, A. (2013). Faster than the speed of print: Reconciling "big data" social media analysis and academic scholarship. *First Monday*, *18*(10). http://journals.uic.edu/ojs/index.php/fm/article/view/4879/3756

Bruns, A., & Stieglitz, S. (2014). Twitter data: What do they represent? *Information Technology*, *56*(5), 240–245.

Center for Democracy & Technology. (n.d.). *Digital decisions*. https://cdt.org/issue/privacy-data/digital-decisions/

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, *163*, 1–13. https://doi.org/10.1016/j.knosys.2018.08.011

Cockburn, A., Dragicevic, P., Besançon, L., & Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Communications of the ACM*, *63*(8), 70–79. https://doi.org/10.1145/3360311

Cohen, R., & Ruths, D. (2013). Classifying political orientation on Twitter: It's not easy! *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 91–99. https://doi.org/10.1609/icwsm.v7i1.14434

Cook, T. D., & Campbell, D. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin.

Council for Big Data, Ethics, and Society. (n.d.). https://bdes.datasociety.net/

Czaplicki, L., Kostygina, G., Kim, Y., Perks, S. N., Szczypka, G., Emery, S. L., Vallone, D., & Hair, E. C. (2020). Characterising JUUL-related posts on Instagram. *Tobacco Control*, *29*(6), 612–617. https://doi.org/10.1136/tobaccocontrol-2018-054824

Desmond, H. (2022). Status distrust of scientific experts. *Social Epistemology*, *36*(5), 586–600. https://doi.org/10.1080/02691728.2022.2104758

Diakopoulos, N. (2016). Accountability in algorithmic decision making: A view from computational journalism. *Communications of the ACM*, *59*(9), 56–62.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . .Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Elmas, T., & Gül, İ. (2023). *Opinion mining from YouTube captions using ChatGPT: A case study of street interviews polling the 2023 Turkish elections* (arXiv:2304.03434). https://doi.org/10.48550/arXiv.2304.03434

Emery, S., Szczypka, G. A., Kim, Y., & Vera, L. (2014). Are you scared yet? Evaluating fear appeal messages in tweets about the Tips Campaign. *Journal of Communication*, *64*(2), 278–295.

Fairness, Accountability, and Transparency in Machine Learning. (n.d.). https://www.fatml.org/

Fineberg, H., Stodden, V., & Meng, X.-L. (2020). Highlights of the U.S. National Academies report on "reproducibility and replicability in science." *Harvard Data Science Review*, *2*(4), 1–6. https://doi.org/10.1162/99608f92.cb310198

Gallagher, R. J., Doroshenko, L., Shugars, S., Lazer, D., & Foucault Welles, B. (2021). Sustained online amplification of COVID-19 elites in the United States. *Social Media + Society*, *7*(2). https://doi.org/10.1177/20563051211024957

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, *38*(1), 16–27.

Goroff, D. L. (2015). Balancing privacy versus accuracy in research protocols. *Science*, *347*, 479–480.

Graham, M., Hale, S. A., & Gaffney, D. (2013). Where in the world are you? Geolocation and language identification in Twitter. In *Proceedings of ICWSM* (pp. 518–521). https://arxiv.org/ftp/arxiv/papers/1308/1308.0683.pdf

Hall Jamieson, K. (2015). Communicating the value and values of science. *Issues in Science and Technology*, *32*, 72–79.

Haluza, D., & Jungwirth, D. (2023). Artificial intelligence and ten societal megatrends: An exploratory study using GPT-3. *Systems*, *11*(3), 120.

Hosanagar, K., & Jair, V. (2018, July 23). We need transparency in algorithms, but too much can backfire. *Harvard Business Review*. https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire

Hsieh, Y. P., & Murphy, J. (2017). Total Twitter error: Decomposing Public Opinion Measurement on Twitter from a Total Survey Error Perspective. In P. P. Biemer, E. D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker, & B. T. West (Eds.), *Total survey error in practice* (pp. 23-46). John Wiley.

Huang, J., Szczypka, G., Kornfield, R., & Emery, S. L. (2014). A cross-sectional examination of marketing of electronic cigarettes on Twitter. *Tobacco Control*, *23*(3), iii26–iii30.

Hunter, R. F., Gough, A., O'Kane, N., McKeown, G., Fitzpatrick, A., Walker, T., McKinley, M., Lee, M., & Kee, F. (2018). Ethical issues in social media research for public health. *American Journal of Public Health*, *108*(3), 343–348.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, *79*(4), 839–880.

Jurgens, D., Finnethy, T., McCorriston, J., Xu, Y. T., & Ruths, D. (2015). Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. In *ICWSM*. AAAI. https://ojs.aaai.org/index.php/ICWSM/article/view/14627/14476

Kabat, G. C. (2017). Taking distrust of science seriously. *EMBO Reports*, *18*, 1052–1055. https://doi.org/10.15252/embr.201744294

Kennedy, B., Tyson, A., & Funk, C. (2022, February 15). Americans' trust in scientists, other groups declines. *Pew Research Center*. https://www.pewresearch.org/science/2022/02/15/americans-trust-in-scientists-other-groups-declines/

Kim, K., Gibson, L. A., Williams, S., Kim, Y., Binns, S., Emery, S. L., & Hornik, R. C. (2020). Valence of media coverage about electronic cigarettes and other tobacco products from 2014 to 2017: Evidence from automated content analysis. *Nicotine & Tobacco Research*, *22*(10), 1891–1900. https://doi.org/10.1093/ntr/ntaa090

Kim, Y., Huang, J., & Emery, S. (2016). Garbage in, garbage out: Data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of Medical Internet Research*, *18*(2), e41. https://doi.org/10.2196/jmir.4738

Kim, Y., Nordgren, R., & Emery, S. (2020). The story of Goldilocks and three Twitter's APIs: A pilot study on Twitter data sources and disclosure. *International Journal of Environmental Research and Public Health*, *17*, 864. https://doi.org/10.3390/ijerph17030864

Kostygina, G., Tran, H., Shi, Y., Kim, Y., & Emery, E. (2016). "Sweeter than a Swisher": Amount and themes of little Cigar and Cigarillo content on Twitter. *Tobacco Control*, *25*(Suppl. 1), i75–i82. https://doi.org/10.1136/tobaccocontrol-2016-053094

Kozinets, R. V. (2020). *Netnography: The essential guide to qualitative social media research*. (3rd ed.). SAGE Publications.

Lazer, D., & Kennedy, R. (2015, October 1). What we can learn from the epic failure of Google flu trends. *Wired*. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Li, H., Mukherjee, A., Liu, B., Kornfield, R., & Emery, S. (2014). Detecting campaign promoters on Twitter using Markov random fields. In *2014 IEEE international conference on data mining* (pp. 290–299). https://dl.acm.org/doi/10.1109/ICDM.2014.59

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLOS Medicine*, *6*(7), Article e1000100.

Liu, J., Siegel, L., Gibson, L. A., Kim, Y., Binns, S., Emery, S., & Hornik, R. C. (2019). Toward an aggregate, implicit, and dynamic model of norm formation: Capturing large-scale media representations of dynamic descriptive norms through automated and crowdsourced content analysis. *Journal of Communication*, *69*(6), 563–588. https://doi.org/10.1093/joc/jqz033

Mehrabi, N., & Pashaei, E. (2021). Application of horse herd optimization algorithm for medical problems. In *2021 international conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 1–6). https://doi.org/10.1109/INISTA52262.2021.9548366

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. (2013). Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *Proceedings of the 7th international AAAI conference on weblogs and social media*. https://arxiv.org/abs/1306.5204

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), Article 89. https://doi.org/10.1186/s13643-021-01626-4

Reed, L., & boyd, d. (2016). *Who controls the public sphere in an era of algorithms? Questions and assumptions*. Data & Society. https://www.datasociety.net/pubs/ap/QuestionsAssumptions_background-primer_2016.pdf

Rhode, P. W., & Strumpf, K. S. (2004). Historical presidential betting markets. *Journal of Economic Perspectives*, *18*(2), 127–141.

Rosenberg, M., Confessore, N., & Cadwalladr, C. (2018, March 17). How Trump consultants exploited the Facebook data of millions. *The New York Times*. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html?action=click&module=Intentional&pgtype=Article

Salah, M., Al Halbusi, H., & Abdelfattah, F. (2023). May the force of text data analysis be with you: Unleashing the power of generative AI for social psychology research. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100006. https://doi.org/10.1016/j.chbah.2023.100006

Schillinger, D., Chittamuru, D., & Ramírez, A. S. (2020). From "infodemics" to health promotion: A novel framework for the role of social media in public health. *American Journal of Public Health*, *110*(9), 1393–1396. https://doi.org/10.2105/AJPH.2020.305746

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social media analyses for social measurement. *Public Opinion Quarterly*, *80*(1), 180–211.

Stryker, J., Wray, R., Hornik, R., & Yanovitzky, I. (2006). Validation of database search terms for content analysis: The case of cancer news coverage. *Journalism & Mass Communication Quarterly*, *83*, 413–430.

Taylor, J., & Pagliari, C. (2018). Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics*, *14*(2), 1–39.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the 8th international conference on weblogs and social media, ICWSM 2014* (pp. 505–514). The AAAI Press. https://arxiv.org/abs/1403.7400

Vincent, J. (2023, January 17). Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content. *The Verge*. https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit

Voytek, B. (2017). Social media, open science, and data science are inextricably linked. *Neuron*, *96*, 1219–1222.

Winter, T., Riordan, B. C., Scarf, D., & Jose, P. E. (2022). Conspiracy beliefs and distrust of science predicts reluctance of vaccine uptake of politically right-wing citizens. *Vaccine*, *40*(12), 1896–1903. https://doi.org/10.1016/j.vaccine.2022.01.039

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review*, *53*, 4335–4385. https://doi.org/10.1007/s10462-019-09794-5

Yom-Tov, E. (2016). *Crowdsourced health: How what you do on the internet will improve medicine*. MIT Press.

Author Biographies

Ganna Kostygina, PhD, is a Principal Research Scientist at the Social Data Collaboratory, NORC at the University of Chicago. Her research agenda centers on advancing communication science and technology for tobacco control and health promotion. Specifically, she has conducted research on topics related to tobacco and substance use prevention, as well as tobacco and alcohol product marketing and counter-marketing on traditional and digital media channels.

Yoonsang Kim, PhD, is a Principal Data Scientist with NORC at the University of Chicago. She oversees study design, statistical analysis, machine learning, data harmonization, and other data science practices. At NORC's Social Data Collaboratory, she serves as a lead biostatistician for social media research to examine the effects of exposure to social media on health behaviors, focusing on data quality assessment and the development of social media. Her primary research interests are substance use, social environment, and marketing of health-related products.

Zachary Seeskin, PhD, is a Senior Statistician with NORC at the University of Chicago, where he works on sample design, estimation, and data analysis for government and public interest surveys. Seeskin contributes to weighting, total survey error analysis, small area estimation, imputation, and adaptive design for such surveys as the National Immunization Survey and the General Social Survey. In addition, his expertise includes analyzing administrative data quality and combining data sources for evidence-building. He further serves as an adjunct faculty member with Northwestern University's School of Professional Studies, teaching in the Public Policy and Administration program.

Felicia LeClere, PhD, is a Distinguished Senior Fellow in NORC's Health Sciences Department. She is currently Project Director for the Medicare Current Beneficiary Survey and the Healthcare Cost and Utilization Project sponsored by the Center for Medicare and Medicaid Services and the Agency for Healthcare Research and Quality, respectively. Her primary interests are in data dissemination and the support of scientific research through the development of infrastructure. She has also conducted research on the health of minorities and immigrants as well as health disparities. Her work has been sponsored by the National Institute of Child Health & Human Development and the National Institute on Drug Abuse.

Sherry Emery, PhD, is a Senior Fellow at the Social Data Collaboratory, NORC at the University of Chicago. Her interdisciplinary research applies the approaches of health communication, economics, and public policy to understand how both traditional and new media influence health behavior. For over two decades, she focused primarily on the roles that tobacco control and other tobacco-related advertising play in shaping attitudes, beliefs, and tobacco use behaviors among youth and adults.