Suvivuo, Sampsa; Tuunainen, Virpi

Challenges and Solutions in Qualitative Big Data Research: A Methodological Literature Review

Published in:
Communications of the Association for Information Systems

DOI:
10.17705/1CAIS.05502

Published: 31/07/2024

Document Version
Publisher's PDF, also known as Version of record

Communications of the Association for Information Systems

Volume 55 Paper 2

7-28-2024

Challenges and Solutions in Qualitative Big Data Research: A Methodological Literature Review

Sampsa Suvivuo
Aalto University School of Business

Virpi K. Tuunainen
Aalto University School of Business

Recommended Citation

Suvivuo, S., & Tuunainen, V. K. (2024). Challenges and Solutions in Qualitative Big Data Research: A Methodological Literature Review. Communications of the Association for Information Systems, 55, 37-76. https://doi.org/10.17705/1CAIS.05502

Challenges and Solutions in Qualitative Big Data Research: A Methodological Literature Review

Cover Page Footnote

This manuscript underwent peer review. It was received 07/06/2022 and was with the authors for 36 months for three revisions. Mathieu Templier served as Associate Editor.

Communications of the Association for Information Systems

Research Article DOI: 10.17705/1CAIS.05502 ISSN: 1529-3181

Challenges and Solutions in Qualitative Big Data Research: A Methodological Literature Review

Sampsa Suvivuo
Department of Information and Service Management
Aalto University School of Business
0000-0002-2146-3923

Virpi Kristiina Tuunainen
Department of Information and Service Management
Aalto University School of Business
0000-0002-5758-6925

Abstract:

The digitalization of our daily lives has considerably increased the amount of digital (trace) data on people's behaviors that are available to researchers. However, qualitative methods that require manually perusing each document struggle with the width and breadth of such data. Although quantitative and qualitative big data share many challenges, we identified the practical challenges encountered by researchers, specifically with qualitative big data, and how these challenges were addressed. We reviewed 169 studies that used qualitative big data and identified three main categories of intertwined challenges: locating relevant data, addressing noise in the data, and preserving data richness. We found that the greater the amount of data and the richer they are, the greater the variety of types and sources of noise. While the volume of the data necessitates the use of algorithms, doing so entails the treatment of data in ways that decrease the richness of qualitative data. Furthermore, simultaneously ensuring high richness and veracity might be difficult because the algorithms are probabilistic, thus compelling researchers to balance the desired levels of volume, variety, and veracity. Although the identified solutions cannot completely solve this tripartite balancing, they can still be used to alleviate different aspects of such a challenge.

Keywords: Qualitative Big Data, Big Data Research, Challenges, Methodological Literature Review.

This manuscript underwent peer review. It was received 07/06/2022 and was with the authors for 36 months for three revisions. Mathieu Templier served as Associate Editor.

1 Introduction

The amount of digital data generated globally doubles every two years (IDC, 2014, p. 267). According to estimates, about 80–90 percent (Harbert, 2021; T. King, 2019) of company data are qualitative or unstructured and have several formats, including text (emails, webpages, social media, blogs, and documents), video, audio, and image formats. A company's internal and customer-facing systems could produce petabytes' worth of data in a matter of hours. From a researcher's perspective, this has led to a situation wherein more data are available than can be feasibly accommodated by qualitative methods that require humans to peruse each record (Berente & Seidel, 2014; Lindberg, 2020; Walsh et al., 2015). Furthermore, sampling "big data" is only a partly satisfactory way to reduce the amount of data and make it more manageable. Often, it is not clear beforehand which part of a dataset contains the most interesting data, and a small sample size may cause researchers to miss temporal shifts, relative volumes, or dimensions in the data. Therefore, the limits set by laborious manual coding must be overcome to better study the social patterns or collective expressions of a phenomenon (Karamshuk et al., 2017). To address the abovementioned problem, big data research has introduced computationally intensive analysis techniques, such as machine learning (ML) and neural networks (NNs), to study unprecedently large and heterogeneous datasets (Grover et al., 2020).

As with any other line of inquiry, research using big data has limitations, many of which are shared by quantitative and qualitative data. Given that the issues common to quantitative and qualitative big data have been extensively described elsewhere (see, e.g., Boyd & Crawford, 2012; Lazer et al., 2014; Mills, 2018), the current review focuses on the following research question:

> **What are the main issues researchers have encountered with qualitative big data, and how have these issues been addressed?**

Accordingly, this paper serves as a methodological literature review (Aguinis et al., 2023) that describes and synthesizes the challenges encountered by researchers while investigating qualitative big data and proposes areas for improvement.

Thus far, the majority of information systems (IS) studies have been built on what can be described as small data. In 2018, 16 percent of articles published in top IS journals were classified as big data research. However, more recently, scholars have begun to build on big data that can uniquely advance the development of theories by revealing anomalies, alternative conceptualizations of constructs, and new field experiments (Grover et al., 2020). While interviews and surveys are artificial situations, online discussion forums, and other digital sources contain records with candor, updated sources, and constantly evolving topics (McKenna, 2019). Furthermore, large-scale interviews and survey studies can be slow and prohibitively costly to conduct. In comparison, researchers using a big data corpus can collect a diverse and encompassing dataset much more efficiently, provided, of course, that the data exists. For example, studies on asthma risk factors have traditionally involved only one or two triggers of the disease. However, using big data obtained by repurposing and integrating multiple public data sources, Zhang and Ram (2020) simultaneously compared 270 risk factors while determining their relative importance, thus capturing rarely studied environmental factors. Nevertheless, many researchers (see, e.g., Dalton & Thatcher, 2014; Davidson et al., 2019; George et al., 2014; Kitchin, 2014) agree that big and small data are complementary and not mutually exclusive.

The remainder of the paper proceeds as follows. In the ongoing section, we provide definitions of big and small data, as well as our definition of qualitative big data. Then, we briefly describe the opportunities and issues common to qualitative and quantitative big data. In Section 2, we present our methodological literature review, the selection of articles, and their coding. Section 3 categorizes and examines the identified challenges and their solutions, while Section 4 contains a discussion of the intertwined nature of these challenges, along with recommendations for practices to alleviate them in such a way that addressing one challenge does not exacerbate or create other challenges. Section 5 concludes the article with the study's implications and limitations.

1.1 Big and Small Data

The nature of big data and what constitutes them are still being discussed (Jones, 2019), although they have often been described with the four Vs: volume, velocity, variety, and veracity (Abbasi et al., 2016; Goes, 2014). Volume describes the enormous quantity of data provided by the disciplines' standards.

Velocity indicates that a dataset is not static but collected in real time or updated regularly. Big data are often collected from multiple sources, or the dataset contains structured and unstructured data simultaneously, thus introducing variety and richness. Finally, big data are often “noisy,” which means they require preparation before analysis, or there could be other forms of uncertainty in the data (e.g., how the data were collected or stored), thus affecting their veracity. However, other definitions also exist. For instance, the original Vs contained only volume, velocity, and veracity (García & Álvarez-Fernández, 2022; Kitchin & McArdle, 2016), while others added value (value can be extracted from the data), visualization (the ability and need to visualize information in a clear and quick manner), and variability (the meaning of data varies depending on context or time) to make up the seven Vs of big data (García & Álvarez-Fernández, 2022). Kitchin and McArdle (2016) observed that for 26 big data datasets that they examined, there was not a single definition (even the original three Vs) that would have perfectly described every dataset, thereby implying that datasets containing big data can be widely diverse.

Big data are often user generated rather than explicitly created, collected, and stored for research purposes. In comparison, small data are defined as data that are deliberately produced by the researcher’s actions and are thus within the capabilities of traditional research methods, while simultaneously being constrained in size, temporality, and flexibility in their generation (Grover et al., 2020; Kitchin & McArdle, 2016). Most often, even if not always, big data consists of digital trace data, which are records of activities in ISs. Whether digital or analog, trace data are byproducts of actions; they are not specifically produced for research purposes but are “found” instead. Thus, trace data are considered longitudinal because they are event-based, and events occur over time. These properties make them different, for example, from surveys or interview data (Howison et al., 2011; Lazer & Radford, 2017). Moreover, trace data are created by the user(s) and typically have either a semistructured or unstructured format (Østerlund et al., 2020). For example, a post on Facebook’s timeline contains a timestamp, a user ID, and possible location data, together with unstructured text and/or image/video.

In this paper, we focus on qualitative big data, which we define as a dataset of unstructured digital data exhibiting a combination of the four V characteristics (volume, velocity, variety, and/or veracity) of big data.

1.2 Opportunities and Challenges of Big Data Research

Earlier research has identified several opportunities and challenges regarding big data, many of which apply to both quantitative and qualitative research. Today, the distinction between offline and online representations of human behavior is disappearing (Goes, 2014) and more and more human behaviors are registered as digital signals through social media, mobile commerce, cloud services, and the Internet of Things (Grover et al., 2020; Mills, 2018). Such a phenomenon has given researchers access to data that would have been extremely laborious, expensive, or outright impossible to collect two decades ago. The volume and variety of big data enable a more accurate representation of behaviors. Given that data on a wider variety of variables can now be easily collected and operationalized, this allows for addressing novel questions regarding digital phenomena or revisiting older questions with new ways and greater granularity of inquiry (Chang et al., 2014; Goes, 2014; Grover et al., 2020). Volume and variety also encourage cross-tradition and interdisciplinary research projects that maximize the multidimensionality of data (see, e.g., Abbasi et al., 2016; Chang et al., 2014; Goes, 2014). Furthermore, the ability to collect highly granular social, cultural, economic, political, and historical data at scale helps computational social scientists investigate a broader range of phenomena, leading to what has been referred to as a “paradigm shift” in the social sciences (Chang et al., 2014; Kitchin, 2014).

Various ways of combining fundamentally quantitative big data (machine pattern recognition) and qualitative small data studies (human pattern detection) are still being developed, but combining them can be beneficial for theory development (Grover et al., 2020; Lindberg, 2020; Østerlund et al., 2020). For instance, qualitative researchers may complement big data studies by generating and refining theories that help explain data and what should be considered data regarding the phenomenon (Mills, 2018). Grover et al. (2020) and Lindberg (2020) noted that algorithms can be used to uncover novel patterns in big data in ways that can offer initial structural frames for deeper theory building by studying small samples and combining machine and human pattern recognition. Hence, abduction—the process of iterating between discovery and justification—is needed to address the “why” behind identified patterns, and this can be done by coming up with reasonable inferences for the causes behind such patterns and using those inferences as a starting point for proposing new hypotheses or small data studies (Grover et al., 2020; Lindberg, 2020). Small data studies can also be scaled up with the help of big data, such as by applying text mining to big data to corroborate the results of a qualitative study with a larger dataset (Davidson et al., 2019; Karamshuk et al., 2017; Lindberg, 2020).

For quantitative research, studying big data has created the need to address the deflated p-value issue. In particular, many statistical methods and techniques have been developed for smaller samples, while for very large samples, “the immense volume of data means that almost everything is significant” (George et al., 2014, p. 323), suggesting that with big data, the p-value alone is not sufficient to determine whether results are significant (Lin et al., 2013). With a sufficiently large dataset, almost all relationships become statistically significant, causing spurious correlations among them (Abbasi et al., 2016; George et al., 2014; Kobayashi et al., 2018; Lin et al., 2013). For qualitative researchers, one consequence of the high volume of data is that they exceed a research team’s capacity to read and digest all the qualitative data, thus affecting either the breadth or depth of the analysis (Davidson et al., 2019). Furthermore, videos are becoming increasingly cheap and easy to produce and distribute. While they are a rich source of information, such richness can overwhelm a researcher with possibilities and interesting details, thus leading to difficulties in determining “what is important and worthy of analysis” (LeBaron et al., 2018, p. 247). Combining qualitative and quantitative approaches has also been proposed to alleviate the effects of high data volume (Davidson et al., 2019; Kobayashi et al., 2018).

There is a misconception—sometimes referred to as “big data hubris” (Lazer et al., 2014)—that big data are automatically better than small data, regardless of the research question. However, the representativeness of big data can be questioned by considering whether they truly represent the population of interest in general or just individuals with access to the Internet (Boyd & Crawford, 2012; Kobayashi et al., 2018; Lazer & Radford, 2017; Mills, 2018). Despite their size, the datasets are considered local because they describe only a phenomenon observed in a particular platform or community; thus, they are not global, and sampling could be suboptimal. Big data might also become a convenience sample collected not because they are the best or the only viable way to study the phenomenon, but because they are easier, faster, and cheaper to collect, especially if variables are selected due to their traceability rather than based on the understanding of the phenomenon being investigated (Agarwal & Dhar, 2014; Lazer & Radford, 2017; Lindberg, 2020; Walsh et al., 2015). The black-box nature of the application programming interfaces (APIs) used in scraping data from online sources also threatens the replicability of studies and the reproduction of datasets (Felt, 2016; Jones, 2019; Lazer et al., 2014).

Furthermore, big data research tends to focus on the “tactical” issues at the expense of answering the “whys” and settling for correlations (Grover et al., 2020). However, observing a correlation does not explain or provide a deeper understanding of why such correlations exist (Hirschheim, 2021). In general, owing to their quantitative origin, big data are better suited to providing the “what”, “where”, and “when”, but not the “why” or “how” (Abbasi et al., 2016; Dalton & Thatcher, 2014). Similarly, Smith (2020) reported that data mining, which is closely associated with big data research, reverses the scientific method by putting data before theory. Furthermore, Smith (2020) views humans as a clever species that can produce plausible explanations for any kind of pattern. The issue can be exacerbated further with the careless application of ML to qualitative data, thereby turning analysis into a mechanistic black-box exercise, wherein data are simply inputted and the algorithm produces some kind of output. Hence, theory and domain knowledge should accompany the use of ML to ensure that the study results are meaningful (Hannigan et al., 2019; Janasik et al., 2009; Kobayashi et al., 2018; Schmiedel et al., 2019).

Likewise, the use of algorithms requires data preparation to transform qualitative data into a better machine-readable form. This process creates a trade-off between standardizing the content for the algorithm and linking it to the theoretical artifact (Hannigan et al., 2019). For this reason, data preparation should not be formulaic, wherein a step is taken because it is a “common” or “standard” practice (Hickman et al., 2022). Grover et al. (2020) also raised the issues of fishing for interesting relationships (r-hacking) and creating hypotheses after the results are known (HARKing), both of which threaten the generalizability and value of big data research.

Accessibility can also be an issue. For instance, the majority of big data are created, collected, and owned by corporations, leading to a so-called “big data divide”, which indicates that there are different possibilities of accessing big data among researchers with varying financial and technical resources (Dalton et al., 2016; Grover et al., 2020; Thatcher, 2014). The research infrastructure or apparatus required to create, collect, store, and analyze big datasets is also influenced by sociotechnical aspects, such as regular database purges, multiple individuals simultaneously accessing and working on the same dataset in parallel, changes in the ways the system is used over time, and the active interpretations made by researchers. This means that big data might not be as objective and exhaustive as initially thought (as per “big data hubris” (Crawford et al., 2014; Howison et al., 2011; Janasik et al., 2009; Mills, 2018; Østerlund et al., 2020)). Furthermore, using data or a dataset collected by someone else might lead to a blurring of the context and circumstances in which the data were generated (Boyd & Crawford, 2012; Chang et al., 2014; Jones, 2019; Kitchin & McArdle, 2016; Mills, 2018; Thatcher, 2014; Walsh et al., 2015). Indeed, combining datasets is difficult because of the nonuniform ways in which datasets are collected and presented (Jones, 2019; Kelling et al., 2009).

Researchers often tend to work under the ideal user assumption, expecting all users to operate in good faith and not try to manipulate the system or engage in opportunistic behaviors (Lazer et al., 2014; Lazer & Radford, 2017). For example, individuals might not be who they purport to be (some may not even be humans at all), or they might create alternative accounts to fabricate support for themselves. In addition, several ethical issues in big data research, including acquiring the informed consent of users in a sample, minimizing potential harm to users (e.g., not generating additional attention to users who have shared or have been targets of adversarial content), and ensuring their privacy and anonymity (Chang et al., 2014; Howison et al., 2011; Kobayashi et al., 2018; Webb et al., 2017), must also be addressed. However, as demonstrated by Daries et al. (2014) in their study on massive open online courses (MOOCs), anonymization may also cause distortions in the data, thereby limiting the possibilities of replicating or extending a study. While their study originally found that 5 percent of students earned certificates in the studied MOOCs, categorizing data and suppressing values that might compromise anonymity cut the number of certificate-earning students in half while proportions of other student groups remained the same, thus altering the dataset (Daries et al., 2014).

All of the abovementioned issues are not necessarily specific to big data studies only. However, these issues are more pronounced due to the size of the datasets used in big data research, and their consequences are potentially more severe than with traditional small data studies (Mills, 2018).

2 Literature Review Methodology

We conducted a methodological literature review (Aguinis et al., 2023) of the extant literature to identify 1) the practical challenges researchers have faced in using qualitative big data and 2) the ways in which these challenges have been addressed. Methodological literature reviews examine the extant literature regarding methodological issues, summarize the relevant studies, and provide recommendations for improved practice (Aguinis et al., 2023).

2.1 Search and Selection of Articles

Big data, be they qualitative or quantitative, have seen vast interest from the research community since 2010. The search term “big data” produced 9,552 results (both journal and conference papers) in the AIS’s electronic library[1] and many more results in other databases that included other disciplines: 77,774 in EBSCOhost (across all databases), 81,653 in Scopus, and 110,352 in Web of Science (across all databases). Adding the qualifier “qualitative” proved less fruitful. In our search, we used search terms such as “qualitative big data”, “qualitative” AND “big data,” and “CAQDAS” (computer-assisted qualitative data analysis), but this resulted in an insufficient number of articles for a review or produced a set of results without cohesion in terms of the fields or themes of the studies.

To identify the correct search terms, we engaged in a manual staged review of articles in the AIS Senior Scholars’ Basket of Eight Journals[2] (European Journal of Information Systems, Information Systems Journal, Information Systems Research, Journal of the Association for Information Systems, Journal of Information Technology, Journal of MIS, Journal of Strategic Information Systems, and MIS Quarterly) and six journals with explicit focus on big data (Big Data & Society, Big Data and Information Analytics, Big Data Research, Frontiers in Big Data, IEEE Transactions of Big Data, and Journal of Big Data). We excluded conference papers from the review because we preferred the journal articles’ longer format, which allowed the authors more room to provide more details on their studies and the methods they used.

Articles published in these journals between 2017 and August 2021 comprised an initial sample of 2,862 articles. We evaluated the headline and abstract of each article against the following selection criteria: 1) an article must be empirical and 2) it must have used qualitative big data in the form of actual text, videos,

[1] https://aisel.aisnet.org/

[2] This was before “the Basket of Eight” was renamed “Senior Scholars’ List of Premier Journals” and then extended to include three more journals: Decision Support Systems, Information & Management, and Information and Organization.

pictures, and/or audio, instead of aggregates or metadata, such as string length, number of posts, ratings, retweets, likes/votes, or follower count. In unclear cases, we studied an article's data collection section more closely to decide whether it should be included in the sample. In the subsequent in-depth reading of all articles, we dropped isolated articles when we were unable to confirm the amount of data they had collected. The largest, completely manually studied sample that we came across comprised 23,000 tweets (Vaast et al., 2017). In this review, we set a threshold for observations at 30,000 to ensure adequate difference between big and small data studies while being mindful not to exclude big data studies from the smaller end of the scale. Of the 2,862 articles, 75 were included in the first phase of the review.

To extend our review, we created a search query (Table 1) based on the most common data sources, methods, and topics or phenomena in the articles reviewed in the first phase. We obtained the second sample of 450 articles by concentrating our search on the titles, abstracts, and keywords in the subject areas of business, management, and accounting (255 articles); decision sciences (129 articles); and economics, econometrics, and finance (66 articles) in the Scopus repository. After controlling for duplicates, we included 94 articles from the sample using the same criteria as in the first phase. This brought the total number of reviewed articles to 169. The reviewed articles are listed in Appendix A, together with their data sources and the initial and final dataset sizes.

Table 1. Query for Qualitative Big Data Articles in the Scopus Repository

Type of data	"big data" AND
Sources	Twitter OR forum* OR blog* OR Weibo OR YouTube OR Airbnb OR Amazon OR Facebook OR Yelp OR Bitcointalk.org OR StockTwits.com AND
Method	social OR text OR aspect OR "machine learning" OR LDA OR LSA OR multi* OR supervised OR unsupervised OR vector OR "deep learning" OR "sentiment analysis" OR analytics OR "content analysis" OR "topic modelling" OR "topic mining" OR "data mining" OR "convolutional neural network" OR "natural language processing" AND
Topic/phenomenon	online OR review* OR recommendation OR crowdsourc* OR "fake news" OR automated OR customer OR cyber* OR digital OR "word of mouth" OR financial OR stock OR health* OR chronic OR medical OR information OR political OR visual

Compiling the search query based on the articles from the first phase poses one benefit and one drawback. Improved accuracy (75 out of 2,862 vs. 94 out of 450) helps us zoom in on the relevant literature among several thousand big data articles. However, we now have a very specific picture of what a study using qualitative big data looks like (i.e., coming from certain kinds of sources and applying specific methods to bounded topics or phenomena). In other words, in the second phase, we excluded large parts of big data research. Nevertheless, we believe that the 169 selected articles comprise a representative—if not exhaustive or comprehensive—sample. Toward the end of our review, we encountered fewer and fewer articles presenting new challenges and solutions, suggesting that we had reached saturation.

2.2 Review and Coding of Articles

In the first phase, we developed a coding scheme to capture each article's keywords and phrases, data source(s), initial and final sample sizes (some studies further narrowed down their initial sample in pursuit of relevant data), perceived challenges, and possible mitigation strategies. These challenges were either explicitly stated in the paper or inferred from careful reading. Although some of the articles analyzed a significantly smaller subset than their initial dataset, we still included them in our review, because identifying relevant data is one of the key challenges of doing big data research. For example, studying the content of a small and clearly delineated Facebook group does not constitute big data research, even though Facebook itself is an important source of big data. Then again, identifying 15,000 relevant tweets with topic modeling out of hundreds of thousands or millions of tweets does.

We also captured data sources to examine their diversity. The single most common source in the first phase was Twitter (today known as X). In particular, 21 studies collected all of their data from Twitter, and seven used Twitter data, among other sources. Other major platforms, such as TripAdvisor, Amazon, Weibo, Facebook, and YouTube, were also represented, in addition to a range of other platforms, databases, and online communities. Twitter was also the most common data source (82) in the second phase, in which we used names of platforms (e.g., "Twitter", "Facebook", and "Weibo") as part of our query. In both phases, the sample sizes were recorded to ascertain whether a study could be described as big data research. With shorter texts (e.g., tweets and reviews), the majority of sample sizes varied from hundreds of thousands to millions, whereas with longer texts (e.g., loan applications, petitions, and news articles), sample sizes varied in tens of thousands but above the set threshold of 30,000 observations. Finally, we categorized the perceived challenges and their solutions—the main focus of this study—under common themes.

3 Identified Challenges with Qualitative Big Data and Their Solutions

Across the 169 reviewed articles, we identified three main categories of challenges: 1) locating relevant data, 2) addressing noise in the data, and 3) preserving data richness. Addressing these challenges often involves supervised or unsupervised ML. In supervised ML, the algorithm is given examples of relevant data or what to look for. The dataset is then split into training data (also referred to as "ground truth" or "golden standard"), which are used to train the algorithm but not for the actual analysis, and testing data, to which the trained algorithm is applied. The training data are annotated by the researcher, who "tells" the algorithm that, for example, a picture contains a cat, a certain message is an example of cyberbullying, or a specific tweet expresses a positive sentiment. Given that the training data are a sample of the whole dataset and their contents and composition are known, the algorithm's accuracy in testing data can be evaluated against the training data results. In comparison, unsupervised learning does not use training data and is applied directly to the data without being explicitly told what to look for. We present a summary of the identified solutions within each category of challenges in Table 2 at the end of this chapter.

Among the reviewed studies, only Triantafyllidou et al. (2017) and Yu et al. (2019) reported challenges related to velocity, as they engaged in real-time image recognition and text analytics rather than data collection and subsequent analyses. Similarly, the reviewed studies do not perceive the process of collecting qualitative big data as a challenge per se. In the era of social media and different digital repositories, collecting data is often a mere technical exercise of web scraping and API usage. However, as previously noted, the representativeness of the data must be considered in terms of whether they sufficiently represent the population of interest (Boyd & Crawford, 2012; Lazer & Radford, 2017; Mills, 2018), as well as the ethical and legal limitations related to, for instance, informed consent and privacy of the subjects (see, e.g., Kobayashi et al., 2018; Webb et al., 2017). Furthermore, the collection of big data is not necessarily quicker than collecting data with interviews or surveys. For example, Schlosser et al. (2021) took 749.3 hours, or a little over 31 days, to retrieve 119.5 million tweets in their study. Among the reviewed works, only Asr and Taboada (2019) reported collecting quality data as a challenge in their study on fake news and misinformation detection. They tackled the challenge of the lack of a single sufficient dataset of fake news by combining several small fake news datasets. In their case, they used a straightforward approach of combining different datasets because the data were uniformly presented as either true or fake news.

3.1 Locating Relevant Data

A distinction can be made between data in principle and data in practice (Jones, 2019): the former refers to data that are recorded but not accessed or analyzed, while the latter refers to data put into actual use. In principle, 10 million tweets comprise a large amount of data, but in practice, only two million tweets might be relevant to a particular study. As the volume of big data exceeds the typical research team's ability to absorb what is in the data or worthy of analysis (Davidson et al., 2019; LeBaron et al., 2018), locating relevant data might pose a challenge. All big data are voluminous, but arguably, the challenges with qualitative data are different from those related to quantitative data, which are typically structured. Furthermore, quantitative data depict information that is most often accurately known to researchers. However, this is not always true with qualitative data. For example, for their 10 gigabytes of plain text data identified as relevant, Huang et al. (2017, p. 1184) also had to collect "a nontrivial amount of irrelevant data", thus highlighting the issue of finding the proverbial needle in the haystack. Two general strategies are used to locate relevant data from a larger corpus of big qualitative data: the dictionary approach, which is divided into general and customized dictionaries, and the concentrated approach, in which the search is focused on events or is based on other descriptors.

3.1.1 General-purpose Dictionaries

Using general dictionaries and creating data-specific dictionaries are different ways of analyzing content and generating keywords. General-purpose dictionaries capture given emotions or are tailored for a certain purpose, but they are not specifically designed for individual researchers' data. For example, Benabderrahmane et al. (2017) used the vocabulary provided by the French Ministry of Labor to conduct a coherent semantic analysis of heterogeneous job boards with widely varying job descriptions. They then used the vocabulary to build a recommendation system that could discern the best online job boards for specific job offers. Mai et al. (2018) used a financial sentiment dictionary with 2,329 negative and 297 positive words to locate positive and negative messages, enabling them to study the connection between social media sentiments and the future monetary value of Bitcoin. Similarly, to detect abuse in in-game chats, Cécillon et al. (2019) used a list of abusive words and a technique called message collapsing, in which letters occurring more than twice consecutively were removed (e.g., from “loooool” to “lool”). Meanwhile, medical dictionaries have been used to identify domain-specific stop words, which are words that do not add meaning to the sentence and the removal of which increases information gain (Sundararaman et al., 2018), to break unstructured questions and answers into entities and to extract high-quality medical information from crowdsourced medical Q&A websites (Y. Li et al., 2020). Medical dictionaries have also been used to map health-related terms in messages against professional health terminologies to determine the level of health literacy in online health community discussions (L. Chen et al., 2019).

3.1.2 Tailored Dictionaries

General-purpose dictionaries might perform poorly in capturing the nuances or particular characteristics of a dataset. For example, a general-purpose sentiment dictionary might not recognize that “long” and “short” can also refer to investors’ sentiments in finance (Sanford, 2022). In cases such as these, researchers have developed tailored dictionaries for their specific use cases to better capture all relevant data. For example, to filter customer complaints from compliments and messages that seek or share information, Gunarathne et al. (2018) tailored a dictionary by building a lexicon of 326 complaint n-grams and 354 compliment n-grams out of 2,000 tweets. N-grams are sequences of co-occurring words used in natural language processing (NLP). In the same vein, while consumers’ search intent varies, the keywords they use might be the same, which is why a study on the effects of keyword ambiguity on search advertising performance generated an intent-based dictionary (Gong et al., 2018). Similarly, considering that researchers who identify words associated with disease outbreaks might miss colloquial and informal terms, Mejia et al. (2019) used a probabilistic naïve Bayes classifier[3] to create a tailored dictionary based on Yelp reviews. In this way, the words and phrases (e.g., “pungency”, “barely edible”, and “wiping nose”) that the authors might have failed to include in the dictionary and were unlikely to be included in any other predefined dictionary were included in the dictionary. However, unless the dictionary also contains phrases (comprising up to two or three words) apart from single words, there is a risk of oversimplifying the language. For example, had the dictionary by Mejia et al. (2019) used only single words, it would have captured common simple words (e.g., “nausea”), but it would have missed the richer and more varied longer phrases described above.

Numerous studies on online health communities have used a manual content analysis approach. However, this approach quickly becomes unfeasible when the volume of data increases and other solutions are needed. For example, L. Chen et al. (2019) manually coded 3,086 replies from an online health community forum and used them as training data for the support vector machine (SVM)[4], and finally used a trained classifier to code the remaining replies. Existing dictionaries may also be combined and improved. For instance, Sun et al. (2021) combined two dictionaries (HowNet and NTUSD) and added high-frequency words they identified themselves to create a sentiment dictionary of 48,878 words. Another group of researchers (Yin et al., 2020) similarly enhanced the dictionary used by the SnowNLP database to process Chinese text with HowNet and NTUSD dictionaries encompassing most Chinese vocabulary. The authors then combined the new dictionary with a domain-specific BosonNLP dictionary to gauge how investors’ daily sentiments affect the liquidity of Chinese stocks.

Topic modeling, an unsupervised ML approach to finding clusters of associated words and phrases (i.e., topics in documents), can also be used to locate relevant data. When key terms do not occur simultaneously, arranging terms under topics can help identify relevant documents (Dowling et al., 2019; Ngoc et al., 2019). For example, rather than trying to produce keywords to locate texts on the taste or mouthfeel of a certain type of coffee, these can be turned into topics (i.e., “bags of words” containing words such as “finishing”, “mouthfeel”, and “aftertaste”), thus allowing the algorithm to identify relevant texts. Another study adopted a similar approach in investigating how information accumulated while helping others in a crowdsourced support community affects the quality of new product ideas (Hwang et al., 2019). To determine who had

[3] Naïve Bayes classifiers are a family of probabilistic algorithms based on Bayes’ theorem. They are “naïve” in the sense that they expect all individual features to have independent and equal contributions to the outcome.
[4] SVM is a supervised method used for linear and nonlinear classification. If data are not initially linearly separable, they are transformed in a way that allows linear separation, and the new characteristics are then used to predict which group a new record should go to.

contributed to which topics and what these topics were, the authors generated 115 topics by applying latent Dirichlet allocation (LDA)[5] to two million messages. Similarly, to identify tweets related to inflation among 11.1 million tweets, Angelico et al. (2022) applied LDA to create 50 topics from which they chose two inflation-related topics with the top 10 words (e.g., “inflation”, “wages”, “deflation”, “euro”, and “price”) from the associated tweets for further examination. Then, they created N-grams from the 1,534,743 remaining tweets with words such as “price,” “expensive,” and “inflation” and manually coded whether the N-gram referred to increasing or decreasing inflation. Finally, to ascertain that the tweets had captured inflation expectations, they compared the results against official surveys and market-based inflation expectations.

3.1.3 Concentrated Search

Relevant data can also be located by better focusing the search, for example, on specific events or other descriptors. For instance, in a study on how a social movement used a popular multiplayer game as a virtual environment for awareness raising, McKenna (2019) identified examples of awareness raising by focusing their search around times when game patches were introduced in addition to keyword searching, rather than perusing and absorbing the complete dataset of forum messages. Similarly, Yue et al. (2019) had to locate relevant content among 2,960,893 posts in 355,222 threads in their study on the impact of online hackers’ forum discussions on the extent of distributed denial-of-service attacks. The authors identified relevant posts by looking for those with mentions of the port numbers listed in the threat databases. Port numbers on computers are used to differentiate transactions over a network (e.g., web service, mail service, and file transfer). In this case, they could be used to determine what service or vulnerability the threat was targeting, allowing the authors to zero in on relevant posts. Luo et al. (2017) focused on the influence of expert blogs on nine specific brands and managed to narrow down 1.25 million Google blog search results for just one brand to 131,759 blog posts for all brands using Technorati, a specialized blog search engine.

3.2 Addressing Noise in the Data

Often a researcher does not have control over how qualitative big data is created (Schmiedel et al., 2019) which is why even data deemed relevant might contain unrelated, incorrect, misclassified, and sometimes fraudulent records. For example, a study comparing COVID-19-related tweets by news organizations and citizens had tweets by other organizations mixed with citizens’ tweets (Han et al., 2021). Together, the irrelevant data items and those of uncertain reliability obscure the relevant data, hampering the algorithms’ performance by making them learn from unrelated or false data. This issue can affect the veracity of the results. Although noise is present in both quantitative and qualitative big data, arguably, noise in qualitative data can be more difficult to address due to the unstructured nature of data and the higher proportion of noise (Schmiedel et al., 2019). While strongly connected to identifying relevant data, different methods of annotating or labeling data can also be considered a form of noise reduction. These methods include filtering out irrelevant data, addressing out-of-vocabulary words, and applying techniques to improve data accuracy.

3.2.1 Filtering Out Irrelevant Data

In principle, many words that are correctly related to a topic can, in practice, also be used in unrelated contexts, indicating that the inclusion of statements or comments with these words in a study could add noise. For example, in a study about asthma triggers and risk factors, a social media comment, “I will call you ‘asthma’, because you take my breath away” (W. Zhang & Ram, 2020, p. 313), includes the correct keyword (“asthma”), but the sentence is clearly not related to asthma triggers or risk factors. Moreover, when studying rumors about a police officer being blinded by a firecracker or nuclear weapons being used by the police during the 2017 G20 summit demonstrations, tweets on police being metaphorically blind, or discussions on antinuclear treaties are irrelevant; thus, they are considered noise in the data (Jung et al., 2020). Similarly, when a study’s aim is to identify the sales and promotion of wildlife products on Twitter using ML, talk of banning the sales of ivory, rather than selling it or discussing wildlife preservation and related laws and policies, might contain relevant keywords but are false positives; thus, these are also considered noise (Xu et al., 2019). In another example, an author had to remove tweets about apple orchards when studying sentiments toward Apple, the tech company (Sanford, 2022).

[5] Latent Dirichlet allocation (LDA) is a frequently used unsupervised ML method for topic modelling that views documents as combinations of topics and topics as combinations of words. Here, topics are created based on words’ probabilistic co-occurrence. However, LDA ignores word order and each word’s grammatical role.

The solutions to these challenges vary. For example, W. Zhang and Ram (2020) combined NLP (e.g., N-grams, ML, and adapting algorithms trained previously in a related domain) and compared the prevalence of “asthma” in tweets to its prevalence in the US adult population. Jung et al. (2020) originally had a dataset of 736,577 tweets but eventually analyzed only 6,095 tweets by concentrating on the most active rumor spreaders and debunkers. Xu et al. (2019) initially conducted a manual search on Twitter to identify keywords for automated data collection, which could then be used in topic modeling, thus enabling the authors to focus on topics that had the strongest signals of the ivory trade. Sanford (2022) created a list of word combinations (N-grams) that should not exist when talking about Apple (the company) (e.g., the word “orchard” preceding or following the word “apple”) and then manually ensured that the few thousand tweets removed were not related to the company.

Meanwhile, not all content on social media platforms such as Twitter is created by humans but by “bots” (i.e., software applications programmed to perform certain tasks). The messages posted by bots may distort data by getting mixed with authentic users’ opinions and sentiments, thus warranting their removal. For example, to remove bot-generated messages, X. Liu et al. (2017) followed bot-detection criteria (Chu et al., 2012; Wang, 2010) that focus on the tweets’ contents and the tweeting accounts’ social graph features. Given that bots are frequently used for spamming, their messages are often repeated or presented simply as links. Furthermore, the degree of regularity in their tweeting is uncharacteristic for humans. Thus, a stark difference likely exists between the number of accounts followed and those following the suspected bot account, as well as the ratio between replies and mentions (Chu et al., 2012; Wang, 2010).

3.2.2 Addressing Out-of-Vocabulary Words

Informal language, abbreviations, misspellings, punctuation errors, nondictionary slang, wordplay, comparative sentences, negations, and double negations, transferred negations, sarcasm, unwanted languages, spam, and emoticons all constitute noise for algorithms if not properly addressed (Ho et al., 2019; Khazraee, 2019; Ordenes et al., 2017; W. Zhang & Ram, 2020). Often, such noise is addressed by preprocessing data through the removal of certain message features, such as misspellings, slang, or emoticons (T. Li et al., 2018). However, in doing so, some part of the richness of the data might be lost. For example, uppercase and extended words (e.g., “FUNNYYY” and “looool”), emoticons, and URLs included in a text could also be used to decipher sentiment and intent instead of simply being processed as noise (El Alaoui et al., 2018). In detecting informal language, Cury (2019) verified data against a dictionary of one million distinct words to spot out-of-vocabulary words for further examination. The author then corrected the words not found in the dictionary to comply with grammar rules, ignored them, or subjected them to further analysis to determine their sentiments. This approach—as opposed to removing all sources of noise—improved the algorithm’s results. The out-of-vocabulary problem posed by informal language and extended words, for example, can also be addressed with a subword feature, which breaks a rare or unknown word into more frequent subwords that operate between a single character and a complete word; thus, for example, “goooood” is recognized in the same way as “good” (C.-Y. Huang et al., 2019).

People can also use different words to describe the same thing, implying that algorithms might have to address multiple synonyms and polysemes, which, in turn, increase computational requirements. Zhou et al. (2018) addressed this issue by reducing keywords from 41,101 to 8,435 using the singular value decomposition [6] technique, which made text processing more efficient and reduced computational requirements. When investigating user connections across eight social networks, Cheung et al. (2018) found that user-generated tags proved unreliable due to language and cultural differences among users. Building on the idea that users who share similar images likely have a connection that can be used to improve recommendation engines, Cheung et al. (2018) used a convolutional NN (i.e., an algorithm that assigns weights for an image’s features and is able to classify images according to those weights) to annotate over two million images. With this approach, they managed to reduce noise in the data and ascertained that connected users tended to share similar images.

3.2.3 Improving Accuracy

A word might have a different meaning and sentiment value depending on the sentence or context in which it occurs (e.g., “I’d kill for some cookies right now” versus “I’d kill without remorse”). However, some

[6] Singular value decomposition is a linear algebra technique for transforming a large dataset into a smaller dataset while preserving most of the information. This method promotes interpretability by focusing on the most important or dominant correlations or features in the data.

algorithmic approaches are unable to take this into account, thus causing incorrect annotations (Cury, 2019; Ngoc et al., 2019; J.-B. Zhang et al., 2017). Furthermore, negations do not necessarily change the meaning completely; for example, “not horrible” or “not too bad” do not mean the same as “amazing” or “great” (Ordenes et al., 2017). In some languages, a word can also denote a person’s name. For instance, in Arabic, the popular names “Saeed” and “Amal” also mean “happy” and “hope”, respectively, and their use depends on the context (Al Shehhi et al., 2019). In such cases, accuracy can be increased by conducting a joint analysis of a word’s local context, such as its syntactic and semantic features in a sentence and its global context, or the nature of the document or the paragraph in which the word occurs (Meng et al., 2020; Rintyarna et al., 2019).

When several different conversations take place in a single discussion thread (Abbasi et al., 2019), this can also generate noise in the data. For example, this could mean that some recorded sentiments are directed toward something a specific commentator said rather than the nominal topic or original post. Motivated by the need to address multiple conversations in discussion forum threads, Abbasi et al. (2019) converted five million posts into 26 million sentences with greater focus and consistency. Zhang et al. (2017) adopted a similar approach by analyzing documents at a paragraph level to better identify the relevant parts.

Another issue related to accuracy is the trade-off between the accuracy and coverage of the dataset; that is, accuracy may suffer if excessive variety is found in the data. For example, to balance accuracy and coverage, Geva et al. (2017) opted to use brand names only for cars without any additional refinements (e.g., model names) in their study, in which they located relevant data from all English-speaking forums indexed by Google’s discussion forum search. “Chevrolet Malibu” or “Chevrolet Spark” are, in principle, more detailed queries than “Chevrolet”, but they introduce noise to the sample in the form of irrelevant results, such as the “City of Malibu” in California. In their study, Toshniwal et al. (2019) removed non-English and extremely short tweets, because they would have confounded the identification of relevant tweets; that is, tweets of few words do not contain enough data, and the algorithm was tailored for content written in English. Furthermore, in the training of a facial recognition algorithm, the image background tends to add noise. In particular, the background does not contribute to facial recognition but adds unnecessary details that the algorithm does not know to ignore, thus decreasing its accuracy. To cancel this effect, backgrounds can simply be cropped out (Hashemi & Hall, 2020).

3.3 Preserving Data Richness

The extant literature has discussed the tendency of big data research to prioritize “tactical” topics and correlations over a richer understanding of the phenomenon (see, e.g., Grover et al., 2020; Hirschheim, 2021; Smith, 2020). In the reviewed articles, Y. Chen et al. (2019, p. 122) called for the development of “new theories for capturing linguistic and other patterns in the rich, abundant content generated by ICT communication”. In comparison, McKenna (2019) observed that social movement theories used to understand social movements online originate from the “pre-online times”, thus necessitating new theoretical approaches better suited for online environments. While qualitative data and research are denoted by rich descriptions, the quantification of qualitative data reduces such richness. Thus, due to its volume, a large unstructured dataset must be (at least partially) transformed into a structured form to facilitate quantitative analysis. Researchers use a range of algorithms to quantify and study voluminous text data, the most prevalent among the reviewed articles being LDA for topic identification and modeling; latent semantic analysis (LSA)[7] for finding common themes among documents, building domain-specific dictionaries, and reducing the dimensionality of data; and SVM for classifying sentiments. However, the algorithms might not, for example, be able to account for the dataset’s temporal aspects or the evolution of topics over time in a discussion forum. Furthermore, the algorithms might disregard grammar and word order or even ignore some of the data features, such as informal language, emoticons, or URLs, which could help determine the author’s sentiments and intent. While humans can account for many of these issues almost subconsciously, algorithms must be explicitly told how to address each issue. Researchers have addressed these challenges by creating training datasets and/or using different methods to improve data classification.

[7] As an unsupervised approach, LSA assumes that words with similar meaning occur frequently together, creating a term–document matrix. The matrix is then reduced in size, leaving only the most important terms. LSA can partially capture polysemy.

3.3.1 Creating Training Datasets

Annotating data is a labor-intensive task and is often prohibitively expensive to perform for all data. This is especially true with highly domain-specific information (e.g., medical data) that requires several expert annotators (Y. Li et al., 2020). Therefore, training datasets tend to be smaller than full datasets (W. Zhang & Ram, 2020). This could mean that small training data might not completely capture the richness or essential features of the whole dataset. Document-level features could include a sentiment, topic, gender, or whatever it is that the researcher is interested in. In turn, dataset-level features could include different sentiments contained in the data, the ratio between them, how the phenomenon has developed over time, and a general understanding of what is contained in the data. For example, if the training dataset is too small, it might not include all possible sentiments, which means that the algorithm may not learn to identify all relevant sentiments. In turn, this might affect the perceived ratio between negative and positive sentiments in the dataset.

Often, thousands of observations are annotated by researchers, junior faculty, or the workforce recruited from the student population to gain familiarity with the data and create training datasets (see, e.g., L. Chen et al., 2019; Gunarathne et al., 2018; Samtani et al., 2017; W. Zhang & Ram, 2020). However, if this is not sufficient, researchers often turn to crowdsourcing from online platforms, such as Amazon Mechanical Turk, Crowdflower, or Clickworker. While crowdsourcing annotations can be quite challenging in terms of label noise or intercoder agreement (e.g., different dog breeds could be confused and receive wrong labels, which is an example used by Y. Chen et al. (2020) in their paper on addressing label noise), the issue subsides when there are several annotators per message or datum. T. Li et al. (2018) had each message manually classified by at least five different annotators, while H. Geva et al. (2019) used three crowdsourced annotators to verify their own annotations.

Furthermore, Gray and Suzor (2020) found that a relatively low amount of manual coding would be needed for transfer learning (i.e., finetuning pretrained general models) to achieve a highly accurate ML classifier and avoid misclassification. The training time is also shorter than training the algorithm from the beginning (Tao & Fang, 2020). However, transfer learning still requires an adequate amount of training data (Mukherjee et al., 2022). Smaller training datasets are needed to achieve a well-working model with multimodal data, such as an image and accompanying text (Lopez et al., 2020; Wan & He, 2019), as opposed to having just a single data type. In learning from multimodal data, the different features extracted from various data types are fused into one representation of features, thus providing the algorithm with richer data to work with than it would receive from working only with texts or images. However, in the case of images, mirroring existing images easily doubles the size of the training data because the algorithm regards the mirrored images as new images (Triantafyllidou et al., 2017).

3.3.2 Improving Data Classification

Big data analysis tends to reduce complex issues and nuances to “clean” charts and other forms of quantification. However, this can hide the underlying messiness and produce particular and reductive interpretations and ways of seeing. Thus, deep thematic knowledge should accompany big data analysis to avoid superficial understanding (Walker & Boamah, 2020). Similarly, X. Liu et al. (2020) reported that their theory-driven algorithm, which is based on the knowledge adoption model, fared better than algorithms without theoretical guidance in terms of predicting the usefulness of information found in online knowledge communities. The authors further proposed that combining ML and extant IS theories yields better explainability.

Certain algorithms lack the ability to understand context. For example, an online petition might be considered unreasonable to begin with (see, e.g., a petition for “fixing bias” in democracy by giving more votes to those who pay more taxes[8]), and this is something the algorithms cannot account for (Y. Chen et al., 2019). On the one hand, unsupervised ML algorithms often focus on term frequencies and might miss less frequent terms; thus, their results are not representative of the whole content but merely a picture of what is popular (Guo et al., 2017; Ngoc et al., 2019). Guo et al. (2017) solved this problem by categorizing contents according to their similarities and then sampling them, thus ensuring that less popular topics were also selected for reading. On the other hand, supervised ML algorithms are limited by the requirements of predefined topics, which is why interesting topics may remain hidden (Gong et al., 2018). However, in unsupervised ML, a topic could emerge when an algorithm is told to create eight topics, but the topic would

[8] https://www.kansalaisaloite.fi/fi/aloite/432

not emerge when asked for less or more than eight; in this case, the number of topics can be manually varied to triangulate the topics in the data and evaluate their robustness (see, e.g., Asr & Taboada, 2019; Gong et al., 2018; Gray & Suzor, 2020). In addition, the topics produced by LDA are static, which means that the algorithm uses all the given data and produces a predetermined number of topics, thereby ignoring the temporal development of the topics. Thus, if researchers are studying the evolution of the most salient topics at different points in time, they can, for example, run LDA several times for every December data in the dataset rather than analyze the entire dataset in one go (X. Liu, 2020).

Sentiment analysis is used to classify text as positive, neutral, or negative or to detect some predetermined emotions, such as trust, surprise, joy, and disgust. The predetermined emotions vary between dictionaries, which means that interesting emotions might go unnoticed, thereby affecting the veracity of the results. For example, Q. Liu et al. (2020) studied crowdsourcing communities for open innovation and analyzed 43,550 product ideas submitted by users to the new product development community of an electronics manufacturer. They then categorized the feedback valences into positive and negative. However, merely classifying sentiments does not tell us the objects of the users' positive or negative feelings, thus decreasing the richness of the data. To a certain degree, this problem can be alleviated by using aspect-based sentiment analysis to connect sentiments to particular aspects (Ho et al., 2019; Rintyarna et al., 2019). For example, beyond classifying tweets about an airline as positive or negative, Ho et al. (2019) connected these sentiments with statements about the airline staff, luggage, comfort, airport, and punctuality, thus producing finer-grained data. Another option would be to use the apriori algorithm to establish association rules between sentiments and different issues, such as cabin crew behavior, food quality, and loss of baggage (Kumar & Zymbler, 2019).

Acknowledging the idea that relying solely on machine-based interpretations of nuances in language and symbols may constrain the depth of insights that can be obtained, researchers have integrated big data with manual content analysis (see, e.g., Bhattacharjya et al., 2016; Brooker et al., 2018; Rossi et al., 2021). While manually studied samples are often significantly smaller, they help address algorithms' weaknesses by identifying more detailed nuances and providing temporal insights.

3.4 Summary of General Solutions

Some of the identified challenges (e.g., the same words being used to describe or search for different things, or the opposite case where diverse words are used to refer to the same thing) are quite specific and only mentioned in one or a few articles. In comparison, more frequently encountered challenges were related to the difficulties in perusing and absorbing what is in the dataset to locate relevant data, filtering out content that contains seemingly relevant words but is unrelated to the study's goals, transforming qualitative data into machine-readable form while retaining specific dimensions or features in the data, and accounting for the different ways of using the language. Table 2 presents an overview of the most common solutions for a given challenge category. The numbers in square brackets refer to the numbering of the reviewed articles listed in Appendix A. A more detailed account of the challenges and solutions can be found in Appendix B.

Table 2. Identified Challenges and Solutions by Category

Challenge	Solutions
Locating relevant data What data should be collected or kept? All available or collected data are not necessarily relevant to answering the research question. The volume of data makes manually perusing each document and absorbing what is in the data unfeasible.	General-purpose dictionaries can be used to locate relevant data when the meaning of a dictionary's content does not vary between contexts (e.g., a list of abusive words [21], medical terms [70, 155], and official labor classification codes [13]), or when the premade dictionary otherwise applies to the data even if not specifically created for it, such as financial sentiments [78]. Tailored dictionaries can be created when there is reason to assume that a general dictionary cannot adequately capture nuances in the data or when one does not exist. A dictionary of the data may be generated through N-grams [47], classifiers [70, 84], or topic modeling [7, 92, 101]. A tailored dictionary can also be created by combining dictionaries [128, 163]. Conducting a concentrated search around events of interest [83], guiding the search with other data [166], or using a specific search tool [74] when perusing and absorbing all available data for relevant data is not possible.

Addressing noise in the data Among relevant data, there might, for example, be misclassified or fraudulent records, or simply a myriad of different ways of using the language. The meanings of certain words might depend on specific contexts, and the algorithm must be specifically told to address this individually.	Irrelevant data can be filtered out so that false positives (i.e., those related to the phenomenon under study but actually referring to a different context) do not confound the algorithms. This has been achieved, for example, by creating N-grams that should not exist in the given context and checking the data for them [115], using topic modeling to focus on the content with the strongest connections to the research question [101], focusing on the most active discussants for manual perusal [60], and using natural language processing (NLP) and comparing whether the prevalence of the phenomena in the data matches the phenomenon's prevalence in the population [138]. Out-of-vocabulary words can be addressed in various ways so that they do not affect the algorithms' performance, such as by removing noisy features [129, 132, 138], addressing informal language by checking words not in the vocabulary, and then changing them to comply with grammar rules and determining their sentiment [31], or breaking extended words into smaller elements [19]. Different ways of using language may be circumvented by focusing instead on shared images [24]. Accuracy can be improved by comparing a word's local and global context to infer its meaning and sentiment [85, 105], breaking texts into smaller portions to better understand different sentiments [1, 59], avoiding complex search terms with too much coverage producing false positives [130], and removing extremely short messages and those appearing in unwanted languages [135].
Preserving data richness Addressing noise usually means removing certain features from the data. The quantification of qualitative data might create reductive interpretations or lead to a superficial understanding of the phenomena under study.	Training datasets can be created so that they capture the richness and essential features of the dataset by using crowdsourcing platforms to increase the amount of annotated data [49, 132], using pretrained ML-models [44], and taking advantage of multimodal data [73, 140]. Data classification can be improved with deep domain knowledge by [139] using theories to guide the process of choosing and creating algorithms [149], ensuring representative sampling [48], triangulating topic models [9, 43, 44], connecting classified sentiments to what it pertains to [54, 67, 105], and combining big data analytics with manual content analysis [14, 17, 106] to avoid a superficial understanding of the phenomenon and deceptively clean representations of messy and nuanced data.

4 Discussion

In line with the general characteristics of big data, the practical challenges identified in our review stem primarily from the volume, variety, and veracity of the data. While voluminous data typically have benefits, such as richness, added rigor, exhaustiveness, and the possibility of uncovering many interesting relationships, the volume also makes it difficult to locate relevant data or form a full understanding of such data. Furthermore, the variety of communication styles and data types (e.g., texts, pictures, and videos), as well as different data sources, require technically diverse skills and introduce various types of noise. As mentioned previously, the volume of data necessitates the use of algorithms, but the accuracy and performance of these algorithms—and, by extension, the veracity of their results—are hampered by several challenges, such as the use of informal language and many conversations taking place simultaneously. Inauthentic user accounts, such as bots, trolls, and vendors writing positive reviews of their own products, further challenge the veracity of data. Challenges related to the dynamic nature of data (velocity) were not widely encountered because the studies were mostly backward-looking and adopted a snapshot view rather than a continuous, real-time approach to data collection. Nevertheless, we note that datasets containing tens or hundreds of millions of entries would not exist without organizations capable of recording the flow of data in the first place. For example, Twitter produces data in the petabyte range and processes around 400 billion events every day (L. Zhang & Malife, 2021). In the following subsections, we discuss the intertwined nature of the challenges we have identified and how this creates trade-offs in qualitative big data research.

Then, we outline recommendations to address the trade-offs without exacerbating the challenges in other categories.

4.1 Intertwined Challenges in Qualitative Big Data Research

We identified three main categories of practical challenges: locating relevant data, addressing noise in the data, and persevering in data richness. While these challenges stem primarily from the volume of data, we acknowledge the intertwined nature of the three categories (Figure 1). The volume of the data makes it difficult to locate relevant data and form a full understanding of such data. Furthermore, increased volume can lead to increased variety (i.e., how the data are structured and the types of data created) and variance. Increased variance may depend, for example, on the motives (e.g., genuine opinions, trolling, or posting faulty reviews) and communication styles of those producing the data (e.g., tweets). This variety not only represents the richness of qualitative data but also introduces multiple sources and types of noise, in which the greater the variety in data, the greater the likelihood of noise present in them. High data volume also impedes manual data perusal and analysis, thus necessitating the use of algorithms for data processing. However, as the algorithms quantify the data by removing, ignoring, or misclassifying what they do not understand, this process can lead to difficulties in fully preserving and maximizing the richness of the qualitative data. Furthermore, when a large portion of the dataset is studied using probabilistic algorithms, the ability to establish veracity is affected. The abovementioned connections among the identified challenges prompt us to interpret the interrelationships from a high level of abstraction in the following ways: (1) prioritizing high volume means concessions regarding variety or veracity, (2) pursuing high veracity means that either the variety or volume must be reduced, and (3) preserving richness means having reduced volume or veracity.

As noted earlier, velocity is not a widely encountered challenge for researchers. However, given that digital trace data—the most common type of quantitative and qualitative big data—are generated continuously, we depict its role as a challenge surrounding the other three challenges and as something that might or might not need to be addressed by researchers, depending on the data collection approach.

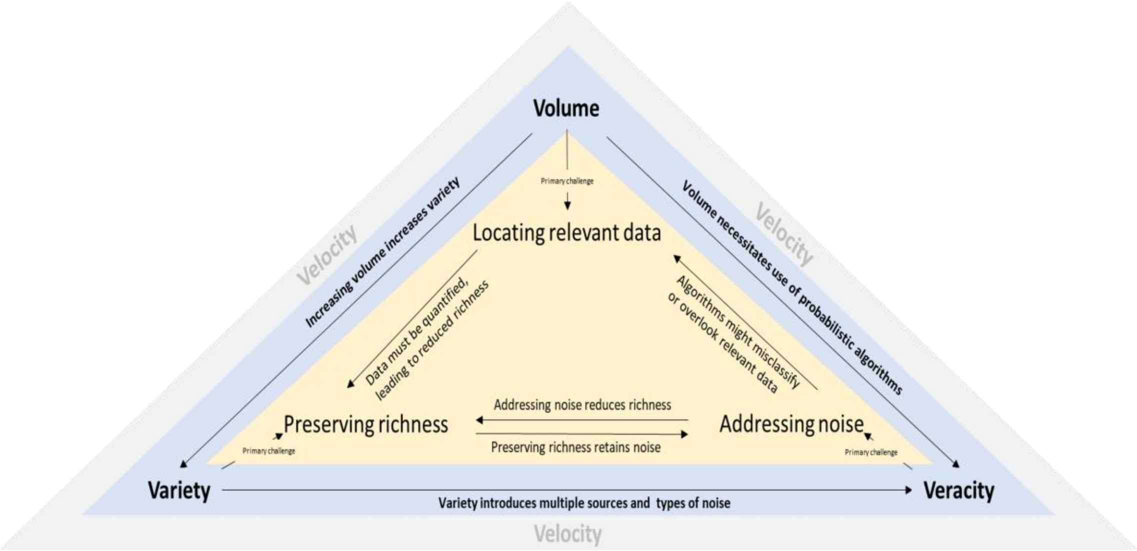
Volume
Primary challenge
Locating relevant data
Velocity
Increasing volume increases variety
Velocity
Volume necessitates use of probabilistic algorithms
Data must be quantified, leading to reduced richness
Algorithms might misclassify or overlook relevant data
Addressing noise reduces richness
Preserving richness
Addressing noise
Preserving richness retains noise
Primary challenge
Primary challenge
Variety
Veracity
Variety introduces multiple sources and types of noise
Velocity

Figure 1. Intertwined Challenges of Qualitative Big Data Research

If locating relevant data requires an algorithm, this means that the data must be quantified in one way or another, thus decreasing richness, for example, by ignoring the effects of context or the premise of a petition. Even if relevant data can be located without algorithms (e.g., by determining all petitions, incident reports, or lending applications to be relevant), the threat to data richness remains because of the need to address what is considered "noise" in the dataset. Therefore, addressing noise in data and preserving data richness appear to be largely opposite goals: one decreases data variety, while the other seeks to preserve it. Irrelevant data are considered noise, and their exclusion does not affect data richness. In fact, the removal of such data improves results because algorithms do not try to learn from or study unrelated pieces of data, and researchers are better able to identify and focus only on the relevant data. Yet, beyond detecting clearly irrelevant data and identifying and defining noise, finding ways to address this is not always a straightforward task. In fact, filtering irrelevant data is not a particularly easy task. For example, while words or messages written in all capital letters, with extra vowels, or those made completely out of emoticons might be considered noise from the algorithms' point of view because the algorithms do not know what to do with them, such data may still contain valid sentiment or intent. Thus, tension exists between established methods that often require removing features like these in data preparation and studies interested in sentiment as part of their research design. At the same time, if noise is not sufficiently addressed, the veracity of an algorithm's output becomes untenable and riddled by numerous instances of false positives and/or misclassifications.

However, the major challenge in working with qualitative big data is the task of preserving the richness of data to answer the "why" and "how" questions. Without addressing this challenge, big data research risks producing correlations, predictions, and superficial knowledge of the phenomenon being investigated but not necessarily a deeper understanding of it (see, e.g., Grover et al., 2020; Hirschheim, 2021; Smith, 2020). To date, locating the right data and treating them for noise appears to have more varied solutions than preserving data richness. Indeed, locating the relevant data and preparing it for analysis are necessary for there to be qualitative big data analysis; thus, they might have been considered more immediate challenges than preserving data richness. Nevertheless, going forward, improving the preservation of data richness is likely to increase in importance for qualitative big data researchers. Due to the intertwined nature of the aforementioned challenges, addressing noise and locating relevant data are connected to preserving their richness. However, they are not direct solutions but rather affect richness as a byproduct. Thus, preserving and using data richness when algorithms are needed due to volume is also a problem of determining how much of a given dataset the algorithms do not understand or capture.

4.2 Recommendations for Addressing the Trade-offs

The studies we reviewed demonstrate the difficulties involved in addressing the abovementioned trade-offs simultaneously. In particular, prioritizing one category might create increased or additional problems in other categories. While solving all three categories of challenges simultaneously may seem impossible, we outline some recommendations below that can be useful in addressing the challenges of one type without increasing problems in the others.

To begin with, addressing the interdependencies among the categories cannot be done using a single correct approach because of variations in the data and study objectives. Identifying the most crucial category and how to balance the systemic effects of the three categories depends on the type of data and the research question(s) set for a specific study. Data preparation should not be overly formulaic (Hickman et al., 2022), and we believe that this rule applies to addressing the challenges identified in this study. Specifically, a challenge should not be addressed without considering how it can possibly affect the other types of challenges. For example, the need to remove noise from the data was frequently mentioned in the reviewed articles but its effect on preserving richness was less discussed. Similarly, annotating data can affect challenges in different categories. Annotation is often required to locate and analyze data because increased volume prevents the perusal of an entire dataset. However, if the annotated data are too small compared with the whole dataset, uncertainty arises as to whether the annotated set captures all the relevant features in the data.

Thus, the first step in addressing the intertwined challenges is to consider all the different challenges and their corresponding systemic effects. We believe that qualitative big data analysis will benefit from explicitly considering how solving one challenge can affect other potential challenges and from examining why one data attribute (e.g., veracity) might be more important than another (e.g., variety) for a given study.

> **Recommendation 1: Consider the trade-offs created by the challenges. Researchers should explicitly consider what can and cannot be optimized regarding the intertwined challenges. They should also consider what is being prioritized at the expense of something else and why and how this affects the study.**

The earlier literature critiques big data research for being unable to answer "how" and "why" questions (Abbasi et al., 2016; Dalton & Thatcher, 2014; Grover et al., 2020) and for producing correlations without understanding why such correlations exist (Hirschheim, 2021; Smith, 2020). The literature recommends combining big and small data studies in a complementary manner (Dalton & Thatcher, 2014; George et al., 2014; Grover et al., 2020; Kitchin, 2014; Lindberg, 2020). This process begins by first looking at the big picture, identifying the underlying correlations among the elements within the dataset, and then focusing on interesting relationships with small data studies to answer the "why's" and "how's" of the phenomenon being investigated. Furthermore, combining big and small data and iterating between them as per abductive methodology (Lindberg, 2020) could also alleviate the issues of r-hacking and HARKing (Grover et al., 2020). Overall, working on qualitative big data is necessarily an iterative process, wherein the tasks of locating relevant data, addressing noise, and preserving data richness alternate and where the researcher shifts between the broad picture of big data and the nuanced information found in small data.

While preserving richness and addressing noise might be conflicting tasks given their largely opposite goals, a qualitative small data study could help minimize the negative effect of addressing noise on data richness. Even if a manual analysis of an entire dataset is impossible, analyzing a small subset of untreated, relevant data could generate richer, more nuanced understandings and insights that are not typically within quantitative methods' purview, while simultaneously evaluating the veracity of results.

> **Recommendation 2: Retain relevant, untreated data. Researchers should ensure, when feasible, that untreated, relevant data are available for analysis to supplement and possibly guide research by allowing for iterations between big and small data insights.**

The choice of an algorithm can also affect the preservation of data richness. For example, employing aspect-based sentiment analysis (see e.g., Ho et al., 2019; Kumar & Zymbler, 2019; Rintyarna et al., 2019) instead of regular sentiment analysis could preserve more details about the information at hand, thus enabling richer analysis. Furthermore, apart from the fairly common practice of creating tailored dictionaries, combining dictionaries could provide additional richness by infusing general-purpose ones with the domain specificity of tailored dictionaries (Sun et al., 2021; Yin et al., 2020). The reviewed articles presented examples of addressing noise by trying to accommodate rather than just removing features deemed noise, such as attempting to recognize and recode out-of-vocabulary words (Cury, 2019), shortening words with more than two same letters in a row (Cécillon et al., 2019), or breaking the out-of-vocabulary words into subwords (C.-Y. Huang et al., 2019), thus retaining data richness. Furthermore, changing the default noise addressing approach from deleting data into using algorithms and approaches that allow for including more of the data deemed noisy could also alleviate the tension between preserving data richness and addressing noise.

> **Recommendation 3: Prefer data preserving algorithms and approaches. Researchers should favor algorithms and approaches that preserve more data for analysis over more reductivistic ones and that best cater to the type of big qualitative data being collected and analyzed.**

Preserving data richness is not limited to how they are manipulated or processed. Both the lens through which researchers interpret data and how prepared they are to identify meaningful observations in the data affect the richness of the picture generated by the given data. Furthermore, concerns have been raised about how the application of ML without theory and domain knowledge makes it difficult to interpret the results or determine if they are meaningful (Hannigan et al., 2019; Kobayashi et al., 2018; Schmiedel et al., 2019; Smith, 2020). In the reviewed articles, Walker and Boamah (2020) likewise articulated that deep knowledge of the domain or of the phenomenon of interest should accompany big data analysis so that researchers can go beyond graphs and correlations. The theory-driven algorithms of X. Liu et al. (2020) also performed better than those without theoretical guidance in improving data classification and preserving data richness. As such, theories play an important role in addressing the systematic effects of the three challenges mentioned above. However, as McKenna (2019) and Y. Chen et al. (2019) noted new theories that are more suitable for online environments might be needed.

> **Recommendation 4: Leverage theories and domain knowledge. Researchers should build on theories and domain knowledge in terms of selecting or creating an algorithm and interpreting the data and the findings of their studies.**

4.3 Limitations

We acknowledge several limitations of this study. In the first phase, a staged review focusing on titles and abstracts, we may have missed articles that should have been included. At the same time, the 30,000-observation threshold for inclusion may have been too high or too low. We also excluded many articles because the amount of data they used could not be ascertained. Moreover, most of the studies in the first phase of the review used texts as their source, while videos or images were used by some. In comparison, no study used audio sources. This might simply represent the current state of qualitative big data, or it might be a shortcoming of our sampling. In either case, as we built our query based on the manually selected articles, the second phase of the review reflects a similar distribution between data types.

Constructing the second phase's query based on the manually selected articles improved our ability to identify relevant articles but also produced more of the same, which is a limitation of our study. Out of the tens of thousands of big data studies in diverse disciplines, we have reviewed a minor portion of similar studies in IS sciences and closely related disciplines. Furthermore, how identified issues and solutions apply to research using large language models, such as ChatGPT, is currently unclear. Nevertheless, even with these limitations, we believe that our review represents the challenges involved in qualitative big data analysis in the field of IS science and related disciplines. Extending the review to disciplines beyond the scope of business schools (e.g., to digital humanities) might prove beneficial because challenges with qualitative big data are universal, and disciplines with different traditions might have alternative solutions.

5 Conclusions

In this review, our purpose was to describe and synthesize the practical challenges of working with qualitative big data and the solutions found in the extant literature. We identified three intertwined categories of challenges: locating relevant data, addressing noise in the data, and persevering data richness. Given that the high volumes of qualitative big data generated today exceed the limits of human understanding, algorithms are needed to locate relevant data and form an understanding of such data. In preparing the data for algorithms, certain features (e.g., noise) might have to be removed. Furthermore, as such algorithms are often unable to account for grammar, the effect of context, or how sensible a message is to begin with, they might overlook many aspects of the data. Consequently, part of the richness the qualitative data are known for may be lost, thereby decreasing the ability of qualitative big data research to answer important "why" and "how" questions. The veracity of the results may also be affected because probabilistic algorithms might misclassify data.

Furthermore, the challenges might have opposite goals; for example, solutions to address noise by removing them can lead to further problems in maintaining data richness. Therefore, addressing one of the three categories might exacerbate issues in the other two categories. Currently, only a few good solutions have been proposed for preserving data richness after applying algorithms or for scaling up a researcher's capacity to absorb volumes of qualitative big data.

Arguably, the three challenges might be simultaneously unsolvable, and a qualitative big data researcher might have to prioritize exhaustive volume, high veracity, or high variety, essentially giving up one attribute to gain more of the others. Beyond these trade-offs, we identified eight general solution categories (e.g., using tailored dictionaries to locate relevant data and filtering out irrelevant data containing relevant keywords but in unrelated contexts) to address these challenges. Given the broad range of potential applications of qualitative big data, it is unlikely that universally suitable instructions for every situation would emerge. Nevertheless, we are able to propose four guiding recommendations. Even though these recommendations cannot completely solve all three intertwined challenges simultaneously, we hope that these can assist qualitative big data researchers in making informed choices on how to prioritize alleviating different problems with as little as possible need to compensate for the other intertwined challenges. In the future, as the development of tools and methods to study qualitative big data progresses, the division between qualitative and quantitative research might become increasingly tenuous (Lindberg, 2020). A more sensible division—if one must be made—might be the distinction between the small and big data parts of a study for the sake of iteration and abduction. This could also help preserve richness in the qualitative data, thus enabling a more nuanced understanding of the phenomenon being investigated and the underlying data.

Acknowledgments

An earlier version of this paper was presented at the Hawaii International Conference on System Sciences (HICSS-54,) in 2021. This research was supported by funding from HSE Research Foundation, Marcus Wallenberg's Foundation, Emil Aaltonen Foundation and KAUTE Foundation. We would also like to thank the associate editor and the anonymous reviewers for their insights and advice in the development of this paper.

References

Abbasi, A., Li, J., Adjeroh, D., Abate, M., & Zheng, W. (2019). Don't mention it? Analyzing user-generated content signals for early adverse event warnings. Information Systems Research, 30(3), 1007-1028.

Abbasi, A., Sarker, S., & Chiang, R. H. L. (2016). Big data research in information systems: Toward an inclusive research agenda. Journal of the Association for Information Systems, 17(2), 1–32.

Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. Information Systems Research, 25(3), 443–448.

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2023). Best-practice recommendations for producers, evaluators, and users of methodological literature reviews. Organizational Research Methods, 26(1), 46-76.

Ahmadi, S., Shokouhyar, S., Shahidzadeh, M. H., & Papageorgiou, I. E. (2022). The bright side of consumers' opinions of improving reverse logistics decisions: A social media analytic framework. International Journal of Logistics Research and Applications, 25(6), 977-1010.

Al Shehhi, A., Thomas, J., Welsch, R., Grey, I., & Aung, Z. (2019). Arabia felix 2.0: A cross-linguistic twitter analysis of happiness patterns in the United Arab Emirates. Journal of Big Data, 6(33), 1-20.

Altaweel, M., & Hadjitofi, T. G. (2020). The sale of heritage on eBay: Market trends and cultural value. Big Data & Society, 7(2), 1-17.

Amat-Lefort, N., Barravecchia, F., & Mastrogiacomo, L. (2022). Quality 4.0: Big data analytics to explore service quality attributes and their relation to user sentiment in Airbnb reviews. International Journal of Quality & Reliability Management, 39(6), 1-19.

Angelico, C., Marcucci, J., Miccoli, M., & Quarta, F. (2022). Can we measure inflation expectations using Twitter? Journal of Econometrics, 228(2), 259-277.

Asr, F. T., & Taboada, M. (2019). Big data and quality data for fake news and misinformation detection. Big Data & Society, 6(1).

Bang, C. C., Lee, J., & Rao, H. R. (2021). The Egyptian protest movement in the Twittersphere: An investigation of dual sentiment pathways of communication. International Journal of Information Management, 58(1-13).

Barchiesi, M. A., & Colladon, A. F. (2021). Big data and big values: When companies need to rethink themselves. Journal of Business Research, 129, 714-722.

Becken, S., Alaei, A. R., & Wang, Y. (2020). Benefits and pitfalls of using tweets to assess destination sentiment. Journal of Hospitality and Tourism Technology, 11(1), 19-34.

Benabderrahmane, S., Mellouli, N., Lamolle, M., & Paroubek, P. (2017). Smart4job: A big data framework for intelligent job offers broadcasting using time series forecasting and semantic classification. Big Data Research, 7, 16-30.

Berente, N., & Seidel, S. (2014). Big data and inductive theory development: Towards computational grounded theory? Paper presented at the 20th Americas Conference on Information Systems.

Bhattacharjya, J., Ellison, A., & Tripathi, S. (2016). An exploration of logistics-related customer service provision on Twitter the case of e-retailers. International Journal of Physical Distribution & Logistics Management & Organizational History, 46(6).

Bhattacharjya, J., Ellison, A. B., Pang, V., & Gezdur, A. (2018). Creation of unstructured big data from customer service the case of parcel shipping companies on Twitter. The International Journal of Logistics Management, 29(2), 723-738.

Bokányi, E., Kondor, D., Dobos, L., Sebők, T., Stéger, J., Csabai, I., & Vattay, G. (2016). Race, religion and the city: Twitter word frequency patterns reveal dominant demographic dimensions in the United States. Humanities & Social Science Communications, 2, 1-9.

Boyd, D., & Crawford, K. (2012). Six provocations for big data. SSRN Electronic Journal, 1-17.

Brooker, P., Barnett, J., Vines, J., Lawson, S., Feltwell, T., Long, K., & Wood, G. (2018). Researching with Twitter timeline data: A demonstration via ‘‘everyday’’ sociopolitical talk around welfare provision. Big Data & Society, 5(1).

Cavique, M., Ribeiro, R., Batista, F., & Correia, A. (2022). Examining Airbnb guest satisfaction tendencies: A text mining approach. Current Issues in Tourism, 25(22), 3607-3622.

Cécillon, N., Labatut, V., Dufour, R., & Linarès, G. (2019). Abusive language detection in online conversations by combining content- and graph-based features. Frontiers in Big Data, 2.

Chae, B. K. (2019). A general framework for studying the evolution of the digital innovation ecosystem: The case of big data. International Journal of Information Management, 45, 83-04.

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. Decision Support Systems, 63, 67-80.

Chen, F., & Neill, D. B. (2015). Human rights event detection from heterogeneous social media graphs. Big Data, 3(1), 34-40.

Chen, J., Yang, Y. C., & Liu, H. (2021). Mining bilateral reviews for online transaction prediction: A relational topic modeling approach. Information Systems Research, 32(2), 541-560.

Chen, K., Li, X., Luo, P., & Zhao, J. L. (2021). News-induced dynamic networks for market signaling: Understanding the impact of news on firm equity value. Information Systems Research, 32(2), 356-377.

Chen, L., Baird, A., & Straub, D. (2019). Fostering participant health knowledge and attitudes: An econometric study of a chronic disease-focused online health community. Journal of Management Information Systems, 36(1), 194-229.

Chen, Y., Deng, S., Kwak, D., Elnoshokaty, A., & Wu, J. (2019). A multi-appeal model of persuasion for online petition success: A linguistic cue-based approach. Journal of the Association for Information Systems, 20(2), 105-131.

Chen, Y., Song, Q., Liu, X., Sastry, P. S., Hu, X., & Chen, Y. (2020). On robustness of neural architecture search under label noise. Frontiers in Big Data, 3, 1-9.

Cheng, M., & Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. International Journal of Hospitality Management, 76, 58-70.

Cheung, M., She, J., & Wang, N. (2018). Characterizing user connections in social media through user-shared images. IEEE Transactions on Big Data, 4(4), 447-458.

Chew, A. W. Z., Pan, Y., Wang, Y., & Zhang, L. (2021). Hybrid deep learning of social media big data for predicting the evolution of covid-19 transmission. Knowledge-Based Systems, 233, 1-21.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? IEEE Transactions on Dependable and Secure Computing, 9(6), 811-824.

Chung, S., Chong, M., Chua, J. S., & Na, J. C. (2018). Evolution of corporate reputation during an evolving controversy. Journal of Communication Management & Organizational History, 23(1), 52-72.

Ciasullo, M. V., Troisi, O., Loia, F., & Maione, G. (2018). Carpooling: Travelers’ perceptions from a big data analysis. The TQM Journal, 30(5).

Colladon, A. F., Grippa, F., & Innarellac, R. (2020). Studying the association of online brand importance with museum visitors: An application of the semantic brand score. Tourism Management Perspectives, 33, 1-9.

Colladon, A. F., Guardabascio, B., & Innarellac, R. (2019). Using social network and semantic analysis to analyze online travel forums and forecast tourism demand. Decision Support Systems, 123, 1-11.

Conway, E., Rosati, P., Monks, K., & Lynn, T. (2019). Voicing job satisfaction and dissatisfaction through Twitter: Employees’ use of cyberspace. New Technology, Work and Employment, 34(2), 139-156.

Crawford, K., Gray, M., & Miltner, K. (2014). Critiquing big data: Politics, ethics, epistemology. International Journal of Communication, 8, 1663–1672.

Cury, R. M. (2019). Oscillation of tweet sentiments in the election of João Doria Jr. for mayor. Journal of Big Data, 1-15.

Dalton, C., Taylor, L., & Thatcher, J. (2016). Critical data studies: A dialog on data and space. Big Data & Society, 1-9.

Dalton, C., & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Psychological Bulletin, 126(21).

Daries, J., Reich, J., Waldo, J., Young, E., Whittinghill, J., Ho, A., Seaton D., & Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. Communications of the ACM, 57(9), 56-63.

Davidson, E., Edwards, R., Jamieson, L., & Weller, S. (2019). Big data, qualitative style: A breadth-and-depth method for working with large amounts of secondary qualitative data. Quality and Quantity, 53, 363-376.

Deng, S., Huang, Z., Sinha, A. P., & Zhao, H. (2018). The interaction between microblog sentiment and stock returns: An empirical examination. MIS Quarterly, 42(3), 895-918.

Domalewska, D. (2021). An analysis of covid-19 economic measures and attitudes: Evidence from social media mining. Journal of Big Data, 8(42), 1-14.

Dowling, M., Wycoff, N., Mayer, B., Wenskovitch, J., Leman, S., House, L., Polys, N., North, C. & Hauck, P. (2019). Interactive visual analytics for sensemaking with big text. Big Data Research, 16, 49–58.

Durahim, A. O., & Coşkun, M. (2015). #iamhappybecause: Gross national happiness through Twitter analysis and big data. Technological Forecasting & Social Change, 99, 92-105.

Edwards, D., Cheng, M., Wong, I. A., & Wu, J. Z. a. Q. (2017). Ambassadors of knowledge sharing co-produced travel information through tourist-local social media exchange. International Journal of Contemporary Hospitality Management & Organizational History, 29(2), 690-708.

El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A., & Kobi, A. (2018). A novel adaptable approach for sentiment analysis on big social data. Journal of Big Data, 5(1).

Feizollah, A., Mostafa, M. M., Sulaiman, A., Zakaria, Z., & Firdaus, A. (2021). Exploring halal tourism tweets on social media. Journal of Big Data, 8(72).

Felt, M. (2016). Social media and the social sciences: How researchers employ big data analytics. Big Data and Society, 3(1), 1-15.

García, C. G., & Álvarez-Fernández, E. (2022). What is (not) big data based on its 7vs challenges: A survey. Big Data & Cognitive Computing, 6(4), 1-29.

George, G., Haas, M., & Pentland, A. (2014). From the editors. Big data and management. Academy of Management Journal, 57(2), 321-326.

Georgiadou, E., Angelopoulos, S., & Drake, H. (2020). Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes. International Journal of Information Management, 51, 1-9.

Geva, H., Oestreicher-Singer, G., & Saar-Tsechansky, M. (2019). Using retweets when shaping our online persona: Topic modeling approach. MIS Quarterly, 43(2), 501-524.

Geva, T., Oestreicher-Singer, G., Efron, N., & Shimshoni, Y. (2017). Using forum and search data for sales prediction of high-involvement projects. MIS Quarterly, 41(1), 65-82.

Goes, B. P. (2014). Editor’s comments – big data and is research. MIS Quarterly, 38(3), iii-viii.

Gong, J., Abhishe, V., & Li, B. (2018). Examining the impact of keyword ambiguity on search advertising performance. MIS Quarterly, 42(3), 805-829.

Gray, J., & Suzor, N. (2020). Playing with machines: Using machine learning to understand automated copyright enforcement at scale. Big Data & Society, 7(1).

Grover, V., Lindberg, A., Benbasat, I., & Lyytinen, K. (2020). The perils and promises of big data research in information systems. Journal of the Association for Information Systems, 21, 1-26.

Gruss, R., Kim, E., & Abrahams, A. (2020). Engaging restaurant customers on Facebook: The power of belongingness appeals on social media. Journal of Hospitality & Tourism Research, 44(2), 201-228.

Guindy, M. A. (2022). Fear and hope in financial social networks: Evidence from covid-19. Finance Research Letters, 46, 1-9.

Gunarathne, P., Rui, H., & Seidmann, A. (2018). When social media delivers customer service: Differential customer treatment in the airline industry. MIS Quarterly, 42(2), 489-520.

Guo, X., Wei, Q., Chen, G., Zhang, J., & Qiao, D. (2017). Extracting representative information on intra-organizational blogging platforms. MIS Quarterly, 41(4), 1105-1127.

Han, C., Yang, M., & Piterou, A. (2021). Do news media and citizens have the same agenda on covid-19? An empirical comparison of Twitter posts. Technological Forecasting & Social Change, 169, 1-10.

Hannigan, T. R., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M. S., Kaplan, S. & Jennings, P. D. (2019). Topic modeling in management research: Rendering new theory from textual data. Academy of Management Annals, 13(2), 576-632.

Harbert, T. (2021). Tapping the power of unstructured data. MIT Sloan School of Management, Ideas made to matter. https://mitsloan.mit.edu/ideas-made-to-matter/tapping-power-unstructured-data

Harrigan, P., Daly, T. M., Coussement, K., Lee, J. A., Soutar, G. N., & Evers, U. (2021). Identifying influencers on social media. International Journal of Information Management, 26, 1-11.

Hashemi, M., & Hall, M. (2020). Criminal tendency detection from facial images and the gender bias effect. Journal of Big Data, 7(2).

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. Organizational Research Methods, 25(1).

Hippel, C. D. v., & Cann, A. B. (2021). Behavioral innovation: Pilot study and new big data analysis approach in household sector user innovation. Research Policy, 50, 1-14.

Hirschheim, R. (2021). The attack on understanding: How big data and theory have led us astray: A comment on Gary Smith’s data mining fool’s gold. Journal of Information Technology, 36(2), 176-183.

Ho, S. M., & Li, W. (2022). “I know you are, but what am I?” Profiling cyberbullying based on charged language. Computational and Mathematical Organization Theory, 28, 293-320.

Ho, S. Y., Choi, K. W. S., & Yang, F. F. (2019). Harnessing aspect-based sentiment analysis: How are tweets associated with forecast accuracy? Journal of the Association for Information Systems, 20(8), 1174-1209.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. Journal of the Association for Information Systems, 12(12), 767-797.

Hu, L., Li, Z., & Ye, X. (2020). Delineating and modeling activity space using geotagged social media data. Cartography and Geographic Information Science, 47(3), 277-288.

Hu, T., Bigelow, E., Luo, J., & Kautz, H. (2017). Tales of two cities: Using social media to understand idiosyncratic lifestyles in distinctive metropolitan areas. IEEE Transactions on Big Data, 3(1), 55-66.

Huang, C.-Y., Tong, H., He, J., & Maciejewski, R. (2019). Location prediction for tweets. Frontiers in Big Data, 2.

Huang, J., Boh, W. F., & Goh, K. H. (2017). A temporal study of the effects of online opinions: Information sources matter. Journal of Management Information Systems, 34(4), 1169-1202.

Huberty, M. (2015). Can we vote with our tweet? On the perennial difficulty of election forecasting with social media. International Journal of Forecasting, 31, 992-1007.

Hwang, E. H., Singh, P. V., & Argote, L. (2019). Jack of all, master of some: Information network and innovation in crowdsourcing communities. Information Systems Research, 30(2), 389-410.

IDC. (2014, April 2014). The digital universe of opportunities: Rich data and the increasing value of the internet of things. EMC Digital Universe with Research & Analysis by IDC. https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.html

Jones, M. (2019). What we talk about when we talk about (big) data. Journal of Strategic Information Systems, 28(1), 3-16.

Janasik, N., Honkela, T., & Bruun, H. (2009). Text mining in qualitative research application of an unsupervised learning method. Organizational Research Methods, 12(3), 436-460

Jung, A.-K., Ross, B., & Stieglitz, S. (2020). Caution: Rumors ahead—A case study on the debunking of false information on Twitter. Big Data & Society, 7(2), 1-15.

Karamshuk, D., Shaw, F., Brownlie, J., & Sastry, N. (2017). Bridging big data and qualitative methods in the social sciences: A case study of Twitter responses to high profile deaths by suicide. Online Social Networks and Media, 33-43.

Kelling, S., Hochachka, W. M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., & Hooker, G. (2009). Data-intensive science: A new paradigm for biodiversity studies. BioScience, 59(7), 613–620.

Khazraee, E. (2019). Mapping the political landscape of Persian Twitter: The case of 2013 presidential election. Big Data & Society, 6(1), 1-15.

Khurana, S., Qiu, L., & Kumar, S. (2019). When a doctor knows, it shows: An empirical analysis of doctors' responses in a Q&A forum of an online healthcare portal. Information Systems Research, 30(3), 872-891.

Kim, W.-H., Park, E., & Kim, S.-B. (2023). Understanding the role of firm-generated content by hotel segment: The case of Twitter. Current Issues in Tourism, 26(1), 122-136

King, K. K., & Wang, B. (2023). Diffusion of real versus misinformation during a crisis event: A big data-driven approach. International Journal of Information Management, 71, 1-14.

King, T. (2019, 28.3.2019). 80 percent of your data will be unstructured in five years. Data Management Solutions Review. https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. Big Data and Society, 1(1).

Kitchin, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. Big Data and Society, 3(1).

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Hartog, D. N. D. (2018). Text mining in organizational research. Organizational Research Methods, 21(3), 733-756.

Kokkodis, M., & Lappas, T. (2020). Your hometown matters: Popularity-difference bias in online reputation platforms. Information Systems Research, 31(2), 412-430.

Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. Journal of Big Data, 6(62), 1-16.

Kwok, L., Tang, Y., & Yuc, B. (2020). The 7 ps marketing mix of home-sharing services: Mining travelers' online reviews on Airbnb. International Journal of Hospitality Management, 90, 1-11.

Kwon, W., Lee, M., Back, K.-J., & Lee, K. Y. (2020). Assessing restaurant review helpfulness through big data: Dual-process and social influence theory. Journal of Hospitality and Tourism Technology, 12(2), 177-195.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. Science, 343(6176), 1203–1205.

Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. Annual Review of Sociology, 43(1), 19–39.

LeBaron, C., Jarzabkowski, P., Pratt, M. G., & Fetzer, G. (2018). An introduction to video methods in organizational research. Organizational Research Methods, 21(239-260).

Lee, M., Kwon, W., & Back, K.-J. (2021). Artificial intelligence for hospitality big data analytics: Developing a prediction model of restaurant review helpfulness for customer decision-making. International Journal of Contemporary Hospitality Management, 33(6), 2117-2136.

Lee, P.-S., West, J. D., & Howe, B. (2018). Viziometrics: Analyzing visual information in the scientific literature. IEEE Transactions on Big Data, 4(1), 117-129.

Lee, S.-Y., Rui, H., & Whinston, A. B. (2019). Is best answer really the best answer? The politeness bias. MIS Quarterly, 43(2), 579-600.

Li, C., Ye, Q., Nicolau, J. L., & Liu, X. (2021). Smiley guests post long reviews! International Journal of Hospitality Management, 96, 1-5.

Li, T., Dalen, J. v., & Rees, P. J. v. (2018). More than just noise? The information content of stock microblogs on financial markets. Journal of Information Technology, 33, 50–69.

Li, Y., Liu, C., Du, N., Fan, W., Li, Q., Gao, J., Zhang, C. & Wu, H. (2020). Extracting medical knowledge from crowdsourced question answering website. IEEE Transactions on Big Data, 6(2), 309-321.

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Too big to fail: Large samples and the p-value problem. Information Systems Research, 24(4), 906-917.

Lindberg, A. (2020). Developing theory through integrating human and machine pattern recognition. Journal of the Association for Information Systems, 21(1), 90-116.

Liu, Q., Du, Q., Hong, Y., Fan, W., & Wu, S. (2020). User idea implementation in open innovation communities: Evidence from a new product development crowdsourcing community. Information Systems Journal, 30(5), 899-927.

Liu, X. (2019). A big data approach to examining social bots on Twitter. Journal of Services Marketing, 33(4), 369-379.

Liu, X. (2020). Analyzing the impact of user-generated content on b2b firms' stock performance: Big data analysis with machine learning methods. Industrial Marketing Management, 86, 30-39.

Liu, X., Burns, A. C., & Hou, Y. (2017). An investigation of brand-related user-generated content on Twitter. Journal of Advertising, 46(2), 236-247.

Liu, X., Shin, H., & Burns, A. C. (2021). Examining the impact of luxury brand's social media marketing on customer engagement: Using big data analytics and natural language processing. Journal of Business Research, 125, 815-826.

Liu, X., Singh, P. V., & Srinivasan, K. (2016). A structured analysis of unstructured big data by leveraging cloud computing. Marketing Science, 35(3), 363–388.

Liu, X., Wang, G. A., Fan, W., & Zhang, Z. (2020). Finding useful solutions in online knowledge communities: A theory-driven design and multilevel analysis. Information Systems Research, 31(3), 731-752.

Liu, Z., Shan, J., Balet, N. G., & Fang, G. (2017). Semantic social media analysis of Chinese tourists in Switzerland. Information Technology & Tourism, 17, 183-202.

London Jr, J., & Matthews, K. (2022). Crisis communication on social media - lessons from covid-19. Journal of Decision Systems, 31(1-2), 150-170.

Lopez, K., Fodeh, S. J., Allam, A., Brandt, C. A., & Krauthammer, M. (2020). Reducing annotation burden through multimodal learning. Frontiers in Big Data, 3.

Luo, X., Gu, B., Zhang, J., & Phang, C. W. (2017). Expert blogs and consumer perceptions of competing brands. MIS Quarterly, 41(2), 371-395.

Ma, J., Tseb, Y. K., Wang, X., & Zhang, M. (2019). Examining customer perception and behaviour through social media research – an empirical study of the united airlines overbooking crisis. Transportation Research Part E, 127, 192-205.

Mahdikhani, M. (2022). Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic. International Journal of Information Management Data Insights, 2(1-14).

Mai, F., Shan, Z., Bai, Q., Wang, X., & Chiang, R. H. L. (2018). How does social media impact Bitcoin value? A test of the silent majority hypothesis. Journal of Management Information Systems, 35(1), 19-52.

Mallipeddi, R. R., Janakiraman, R., Kumar, S., & Gupta, S. (2021). The effects of social media content created by human brands on engagement: Evidence from Indian general election 2014. Information Systems Research, 32(1), 212-237.

Marine-Roig, E., & Clavé, S. A. (2015). Tourism analytics with massive user-generated content: A case study of Barcelona. Journal of Destination Marketing & Management, 4, 162-172.

Marine-Roig, E., & Huertas, A. (2020). How safety affects destination image projected through online travel reviews. Journal of Destination Marketing & Management, 18.

Martin-Fuentes, E., Fernandez, C., Mateu, C., & Marine-Roig, E. (2018). Modelling a grading scheme for peer-to-peer accommodation: Stars for Airbnb. International Journal of Hospitality Management, 69, 75-83.

McKenna, B. (2019). Creating convivial affordances: A study of virtual world social movements. Information Systems Journal, 30(1), 185-214.

Mejia, J., Mankad, S., & Gopal, A. (2019). A for effort? Using the crowd to identify moral hazard in New York city restaurant hygiene inspections. Information Systems Research, 30(4), 1363-1386

Meng, Y., Huang, J., Wang, G., Wang, Z., Zhang, C., & Han, J. (2020). Unsupervised word embedding learning by incorporating local and global contexts. Frontiers in Big Data, 3(9), 1-12.

Mills, K. A. (2018). What are the threats and potentials of big data for qualitative research? Qualitative Research, 18(6), 591-603.

Mishra, N., Singh, A., Rana, N. P., & Dwivedi, Y. K. (2017). Interpretive structural modelling and fuzzy micmac approaches for customer centric beef supply chain: Application of a big data technique. Production Planning & Control, 28(11-12), 945-963.

Montejo-Ráez, A., Díaz-Galiano, M. C., Martínez-Santiago, F., & Ureña-López, L. A. (2014). Crowd explicit sentiment analysis. Knowledge-Based Systems, 69, 134–139.

Mousavi, R., Johar, M., & Mookerjee, V. S. (2020). The voice of the customer: Managing customer care in Twitter. Information Systems Research, 31(2), 340-360.

Mukherjee, S., Kumar, R., & Bala, P. K. (2022). Managing a natural disaster: Actionable insights from microblog data. Journal of Decision Systems, 31(1-2), 134-149.

Neu, D., Saxton, G., Rahaman, A., & Everett, J. (2019). Twitter and social accountability: Reactions to the Panama papers. Critical Perspectives on Accounting, 61, 38-53.

Ngoc, T. N. T., Thu, H. N. T., & Nguyen, V. A. (2019). Mining aspects of customer’s review on the social network. Journal of Big Data, 6(22), 1-21.

Nguyen, T., Larsen, M. E., O’Deb, B., Nguyen, D. T., Yearwood, J., Phung, D., Venkatesh, S. & Christensen, H. (2017). Kernel-based features for predicting population health indices from geocoded social media data. Decision Support Systems, 102, 22-31.

Nian, T., Hu, Y., & Chen, C. (2021). Examining the impact of television-program-induced emotions on online word-of-mouth toward television advertising. Information Systems Research, 32(2), 605-632.

Obschonka, M., Lee, N., Rodríguez-Pose, A., Eichstaedt, J. C., & Ebert, T. (2020). Big data methods, social media, and the psychology of entrepreneurial regions: Capturing cross-county personality traits and their impact on entrepreneurship in the USA. Small Business Economics, 5(5), 567–588.

Oh, Y.-K. (2020). Determinants of online review helpfulness for Korean skincare products in online retailing. Journal of Distribution Science, 18(10), 65-75.

Oh, Y. K., & Yi, J. (2021). Asymmetric effect of feature level sentiment on product rating: An application of bigram natural language processing (NLP) analysis. Internet Research, 32(3), 1023-1040.

Ordenes, F. V., Ludwig, S., Ruyter, K. D., Grewal, D., & Wetzels, M. (2017). Unveiling what is written in the stars: Analyzing explicit, implicit, and discourse patterns of sentiment in social media. Journal of Consumer Research, 43, 875-894.

Østerlund, C., Crowston, K., & Jackson, C. (2020). Building an apparatus: Refractive, reflective, and diffractive readings of trace data. Journal of the Association for Information Systems, 20(1), 1-22.

Park, S. B., Jang, J., & Ok, C. M. (2016). Analyzing Twitter to explore perceptions of Asian restaurants. Journal of Hospitality and Tourism Technology, 7(4), 405-422.

Park, Y.-E., & Javed, Y. (2020). Insights discovery through hidden sentiment in big data: Evidence from Saudi Arabia's financial sector. Journal of Asian Finance, Economics and Business, 7(6), 457-464.

Pons, A., Vintrò, C., Rius, J., & Vilaplana, J. (2021). Impact of corporate social responsibility in mining industries. Resources Policy, 72.

Pournarakis, D. E., Sotiropoulos, D. N., & Giaglis, G. M. (2017). A computational model for mining consumer perceptions in social media. Decision Support Systems, 93, 98–110.

Qiao, D., Lee, S.-Y., Whinston, A. B., & Wei, Q. (2020). Financial incentives dampen altruism in online prosocial contributions: A study of online reviews. Information Systems Research, 31(3), 1361-1375.

Resce, G., & Maynard, D. (2018). What matters most to people around the world? Retrieving better life index priorities on Twitter. Technological Forecasting & Social Change, 137, 61-75.

Rill, S., Reinel, D., Scheidt, J., & Zicari, R. V. (2014). Politwi: Early detection of emerging political topics on Twitter and the impact on concept-level sentiment analysis. Knowledge-Based Systems, 69, 24-33.

Rintyarna, B. S., Sarno, R., & Fatichah, C. (2019). Evaluating the performance of sentence level features and domain sensitive features of product reviews on supervised sentiment analysis tasks. Journal of Big Data, 6(84), 1-19.

Rossi, R., Nairn, A., Smith, J., & Inskip, C. (2021). "Get a £10 free bet every week!"—Gambling advertising on Twitter: Volume, content, followers, engagement, and regulatory compliance. Journal of Public Policy & Marketing, 40(4), 487-504.

Samtani, S., Chinn, R., Chen, H., & Jr., J. F. N. (2017). Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence. Journal of Management Information Systems, 34(4), 1023-1053.

Sanford, A. (2022). Does perception matter in asset pricing? Modeling volatility jumps using Twitter-based sentiment indices. Journal of Behavioral Finance, 23(3), 262–280.

Sarin, P., Kar, A. K., & Ilavarasan, V. P. (2021). Exploring engagement among mobile app developers – insights from mining big data in user generated content. Journal of Advances in Management Research, 18(4), 585-608.

Saxton, G. D., & Neu, D. (2022). Twitter-based social accountability processes: The roles for financial inscriptions-based and values-based messaging. Journal of Business Ethics, 181, 1041-1064.

Schlosser, S., Toninelli, D., & Cameletti, M. (2021). Comparing methods to collect and geolocate tweets in Great Britain. Journal of Open Innovation: Technology, Market, and Complexity, 7(1), 1-20.

Schmiedel, T., Müller, O., & Brocke, J. v. (2019). Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture. Organizational Research Methods, 22(4), 941-968.

Schneider, M. J., & Guptab, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. International Journal of Forecasting, 32, 243-256.

See-To, E. W. K., & Yang, Y. (2017). Market sentiment dispersion and its effects on stock return and volatility. Electronic Markets, 27, 283–296.

Serrano, L., Ariza-Montes, A., Nader, M., Sianes, A., & Law, R. (2021). Exploring preferences and sustainable attitudes of Airbnb green users in the review comments and ratings: A text mining approach. Journal of Sustainable Tourism, 29(7), 1134-1152.

She, C., & Michelon, G. (2019). Managing stakeholder perceptions: Organized hypocrisy in CSR disclosures on Facebook. Critical Perspectives on Accounting, 61, 54-76.

Shi, D., Guan, J., Zurada, J., & Manikas, A. (2017). A data-mining approach to identification of risk factors in safety management systems. Journal of Management Information Systems, 34(4), 1054-1081.

Shokouhyar, S., Dehkhodaei, A., & Amiri, B. (2021). A mixed-method approach for modelling customer-centric mobile phone reverse logistics: Application of social media data. Journal of Modelling in Management & Organizational History, 17(2).

Singh, A., & Glińska-Neweś, A. (2022). Modeling the public attitude towards organic foods: A big data and text mining approach. Journal of Big Data, 9(2), 1-21.

Singh, A., Shukla, N., & Mishra, N. (2018). Social media data analytics to improve supply chain management in food industries. Transportation Research Part E, 114, 398-415.

Smith, G. (2020). Data mining fool’s gold. Journal of Information Technology, 35(3), 182-194.

Song, L., Lau, R. Y. K., Kwok, R. C.-W., Mirkovski, K., & Dou, W. (2017). Who are the spoilers in social media marketing? Incremental learning of latent semantics for social spam detection. Electronic Commerce Research, 17, 51-81.

Subroto, A., & Apriyana, A. (2019). Cyber risk prediction through social media big data analytics and statistical machine learning. Journal of Big Data, 6(50), 1-19.

Subroto, A., & Christianis, M. (2021). Rating prediction of peer-to-peer accommodation through attributes and topics from customer review. Journal of Big Data, 8(9), 1-29.

Sun, Y., Zeng, X., Zhou, S., Zhao, H., Thomas, P., & Hu, H. (2021). What investors say is what the market says: Measuring China’s real investor sentiment. Personal and Ubiquitous Computing, 25, 587–599.

Sundararaman, A., Ramanathan, S. V., & Thati, R. (2018). Novel approach to predict hospital readmissions using feature selection from unstructured data with class imbalance. Big Data Research, 13, 65-75.

Tang, L., Griffith, L., Stevens, M., & Hardie, M. (2020). Social media analytics in the construction industry comparison study between China and the United States. Engineering, Construction and Architectural Management, 27(8), 1877-1889.

Tao, J., & Fang, X. (2020). Toward multi-label sentiment analysis: A transfer learning based approach. Journal of Big Data, 7(1), 1-26.

Thatcher, J. (2014). Living on fumes: Digital footprints, data fumes, and the limitations of spatial big data. International Journal of Communication, 8, 1765–1783.

Toshniwal, D., Somani, S., Aggarwal, R., & Malik, P. (2019). Global awareness landscape for ailments—a Twitter based microscopic view into thought processes of people. Frontiers in Big Data, 2(18).

Triantafyllidou, D., Nousi, P., & Tefas, A. (2017). Fast deep convolutional face detection in the wild exploiting hard sample mining. Big Data Research, 11, 65-76.

Vaast, E., Safadi, H., Lapointe, L., & Negoita, B. (2017). Social media affordances for connective action: An examination of microblogging use during the Gulf of Mexico oil spill. MIS Quarterly, 41(4), 1179–1206.

Villegas, C. A., & Martinez, M. J. (2022). Lessons from Harvey: Improving traditional damage estimates with social media sourced damage estimates. Cities, 121, 1-13.

Walker, M. A., & Boamah, E. F. (2020). The digital life of the #migrantcaravan: Contextualizing Twitter as a spatial technology. Big Data & Society, 7(2), 1-18.

Walsh, I., Holton, J. A., Bailyn, L., Fernandez, W., Levina, N., & Glaser, B. (2015). What grounded theory is…a critically reflective conversation among scholars. Organizational Research Methods, 18(4), 581–599.

Wan, Z., & He, H. (2019). Answernet: Learning to answer questions. IEEE Transactions on Big Data, 5(4), 540-549.

Wang, A. H. (2010, June 21-23, 2010). Detecting spam bots in online social networking sites: A machine learning approach. Paper presented at the Data and Applications Security and Privacy XXIV.

Wang, R., Browning, M. H. E. M., Qin, X., He, J., Wu, W., Yao, Y., & Liu, Y. (2022). Visible green space predicts emotion: Evidence from social media and street view data. Applied Geography, 148, 1-9.

Wang, R., Hao, J.-X., Law, R., & Wang, J. (2019). Examining destination images from travel blogs: A big data analytical approach using Latent Dirichlet Allocation. Asia Pacific Journal of Tourism Research, 24(11).

Wang, Y.-Y., Guo, C., Susarla, A., & Sambamurthy, V. (2021). Online to offline: The impact of social media on offline sales in the automobile industry. Information Systems Research, 32(2), 582-604.

Wang, Z., Jiang, C., Zhao, H., & Ding, Y. (2020). Mining semantic soft factors for credit risk evaluation in peer-to-peer lending. Journal of Management Information Systems, 37(1), 282-308.

Webb, H., Jirotka, M., Stahl, B. C., Housley, W., Edwards, A., Williams, M., Procter, R., Rana, O., & Burnap, P. (2017). The ethical challenges of publishing Twitter data for research dissemination. Paper presented at the WebSci 2017.

Winkler, M., Abrahams, A. S., Gruss, R., & Ehsani, J. P. (2016). Toy safety surveillance from online reviews. Decision Support Systems, 90, 23-32.

Wu, L., Li, J., & Qi, J. (2019). Characterizing popularity dynamics of hot topics using micro-blogs spatio-temporal data. Journal of Big Data, 6(101), 1-16.

Xi, Y., Ma, C., Yang, Q., & Jiang, Y. (2022). A cross-cultural analysis of tourists' perceptions of Airbnb attributes. International Journal of Hospitality and Tourism Administration, 23(4), 754-787.

Xie, P., Chen, H., & Hu, Y. J. (2020). Signal or noise in social media discussions: The role of network cohesion in predicting the Bitcoin market. Journal of Management Information Systems, 37(4), 933-956.

Xing, Y., Wang, X., Qiu, C., Li, Y., & He, W. (2022). Research on opinion polarization by big data analytics capabilities in online social networks. Technology in Society, 68, 1-12.

Xu, Q., Li, J., Cai, M., & Mackey, T. K. (2019). Use of machine learning to detect wildlife product promotion and sales on Twitter. Frontiers in Big Data, 2.

Xu, S., Yin, B., & Lou, C. (2022). Minority shareholder activism and corporate social responsibility. Economic Modelling, 116, 1-14.

Yang, K.-C., Pierri, F., Hui, P.-M., Axelrod, D., Torres-Lugo, C., Bryden, J., & Menczer, F. (2021). The covid-19 infodemic: Twitter versus Facebook. Big Data & Society, 8(1), 1-16.

Yang, X., Zhu, Y., & Cheng, T. Y. (2020). How the individual investors took on big data: The effect of panic from the internet stock message boards on stock price crash. Pacific-Basin Finance Journal, 59, 1-11.

Yi, J., & Oh, Y. K. (2022). Does brand type affect what consumers discuss? A comparison of attribute-based reviews of value and premium brands of an innovative product. Internet Research, 32(2), 606-619.

Yin, H., Wu, X., & Kong, S. X. (2020). Daily investor sentiment, order flow imbalance and stock liquidity: Evidence from the Chinese stock market. International Journal of Finance & Economics, 27(4), 7816-4836.

Yoon, S., Kuang, D., Broadwell, P., Lee, H., & Odlum, M. (2017). What can we learn about the Middle East Respiratory Syndrome (MERS) outbreak from tweets? Big Data & Information Analytics, 2(3), 203-207.

Yu, W., Li, J., Bhuiyan, M. Z. A., Zhang, R., & Huai, J. (2019). Ring: Real-time emerging anomaly monitoring system over text streams. IEEE Transactions on Big Data, 5(4), 506-519.

Yue, W. T., Wang, Q.-H., & Hui, K.-L. (2019). See no evil, hear no evil? Dissecting the impact of online hacker forums. MIS Quarterly, 43(1), 73-95.

Zhang, D., Wang, D., Vance, N., Zhang, Y., & Mike, S. (2019). On scalable and robust truth discovery in big data social media sensing applications. IEEE Transactions on Big Data, 5(2), 195–208.

Zhang, J.-B., Sun, Y.-X., & Zhan, D.-C. (2017). Multiple-instance learning for text categorization based on semantic representation. Big Data & Information Analytics, 2(1), 69–75.

Zhang, L., & Malife, C. (2021). Processing billions of events in real time at Twitter. https://blog.twitter.com/engineering/en_us/topics/infrastructure/2021/processing-billions-of-events-in-real-time-at-twitter-

Zhang, N., Liu, R., Zhang, X.-Y., & Pang, Z.-L. (2021). The impact of consumer perceived value on repeat purchase intention based on online reviews: By the method of text mining. Data Science and Management, 3, 22-32.

Zhang, W., & Ram, A. (2020). A comprehensive analysis of triggers and risk factors for asthma based on machine learning and large heterogeneous data sources. MIS Quarterly, 44(1), 305-349.

Zhang, Y., Ridings, C., & Semenov, A. (2023). What to post? Understanding engagement cultivation in microblogging with big data-driven theory building. International Journal of Information Management, 71, 1-19.

Zhang, Z., Barbary, K., Frank Austin Nothaft, Sparks, E. R., Zahn, O., Franklin, M. J., Patterson, D. A. & Perlmutter, S. (2020). Kira: Processing astronomy imagery using big data technology. IEEE Transactions on Big Data, 6(2), 369-381.

Zhou, F., Lim, M. K., He, Y., & Pratap, S. (2019). What attracts vehicle consumers' buying. A Saaty scale-based VIKOR (SSC-VIKOR) approach from after-sales textual perspective? Industrial Management & Data Systems, 120(1), 57-78.

Zhou, S., Qiao, Z., Du, Q., Wang, G. A., Fan, W., & Yan, X. (2018). Measuring customer agility from online reviews using big data text analytics. Journal of Management Information Systems, 35(2), 510-539.

Appendix A: Reviewed Articles

This table presents the reviewed articles with their data sources as well as their initial and final sample sizes. The initial sample size is the amount of data immediately after data collection, while the final sample size is the amount of data after data preparation and the location of relevant data. In cases where the study does not take a smaller sample of relevant data but can use all the initial data, or where we are unsure of how much data the final sample contains after data preparation and other operations, the cell for the final sample size is left empty.

Table 1A. Reviewed Articles' Data Sources and Amount of Data

#	Article	Source	Initial sample size	Final sample size
1	Abbasi et al., 2019	Twitter, forums, search query logs, FAERS	12 000 000 tweets, 5 000 000 posts, "Millions of searches", 6 million reports	
2	Ahmadi et al., 2022	Twitter	74 287 035 tweets	37 878 062 tweets
3	Akshit Singh et al., 2018	Twitter	1 338 638 tweets	23 422 tweets
4	Al Shehhi et al., 2019	Twitter	> 17 000 000 tweets	
5	Altaweel & Hadjitofi, 2020	eBay	108 559 sold items	
6	Amat-Lefort et al., 2022	Airbnb	2 735 437 reviews	
7	Angelico et al., 2022	Twitter	11 100 000 tweets	1 534 743 tweets
8	Anupam Singh & Glińska-Neweś, 2022	Twitter	43 724 tweets	
9	Asr & Taboada, 2019	Various fake news datasets	200 648 news articles and short statements	
10	Bang et al., 2021	Twitter	187 131 tweets	10 716 tweets
11	Barchiesi & Colladon, 2021	Twitter	> 94 000 tweets	
12	Becken et al., 2020	Twitter	198 324 tweets	
13	Benabderrahmane et al., 2017	Job boards	>3 000 000 job offers	
14	Bhattacharjya et al., 2016	Twitter	203 349 tweets	16 998 tweets
15	Bhattacharjya et al., 2018	Twitter	706 582 tweets	
16	Bokányi et al., 2016	Twitter	335 000 000 tweets	
17	Brooker et al., 2018	Twitter	1 398 948 tweets	53 990 tweets
18	C. Li et al., 2021	Twitter	98 323 reviews	
19	C.-Y. Huang et al., 2019	Twitter	> 8 000 000 tweets	
20	Cavique et al., 2022	Airbnb	590 070 reviews	
21	Cécillon et al., 2019	SpaceOrigin (game)	4 029 343 in-game chat messages	
22	Chae, 2019	Twitter	1 671 657 tweets	
23	Cheng & Jin, 2019	insideairbnb.com	181 263 reviews	170 124 reviews
24	Cheung et al., 2018	Flickr, Twitter, Tencent, Weibo, Skyrock, 163 Weibo, Pinterest, Digu and Duitang	2 275 412 images	
25	Chew et al., 2021	Twitter, Kaggle	> 100 000 000 tweets	12 466 981 tweets
26	Chung et al., 2018	Twitter	2 612 018 tweets	
27	Ciasullo et al., 2018	Twitter	993 778 tweets	
28	Colladon et al., 2020	Forum	>2 830 000 posts	
29	Colladon et al., 2019	Tripadvisor	>2 667 301 posts	

30	Conway et al., 2019	Twitter	2 121 139 tweets	817 235 tweets
31	Cury, 2019	Twitter	77 179 tweets	76 690 tweets
32	D. Zhang et al., 2019	Twitter	895 628 tweets	
33	Deng et al., 2018	StockTwits.com, Reuters.com	17 835 174 messages, 3 257 797 news articles	
34	Domalewska, 2021	Twitter, Facebook	109 022 tweets, 557 473 posts	
35	Dowling et al., 2019	LexisNexis	>30 000 news articles	
36	Durahim & Coşkun, 2015	Twitter	> 35 000 000 tweets	
37	Edwards et al., 2017	Tripadvisor	115 847 threads with 8 346 conversations	
38	El Alaoui et al., 2018	Twitter	3 720 000 tweets	
39	F. Chen & Neill, 2015	Twitter	96 000 000 tweets	
40	F. Zhou et al., 2019	Car forum	160 000 posts	
41	Feizollah et al., 2021	Twitter	85 259 tweets	33 880 tweets
42	Georgiadou et al., 2020	Twitter	13 018 367 tweets	
43	Gong et al., 2018	Search engine	"close to 8 million" search impressions	10 000 impressions
44	Gray & Suzor, 2020	YouTube	76 661 274 videos' metadata	12 943 693 videos' metadata
45	Gruss et al., 2020	Facebook	174 706 posts	
46	Guindy, 2022	Twitter	~ 45 000 000 tweets	
47	Gunarathne et al., 2018	Twitter	>3 000 000 tweets	173 662 tweets
48	Guo et al., 2017	Internal blogging system	5000 articles and 10 000 comments produced every day in the system	
49	H. Geva et al., 2019	Twitter	3 388 core users + 2 million followers 464 expert users + 700 000 followers	
50	Han et al., 2021	Twitter	129 965 tweets	34 352 tweets
51	Harrigan et al., 2021	Twitter	556 150 tweets	
52	Hashemi & Hall, 2020	Face Recognition Database, FEI Face Database, Georgia Tech face database, Face Place, Face Detection Data Set, and Benchmark Home	39 713 RGB facial images	
53	Hippel & Cann, 2021	Reddit	3 090 000 comments	178 000 comments
54	S. Y. Ho et al., 2019	Twitter	245 495 tweets	
55	Huberty, 2015	Twitter	1 560 000 tweets	
56	Hwang et al., 2019	Crowdsourced customer support community	1 869 951 posts	
57	J. Chen et al., 2021	Airbnb	3 195 244 reviews	
58	J. Huang et al., 2017	+1 500 websites, comments from 4 movie aggregator sites, discussion forums, Twitter	10 gigabytes of relevant plain text data	
59	J.-B. Zhang et al., 2017	SougouC corpus 20newsgroup dataset	40 000 documents 18 828 documents	

60	Jung et al., 2020	Twitter	736 577 tweets	6 095 tweets
61	K. Chen et al., 2021	219 news portals, 4 discussion boards, 7 blog websites	567 352 news articles	211 297 news articles
62	King & Wang, 2023	Twitter	> 42 000 000 tweets	3 589 tweets
63	Khazraee, 2019	Twitter	3 006 528 tweets	
64	Khurana et al., 2019	Practo.com	131 201 observations	
65	Kim et al., 2023	Twitter	175 358 tweets	142 075 tweets
66	Kokkodis & Lappas, 2020	Online reputation platform	763 658 reviews	
67	Kumar & Zymbler, 2019	Twitter	146 731 tweets	120 766 tweets
68	Kwok et al., 2020	Airbnb	1 148 062 reviews	
69	Kwon et al., 2020	Yelp	4 177 377 reviews	
70	L. Chen et al., 2019	Online health community	867 799 posts	
71	L. Hu et al., 2020	Twitter	> 20 000 000 tweets	
72	London Jr & Matthews, 2022	Twitter	91 658 tweets	33 805 tweets
73	Lopez et al., 2020	PadChest dataset, Indiana Chest X-ray dataset	>167 470 images	
74	Luo et al., 2017	Expert blogs	131 759 blog posts	
75	M. Lee et al., 2021	Yelp	> 4 000 000 reviews	1 483 858 reviews
76	Ma et al., 2019	Twitter	55 083 tweets	
77	Mahdikhani, 2022	Twitter	1 251 216 tweets	
78	Mai et al., 2018	Bitcointalk.org, Twitter	343 769 posts, 3 348 965 tweets	
79	Mallipeddi et al., 2021	Twitter	64 783 tweets	
80	Marine-Roig & Clavé, 2015	Travel blogs, online travel reviews	117 487 reviews	
81	Marine-Roig & Huertas, 2020	Airbnb, Wikimedia	152 702 reviews, 13 546 751 abstracts	
82	Martin-Fuentes et al., 2018	Booking.com	18 000 000 reviews	
83	McKenna, 2019	Discussion forum	128 773 posts	
84	Mejia et al., 2019	Yelp New York City Open Data program	Approx. 1 300 000 reviews 24 625 restaurants' inspection data	
85	Meng et al., 2020	20Newsgroup, Reuters-21578	41 578 news articles	
86	Mishra et al., 2017	Twitter	1 338 637 tweets	26 269 tweets
87	Montejo-Ráez et al., 2014	Twitter	1 863 758 tweets	1 516 184 tweets
88	Mousavi et al., 2020	Twitter	612 900 tweets	
89	Mukherjee et al., 2022	Twitter	42 000 tweets	
90	N. Zhang et al., 2021	Twitter	43 752 reviews	
91	Neu et al., 2019	Twitter	113 882 tweets	
92	Ngoc et al., 2019	Tripadvisor.com, beer review, Trung Nguyen's coffee review	193 661 hotel reviews, 50 000 beer reviews, 1200 coffee reviews	

93	Nguyen et al., 2017	Twitter	1 961 536 285 tweets	768 791 808 tweets
94	Nian et al., 2021	Twitter	>1 200 000 tweets	
95	Obschonka et al., 2020	Twitter	1,5 billion tweets	
96	Ordenes et al., 2017	Twitter	45 842 tweets	43 687 tweets
97	P.-S. Lee et al., 2018	PubMed.org	10 233 004 figures	6 897 810 figures
98	Pons et al., 2021	Twitter	2 000 000 tweets	
99	Pournarakis et al., 2017	Twitter	> 280 000 tweets	221 958 tweets
100	Q. Liu et al., 2020	MIUI new product development community	110 383 product ideas' feedback valences	43 550 product idea’s feedback valences
101	Q. Xu et al., 2019	Twitter	138 357 tweets	
102	Qiao et al., 2020	Amazon reviews dataset	514 554 reviews	
103	Resce & Maynard, 2018	Twitter	7 905 317 tweets	
104	Rill et al., 2014	Twitter	4 000 000 tweets	
105	Rintyarna et al., 2019	Amazon product data	142 800 000 product reviews	
106	Rossi et al., 2021	Twitter	888 745 tweets	
107	Rui Wang et al., 2019	Travel blogs	140 286 posts	
108	Ruoyu Wang et al., 2022	Weibo, Tencent Map	158 108 posts, 202 542 images	
109	Park et al., 2016	Twitter	86 015 tweets	
110	Ho & Li, 2022	Twitter	420 000 tweets	190 487 tweets
111	S. Xu, Yin, & Lou, 2022	GubaEastmoney (forum)	146 634 124 posts	
112	S. Zhou et al., 2018	Apple App Store	>3 500 000 reviews	3 000 305 reviews
113	S.-Y. Lee et al., 2019	StackExchange.com	1 193 394 posts	
114	Samtani et al., 2017	7 discussion forums	431 518 posts	
115	Sanford, 2022	Twitter	4 billion tweets	3 988 000 tweets
116	Sarin et al., 2021	Twitter	89 908 tweets	
117	Saxton & Neu, 2022	Twitter	5 099 524 tweets	297 424 tweets
118	Schlosser et al., 2021	Twitter	119 505 204 tweets	
119	Schneider & Guptab, 2016	Amazon	33 507 reviews	
120	See-To & Yang, 2017	Twitter	1 170 414 tweets	24 516 tweets
121	Serrano et al., 2021	Airbnb	176 852 704 comments	13 181 297 comments
122	She & Michelon, 2019	Facebook	21 166 posts with 1 525 955 comments	
123	Shi et al., 2017	Aviation Safety Reporting System ASRS	168 227 incident reports	158 047 incident reports
124	Shokouhyar et al., 2021	Twitter	74 287 035 tweets	37 878 062 tweets
125	Song et al., 2017	YouTube	6 431 471 comments	
126	Subroto & Apriyana, 2019	CVE database, Twitter	83 015 vulnerabilities, 25 599 tweets	
127	Subroto & Christianis, 2021	Airbnb	66 630 reviews	55 377 reviews

128	Sun et al., 2021	Finance forums	>200 000 000 forum posts	
129	Sundararaman et al., 2018	MIMIC-III dataset	>58 000 hospital admission	11 318 hospital admissions
130	T. Geva et al., 2017	Search engine queries, discussion forums	Not stated but surmised to be well above the threshold	
131	T. Hu et al., 2017	Foursquare, Twitter	1 999 676 check-ins	
132	T. Li et al., 2018	Twitter	1 278 604 tweets	1 161 831 tweets
133	Tang et al., 2020	Twitter, Weibo	275 325 tweets, 80 793 posts	
134	Tao & Fang, 2020	Yelp, Wine Reviews winemag.com, Rotten Tomatoes	10 000 sentences, 80 638 wine reviews, 48 755 movie reviews	
135	Toshniwal et al., 2019	Twitter	19 301 623 tweets	
136	Triantafyllidou et al., 2017	MTFL, WIDER FACE	600 000 facial pictures	
137	Villegas & Martinez, 2022	Twitter	81 190 tweets	
138	W. Zhang & Ram, 2020	Twitter	17 175 642 tweets	9 096 self-reported asthma patients
139	Walker & Boamah, 2020	Twitter	109 607 tweets	
140	Wan & He, 2019	MS COCO dataset, VQA dataset	204 721 images, 614 163 questions	
141	Winkler et al., 2016	Amazon	2 234 519 reviews	1 050 000 reviews
142	Wu et al., 2019	Sina Weibo hot topics	1259 hot topics + 138 609 microblogs	
143	Xi et al., 2022	Airbnb	240 484 reviews	195 704 reviews
144	Xia Liu, 2019	Twitter	28 949 448 tweets	
145	Xia Liu, 2020	Twitter	84 000 000 tweets	61 000 000 tweets
146	Xia Liu et al., 2017	Twitter	1 728 880 tweets	
147	Xia Liu et al., 2021	Twitter	3 780 000 tweets	
148	Xiao Liu et al., 2016	Twitter, Google Trends, Wikipedia, IMDB, Huffington Post	Nearly 2 billion tweets, 113,3 million searchers, 50,7 billion page views, 4 300 reviews, 5,5 million articles	
149	Xiaomo Liu et al., 2020	Sun Forums, Apple Discussions	712 531 posts	
150	Xie et al., 2020	Bitcoingtalk.org	>500 000 messages	
151	Xing et al., 2022	Weibo	108 061 posts	
152	Y. Chen et al., 2019	Change.org	45 377 petitions	
153	Y. Chen et al., 2020	CIFAR-10, CIFAR-100	120 000 images	
154	Y. K. Oh & Yi, 2021	Amazon	49 130 reviews	
155	Y. Li et al., 2020	xywy.com	75 468 medical questions + answers	
156	Y. Zhang et al., 2023	Weibo	55 358 posts	
157	Y.-E. Park & Javed, 2020	Twitter	249 710 tweets	
158	Y.-K. Oh, 2020	Amazon	69 633 reviews	
159	Y.-Y. Wang et al., 2021	Facebook Twitter	785 176 posts 3 006 274 tweets	
160	K.-C. Yang et al., 2021	Twitter Facebook Youtube	53 000 000 tweets 37 000 000 posts 16 669 videos	
161	X. Yang et al., 2020	Eastmoney.com	> 5 800 000 posts	990 415 posts

162	Yi & Oh, 2022	Amazon	106 980 reviews	
163	Yin et al., 2020	Eastmoney.com	6 329 256 posts	5 284 941 posts
164	Yoon et al., 2017	Twitter	744 076 tweets	
165	Yu et al., 2019	Weibo	152 964 368 microblogs	
166	Yue et al., 2019	Discussion forums	2 960 893 posts	
167	Z. Liu et al., 2017	Weibo	103 778 messages	
168	Z. Wang et al., 2020	Lending Club	40 010 loan observations	
169	Z. Zhang et al., 2020	Sloan Digital Sky Survey Data Release 7	1 TB worth of images	

Appendix B: Challenges and Solutions

The tables below depict the identified challenges and their respective solutions. The numbers in brackets refer to the numbers assigned to each reviewed article in Appendix A. Some papers acknowledge a challenge as something affecting the study but do not suggest or implement a solution (i.e., the challenge and its solution(s) are not always presented in the same article). Furthermore, some studies are quite straightforward, such as using sentiment analytics software (e.g., Sentistrength, StanfordNLP, LIWC), a lexicon (e.g., NRC emotion lexicon, VADER), or applying a commonly used algorithm such as LDA, SVM or Naïve Bayes without mentioning any challenges. For clarity's sake, we have included here references to articles reporting or encountering more particular challenges that either provide original solutions or describe applying known solutions in a detailed manner.

Table B1 Challenges and Solutions Regarding Locating Relevant Data

Challenge	Solutions
Due to the volume of data, it is not possible to manually identify relevant data such as reviews of unsafe toys [141], human rights events [39], community building posts [45], corporate social responsibility posts [122], tweets about luxury brands [147] or inflation [7], wildlife product promotion [101], abusive messages [21], or customer complaints [47].	Creating keywords associated with injury based on recall reports by authorities [141] or GRI reporting guidelines [122], creating a tailored vocabulary and only retaining tweets that have at least one term, conceiving the Twitter network as a heterogeneous graph [39], creating a supervised machine learning classifier based on random sample coded by volunteers [45] or building a lexicon of compliment and complaint n-grams to build a classifier [47], using Twitter's mention mechanism to find tweets specifically mentioning certain brands [147], using a predefined list of insults and symbols considered abusive, counting the abusive words and calculating a tf-idf score [21] or using topic modeling to identify contents relevant for the research aim [7, 101].
All collected data is relevant, but the study is interested in a particular aspect such as risk factors from incident reports [123], answer's level of health literacy [70] or level of politeness [113], strong sentiments towards Bitcoin [78], tweets signifying market sentiment [120], low- and high-credibility social media content [160], checking scientific articles for the presence of figures and their type [97] or determining job category [13].	Combining topic modeling with a classification model [123], using general-purpose dictionaries by mapping health-related terms expressed in messages to professional health terminologies [70], studying tweets through finance sentiment dictionary of 2,329 negative and 297 positive sentiment words [78] or matching job offers' frequent term vectors with their corresponding official labor classification codes [13]. Manually coding training data for a classifier [70, 97], quantifying politeness level through a combination of linguistic markers likely associated with impoliteness [113], looking for Twitter's cashtags with words "bullish" and "bearish" [120], classifying misinformation at source-level comparing URLs to known low-credibility domains [160].
Identification of data of interest among millions of entries when search terms produce too many results [74, 83, 131, 166].	Concentrating search around events of interest [83, 131], guiding the search with other data [166], or using a specific search tool [74].
Relevant search terms do not occur simultaneously in documents or are otherwise sparse [35, 92].	Arranging terms under topics and using the topics to find relevant documents [35, 92].
No applicable dictionary or lexicon exists to gain an understanding of what is in the data [34, 70, 84, 114].	Creating a Polish lexicon by defining the semantic orientation of 3280 lexical items [34], manually coding a small sample and using it to train a classifier [70, 114], or using an unsupervised classifier that does not require training data [84].
Manually creating key terms or a dictionary might miss colloquial and informal terms [84], lexicons have	Using a naïve Bayes classifier to develop a hygiene dictionary to identify hygiene-related concerns expressed in colloquial and informal speech [84], combining a tailored dictionary of hand-picked frequent words with a basic sentiment lexicons [128].

difficulties in keeping up with the pace languages develop [128].	
Identifying content produced by a particular kind of user such as a journalist, organization, or individual [117] or “green” Airbnb users [121].	Studying differences in account behavior with Harvard IV/General Inquirer dictionaries, journalists and organizations are expected to use more financial inscriptions and avoid value-based inscriptions [117], iterative compiling of a list of words associated with sustainable lifestyles [121].
Few Twitter users enable geolocation limiting available data for studies requiring geolocated data [12, 19, 137].	Multi-head self-attention model for text representation combined with joint training of city and country labels [19] or matching tweet contents with addresses such as county, street number, street name, city, zip code, etc. [137] to determine tweeters’ location.
All words deemed abusive do not indicate cyberbullying intent making the bag-of-words approach ineffective in detecting cyberbullying [110].	Using Linguistic Inquiry and Word Count (LIWC) to create charged language-action cues [110].
No widely accepted list of stop words indicating the end of the sentence in Chinese [108].	Creating a list of stop words based on the frequencies in data [108].
Consumers’ search intents vary while the keywords used might be the same [43].	Focusing on the search goal and using a dictionary relevant to it [43].

Table B2 Challenges and solutions regarding addressing noisy data

Challenge	Solutions
Polysemy, depending on the context, the word’s sentiment may change [4, 26, 31, 85, 105, 141, 154].	Removing polysemous first names from data [4], using a lexicon of positive and negative words to count positivity degree [31], extracting sentence/local and domain/global level features [85, 105], creating a tailored dictionary for given context [141, 154].
Filtering out unrelated content that includes relevant keywords [60, 101, 115, 138].	Focusing on the most active discussants for manual perusal [60], using topic modeling to focus on the content with the strongest connections to the research question [101], creating n-grams that should not exist in the given context and checking the data for them [115], and use of natural language processing and comparing if the prevalence of the phenomena in the data matches the phenomenon’s prevalence in the population [138].
Presence of informal language, abbreviations, misspellings, punctuation errors, nondictionary slang, wordplay, emoticons, and URLs [10, 31, 126, 129, 132, 138].	Hiring bilingual annotators familiar with the phenomenon [10], checking the most frequent out-of-vocabulary-words and correcting, ignoring, or analyzing them [31], removing noisy features [126, 129, 132], combining natural language processing, machine learning, domain adaptation [138].
Addressing extended words [38] such as “Goooooood” [19], or “loooool” [21] or “hungryyyy” [98].	Removing any characters that repeat more than twice [21, 98], breaking unknown words into smaller parts with a subword feature [19], or preserving the extended words and adjusting their weights accordingly [38].

Differentiating between humans and bots as content creators [71, 144, 145, 146].	Looking at the tweet object's source field to determine if the tweet is sent by a bot [71], looking at behavioral cues such as the timing of tweets and whether the same message is being tweeted constantly, if the messages consist mostly of links, and if there is a discrepancy between accounts followed versus followers or replies and mentions [144,145,146].
The bigger the dataset's coverage, the less accurate search results will become [52, 130, 135].	Removing backgrounds from images to make face detection easier [52], keeping search terms as simple and minimal as possible [130], and excluding extremely short messages [135].
Ensuring the veracity of collected data [39, 62, 150].	Considering Twitter as a heterogeneous graph [39], comparing tweet's contents against other data sources to see if it is true or not [62], focusing on network cohesion as directly connected nodes in a network tend to share similar information decreasing diversity in data [150].
There could be several discussions in one thread or document [1, 59].	Breaking posts and documents into sentences and paragraphs, respectively [1, 59].
Sentence contains opposite sentiments [143] such as "This earbud has good sound quality but poor battery life" [154].	Use of a neural network-based sentiment analysis model [143], dividing the sentence into phrases containing a single feature-sentiment bigram using conjunctions and punctuation [154].
Issues relating to a given phenomenon may be presented in various forms, as it is highly likely that different terms are used to refer to the same topic [24, 112].	Analyzing uploaded images instead of their user-given labels [24], constructing a dictionary directly from the text using the naïve Bayes classifier, and applying SVD to reduce the number of keywords [112].
User provided labels are sometimes incorrect [5, 24].	Ignoring given labels and labeling the data with named entity recognition using a conditional random field approach [5], using a convolutional neural network to generate labels to detect social signals from user-shared images [24].
Improving classifier algorithms' performance [89], and enhancing the accuracy of sentiment quantification [90, 163].	Combining themes from topic modeling with linguistic features to improve supervised classifiers [89], combining four sentiment dictionaries to create an encompassing and domain-specific sentiment dictionary [163], or combining dictionaries and adding inductively generated words [90].

Table B3 Challenges and Solutions Regarding Preserving Data Richness

Challenge	Solutions
When content is classified (e.g., into sentiments), some of the "rich descriptions" are inevitably lost [41, 54, 65, 67, 84, 100, 105, 121, 134].	Generating dictionaries from the data to supplement readymade dictionaries [84], using aspect-based sentiment analysis [54, 105, 121, 134], estimating sentiment scores from topics' most important words [65], or establishing association rules between sentiments and issues [67].
Algorithms have difficulties in capturing subtleties of human language limiting the depth of insight [14, 17, 91, 106, 107].	Blended approach combining big data analytics and manual content analysis to provide nuance through significantly smaller subsamples [14, 106, 107] or analyzing a limited number of relevant user-timelines to achieve descriptive rather than explanatory depth [17].

Training data tends to be small compared to the collected data and expensive to produce for highly domain-specific information [44, 73, 134, 138, 140, 155].	Employing transfer learning to fine-tune existing models requires less training data [44, 134], multimodal data requires less training data for a well-working model [73, 140], training a distant supervision model with weak labels [138], creating automatic medical knowledge extraction system [155].
Increasing the size of [63, 84, 136] or verifying researcher annotated training data [49,132] or model [57, 96].	Recruiting people from crowdsourcing platforms to annotate training data [63, 84], mirroring images to double to size of training data [136], having multiple crowdsourced annotators to verify annotations [49, 132], to check topics' coherence [57] or to compare model's results against human coded sample [96].
Ensuring that generated topics are meaningful [8, 77, 92, 93, 144, 156], not too fine- or coarse-grained [9], and robust so that no interesting topics remain hidden [43, 44].	Topic significance [8] or coherence [77] testing, recruiting students to validate topic interpretations [144], triangulating the number of topics and having topics checked by bilingual speakers [156], triangulating the number of topics [7, 9, 43, 44].
Unbalanced dataset prevents the classification of data [120] and introduces bias to results [127].	Random sampling of a bigger dataset to make it match the smaller dataset [120], using undersampling where entries of a bigger dataset are deleted until balance is achieved [127].
Extracting a representative sample, not what is popular [48, 92].	Categorizing content based on their similarities with one another and then taking a sample [48], expanding core terms based on conditional probability [92].
Language related challenges such as algorithms being tuned only for English [41, 81], inaccurate language detection of short texts [63], and identifying topics in a multi-language dataset [103].	Using Google's translation tool to translate text into English [81], training a Language-Aware String Extractor to recognize different languages in short texts [63] or detecting the language based on n-grams generated from Wikimedia abstracts [81] and detecting topics in multi-language data with a Word2Vec algorithm pretrained in 294 languages [103].
Big data analytics reduce complex issues into ostensibly clear but superficial presentations [139] or do not extract theoretically meaningful features [149].	Supplementing analytics with deep thematic knowledge [139], theory-driven text analytics by combining machine learning with extant theories [149].
Standard LDA does not allow the use of supervised labels to incorporate expert knowledge [20].	Semi-supervised LDA where a sample of documents are read and tagged by humans [20].
Latent Dirichlet Allocation produces static topics that do not account for how discussion evolves over the years [145].	Applying topic modeling separately to temporally different subsets such as for every December in the data rather than the whole data [145].

About the Authors

Sampsa Suvivuo is a doctoral researcher pursuing a PhD in information systems in Aalto University School of Business where he also received his M.Sc. in Economics and Business Administration, majoring in information systems. His current research focuses on the digital sustainability of the gig economy and the resilience of online labor platforms. His previous research has been published in Hawaii International Conference on System Sciences and European Conference on Information Systems.

Virpi Kristiina Tuunainen is a Professor of Information Systems Science at the Department of Information and Service Management and the Associate Dean of Research and International Cooperation at Aalto University School of Business. Her current research focuses on users' trust in artificial intelligence (AI), the division of labor between humans and AI in organizations, and the implications of AI on consumer experiences. She has served as the VP of Publications of the AIS (Association for Information Systems) from 2013 to 2016, and she received the AIS Fellow Award in 2016. She currently serves as the Editor for EJIS. Her work has appeared in journals such as MIS Quarterly, European Journal of Information Systems, Journal of Management Information Systems, Information & Management, MIS Quarterly Executive, and Communications of the Association for Information Systems.