JOURNAL OF INFORMATION SYSTEMS AND TECHNOLOGY

Journal of Information Systems and Technology
Vol., No. (2025), 9-17
p-ISSN: XXXX e-ISSN: XXXX
Journal homepage: https://athallahpublishing.com/index.php/jistech

Research Paper

# Information Systems Perspective on Data Extraction in Social Media: Toward a Theoretical Framework

Cici Lestari Farida

Universitas Ahmad Dahlan Yogyakarta, Indonesia
*Corresponding author: cici95@gmail.com

ARTICLE INFO

**Keywords**
Data Extraction
Information Systems
Social Media

## ABSTRACT

The exponential growth of social media platforms has generated vast amounts of unstructured data, offering both opportunities and challenges for research and practice in the field of information systems. Effective data extraction from social media is not merely a technical problem but also an issue of integrating computational methods with organizational, social, and ethical considerations. This paper proposes a theoretical framework that situates data extraction within the broader context of information systems, highlighting the interplay between technological infrastructures, algorithmic techniques, and socio-organizational dynamics. By reviewing existing approaches to social media data extraction, including text mining, natural language processing, and big data analytics, the framework provides a structured lens for understanding the complexities of transforming unstructured social media content into meaningful insights. The study also addresses limitations such as data reliability, privacy concerns, and platform dependency, while emphasizing the importance of interdisciplinary perspectives. Ultimately, the framework seeks to advance the theoretical foundations of information systems research on social media data, bridging the gap between computational methodologies and organizational knowledge creation.

## Introduction

The proliferation of social media platforms has fundamentally transformed the ways individuals communicate, share information, and interact in digital spaces. With billions of daily posts, comments, images, and videos, social media generates vast volumes of unstructured and dynamic data. For researchers and practitioners, this unprecedented data stream offers immense opportunities to study social behavior, organizational dynamics, market trends, and public discourse. However, extracting meaningful insights from such data is a complex endeavor that raises technical, methodological, and ethical challenges. From the perspective of information systems (IS), social media data extraction is not solely a technical process but a multidimensional issue that intersects with organizational objectives, information management, and decision-making processes. Traditional computational methods such as text mining, natural language processing (NLP), and big data analytics have been widely applied to process and analyze social media content. Yet, the integration of these methods into information systems research requires a more holistic theoretical lens that accounts for the interplay between technological capabilities, socio-organizational contexts, and ethical considerations.

Despite the growing body of research on social media analytics, there remains a lack of comprehensive theoretical frameworks that situate data extraction within the broader domain of information systems. Current approaches often emphasize algorithmic performance and scalability while overlooking the organizational, contextual, and normative dimensions that influence how extracted data is utilized. This gap highlights the need for a framework that not only addresses technical challenges but also integrates IS perspectives, thereby bridging computational methods with organizational knowledge creation and decision support. The aim of this paper is to propose such a theoretical framework, positioning data extraction from social media as both a technical and socio-organizational process. By reviewing existing computational techniques and aligning them with IS principles, the study seeks to advance a structured perspective that enhances the theoretical foundations of IS research. In doing so, the paper contributes to a deeper understanding of how social media data can be effectively and responsibly harnessed to support organizational objectives and knowledge development.

Social media has seen explosive growth in popularity over the past decade, with more and more users creating digital footprints of their activities on various platforms [1]. Such rapid expansion of social media signals a transformative shift in the way human interactions and communications are evolving in the digital age. Over four billion people are connected to social media platforms spending a lot of time (an average of two and a half hours per day) on these platforms. In particular, over two hundred and thirty-seven million users are active on X (formerly Twitter) daily, posting over five hundred million tweets. The extensive digital footprints created by users don't just serve as mere records of online activities; they are reflective of the users' identities, preferences, and socio-cultural inclinations. This has opened up a vast potential for research, offering an unprecedented amount of data to be studied and analysed. Social media data can help researchers gain insights into user behaviour, trends, and other valuable information [6]. Consequently, the lines between online and offline worlds blur, demanding a deeper understanding of the implications of such intensive social media engagement.

Users within social media space are connected through various types of relationships that includes their locations, biographical data, and the content they post. Through these relationships, users can access a wide range of content shared on different platforms. Furthermore, the data generated by these users can be used to inform decisions and strategies related to marketin, digital communication, and social media analytics. However, there are some issues that come along with dealing with social media data. Chief among these is the issue of data quality data collected from social media is often not as reliable and accurate as data gathered from other sources. Consequently, if not extracted and appropriately processed, data from social media sources may lead to conclusions that may be unreliable. As such, researchers must take extra care when dealing with social media data and use the appropriate methods to ensure accuracy and reliability of research results.

In the quest for meaningful insights, the integrity and relevance of the data collected becomes paramount. While social media offers a vast pool of data, it is also riddled with noise and irrelevant data which, if not carefully filtered, can skew results. Despite the acknowledged need for a purer dataset, a systematic methodology for the collection and processing of social media data remains a gap in research. Such an approach, if developed, would not only enhance the process of data analysis but would also significantly advance the field of social media data analysis. In this paper, we argue that the quality of insights derived is intrinsically linked to the quality of the data extracted. Compared to different data collection techniques, such as questionnaires, interviews or focus groups, social media analysis works with unique data. It provides researchers with access to incredibly large sample size, with the potential for access to over 414,000 tweets or 1.3 million Instagram posts. As such, it is an invaluable tool for researchers, allowing them to expand knowledge, create research questions for future qualitative research, and increase validity via triangulation when used as an alternative research method.

The absence of a standardized and systematic approach for collecting and processing social media data, compromises the quality and credibility of subsequent data analysis and hinders the realization of the full potential inherent in social media data. Researchers recognize the vast potential of social media data for understanding human behaviour and societal dynamics, but the lack of a comprehensive framework for data extraction and processing poses a significant challenge. The research aims to develop a systematic framework to address this gap and enhance the quality and reliability of data obtained from social media platforms. Data obtained from social media platforms may subsequently be used for different types of analysis. The data may also exhibit various data quality issues. As a result, the types of analysis that can be conducted using social media data is of prime significance because the proposed framework needs to cater to diverse analytical needs. By understanding these different types of analyses, we can better design a framework that is versatile, ensuring that extracted data is fit for various research purposes and methodologies. Furthermore, given the unfiltered, spontaneous nature of user-generated content on social media platforms, we argue that our framework should not only extract data but should also consider how to render it usable for meaningful analysis. Ignoring quality parameters could compromise research integrity, leading to flawed conclusions. In light of these considerations, our discussion of both analytical possibilities and data quality challenges sets the stage for introducing our proposed framework.

## Method

This study adopts a conceptual and theory-building research design, aiming to develop a framework that explains the role and process of data extraction in social media from an Information Systems (IS) perspective. A comprehensive literature review was conducted to identify relevant studies across the domains of information systems, computer science, data science, and social informatics. The review included peer-reviewed journal articles, conference proceedings, and authoritative reports published within the last 15 years. The selection focused on works addressing social media analytics, data extraction techniques (e.g., natural language processing, text mining, big data analytics), and IS theories related to information management, organizational use of technology, and ethical considerations.

The reviewed literature was analyzed and mapped against existing IS theoretical perspectives such as socio-technical systems theory, information processing theory, and knowledge management frameworks. This step aimed to position computational methods for social media data extraction within broader organizational and socio-technical contexts, identifying how technical processes interact with social, organizational, and ethical dimensions. Based on insights from the literature and theoretical mapping, a conceptual framework was developed to articulate the Information Systems perspective on social media data extraction. The framework integrates technological capabilities (e.g., algorithms, infrastructures), organizational goals (e.g., decision-making, knowledge creation), and contextual factors (e.g., ethical norms, regulatory constraints). Iterative refinement was carried out by comparing the proposed framework against established IS models to ensure theoretical consistency and conceptual clarity. This methodology ensures a structured and rigorous approach to theory building. Rather than testing hypotheses through empirical data collection, the study contributes to knowledge by synthesizing interdisciplinary insights into a coherent theoretical foundation. The resulting framework is intended to guide future empirical research and inform practitioners on the responsible and effective integration of social media data extraction within information systems.

## Results and Discussion

In this section, we present a framework for the extraction and processing of social media data (Figure 1). The proposed framework consists of a carefully designed sequence of steps. When followed, these steps enhance the credibility of social media data extraction, ensuring the acquisition of comprehensive, relevant, and reliable data.

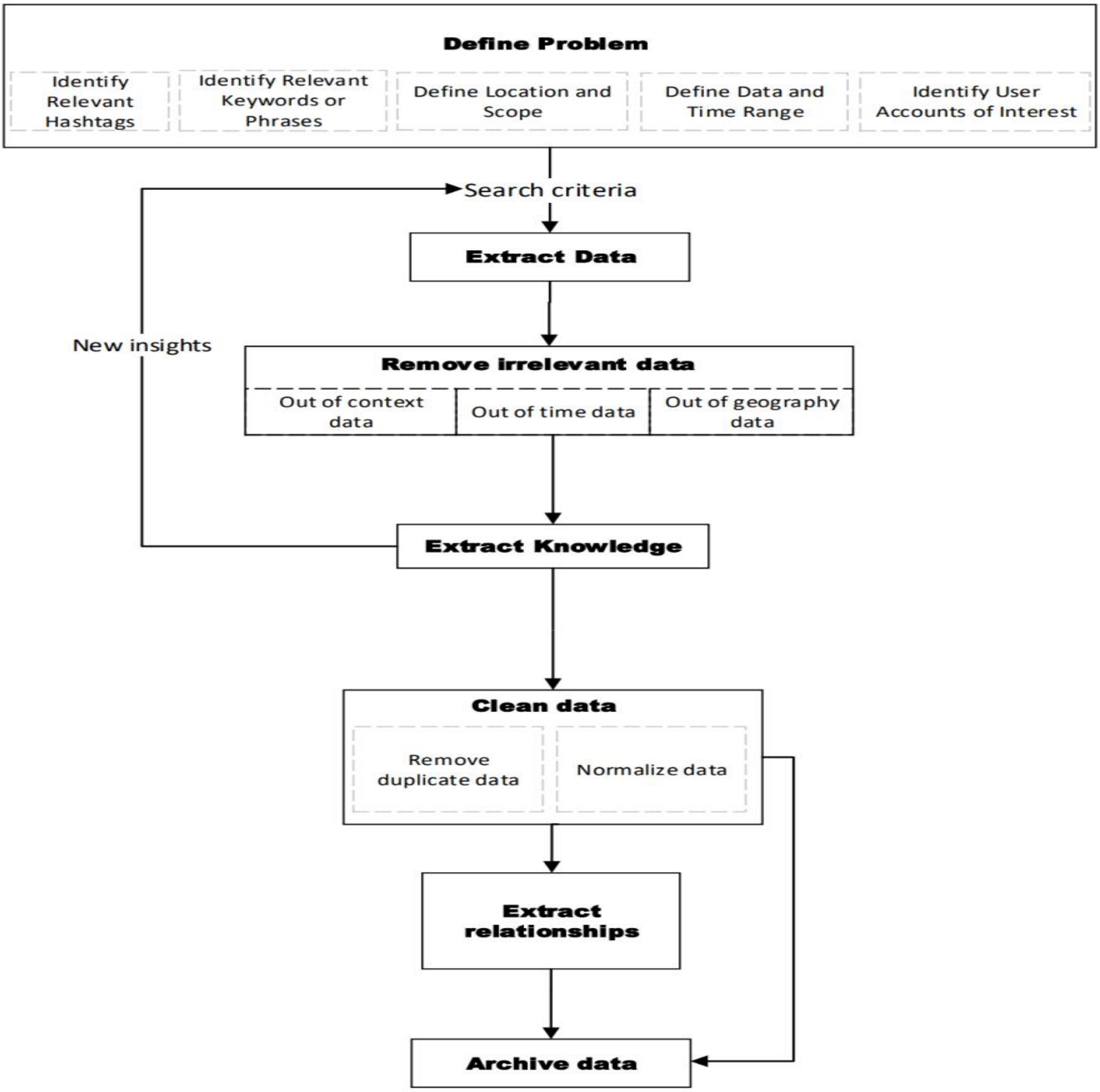

Figure 1. Framework for extracting and processing social media data

### Define the Problem

The first step in the knowledge discovery process is to define the problem. In the case of social media data, this involves understanding the research question, the goals and the objectives of the research. In addition, it is crucial to identify constraints or limitations and determining the data that is available to be used. In essence, this step seeks to answer the question, "Given a specific research question, how do we define a search criterion such that the data collected is sufficient to answer the research questions?". In the subsections below, we discuss some important ways of defining a search criteria.

### Identify Relevant Hashtags

Hashtags are used on various social media platforms to search, organize topics, categorize content, find conversations about relevant topics, and join in on discussions. In essence, hashtags make it easier for users to quickly find topics of interest and engage with other users who are talking about the same topics. This helps to facilitate conversations, promote events, and allow users to locate relevant content easily. Additionally, hashtags can be used to draw attention to a tweet or a topic and indicate that a tweet is part of a larger conversation. Hashtags may also be used to start conversations, to join conversations, or to amplify existing conversations. By using hashtags, actors can participate in larger conversations and have their voices heard. When searching for data on social media, it is thus, crucial to identify all the relevant hashtags associated with the phenomenon under investigation.

However, searching social media platforms using hashtags is not always the best method for obtaining information, as the search results are limited to those posts that include the specific hashtag. This limits the scope of the search and the potential for obtaining a full picture of a particular topic. Additionally, the use of hashtags can be manipulated to skew the search results in a certain direction, as users may choose to include only certain hashtags in their posts and ignore any other related terms. This can lead to biased search results that do not accurately reflect the full range of opinions on the topic of interest. Finally, the same hashtags may have been used in unrelated conversations. Consequently, using hashtags alone may result in the extraction of irrelevant data. Therefore, while the use of hashtags can be a valuable tool in searching social media platforms, it is not always the best method for obtaining comprehensive, accurate information

### Identify Relevant Keywords or Phrases

Searching social media using keywords is often better than searching with hashtags because it allows one to be more specific in finding content. A hashtag may be too general, resulting in irrelevant or off-topic results, whereas a keyword search can be tailored to the exact query. Additionally, a keyword search allows for the discovery and extraction of content that does not include a specific hashtag but may still be relevant to the research question under investigation. It is thus crucial to identify a set of relevant keywords or phrases when conducting research on social media platforms like Twitter. By using well-chosen keywords, one can optimize the search results, making the research more efficient and comprehensive.

### Define Location and Scope

Searching social media using location is a powerful tool for conducting research as it can provide valuable insights into people's attitudes, beliefs, and behaviors within a certain geographical area. This is particularly helpful for researchers looking to understand the experiences of people from a particular geographic region, as it allows for a more thorough analysis than national-level data. For example, researchers can compare the sentiment of tweets from different regions to better understand how different populations are responding to an event or issue. Location-based searches can also provide a better understanding of how a local issue plays out in real-time, including how people are talking about it and how their views may change over time. Additionally, location-based searches can help researchers identify influencers and opinion leaders in a given region, allowing them to better understand the dynamics of the local conversation. However, not all searches should be restricted to a particular geographical area. There are instances where the research question may seek to investigate a global phenomenon.

### Date and Time Range

By searching within a specific timeframe, researchers can ensure that the search results are up-to-date and relevant to their research topic. Additionally, searching within specific dates allows researchers to find the most recent tweets related to their topic, which can provide valuable insights into current trends and events. For instance, by examining the conversation over time, researchers can gain an understanding of how public sentiment changes and how the conversation evolves. Furthermore, searching within specific dates can help researchers identify patterns and connections between different tweets, as they can compare the results of different searches and identify any commonalities between them. Searching X(formerly Twitter) within specific dates can allow researchers to observe the influence of certain individuals or organizations on the conversation and to identify key influencers who have the ability to shape public opinion. This method can also be used to analyze the effectiveness of a particular promotional campaign or marketing strategy.

Setting specific time frames when collecting data on social media is paramount for a variety of research purposes. For instance, event-centred studies necessitate data extraction from a period directly surrounding the event to grasp public sentiment. Similarly, understanding the evolution or emergence of a trend over time requires segmenting data into distinct periods. Comparative analyses, seasonal studies, and longitudinal investigations all benefit from well-defined time windows, ensuring accuracy and relevance. Additionally, given the vast volumes of daily data on platforms like Twitter or Facebook, narrowed time frames offer more manageable and focused datasets

## Conclusion

In summary, this paper has introduced a comprehensive framework for the extraction and processing of social media data. The framework involves a series of well-defined steps, including problem definition, data extraction, cleaning, knowledge extraction, and relationship analysis. The key emphasis is on the importance of clearly defining the research problem and setting search criteria, the process of data extraction, the crucial step of data cleaning to remove irrelevant information, and the iterative nature of refining the search criteria. Data normalization and the removal of duplicate data are highlighted as essential for data preparation. The framework also emphasizes the significance of extracting relationships among actors in social media data. By following these systematic steps, researchers can enhance the credibility of their work, ensuring that the data they gather is both comprehensive and reliable. While the framework is tailored for Twitter, its principles and processes can be adapted for other social media platforms. In the digital age, where social media is a primary source of information and communication, this framework prioritizes data quality through cleaning and processing steps, equipping researchers to navigate this complex landscape, uncover valuable insights, and inform decision-making and innovation.

## References

Avgerou, C. (2019). Developing information systems: Concepts, issues, and practice. *Palgrave Macmillan.* Cao, G., Duan, Y., & El Banna, A. (2019). Data-driven approaches for sustainable digital transformation: A theoretical framework. *Information Systems Frontiers, 21*(5), 1105–1120. https://doi.org/10.1007/s10796-018-9879-2

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons, 53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences.* SAGE.

Kwon, K. H., & Gruzd, A. (2017). Is aggression contagious online? A case of swearing on Donald Trump’s campaign videos on YouTube. *Online Information Review, 41*(5), 782–797. https://doi.org/10.1108/OIR-02-2016-0041

Orlikowski, W. J., & Iacono, C. S. (2001). Research commentary: Desperately seeking the "IT" in IT research—A call to theorizing the IT artifact. *Information Systems Research, 12*(2), 121–134. https://doi.org/10.1287/isre.12.2.121.9700

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management, 39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Vaast, E., & Kaganer, E. (2013). Social media affordances and governance in the workplace: An examination of organizational policies. *Journal of Computer-Mediated Communication, 19*(1), 78–101. https://doi.org/10.1111/jcc4.12032

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems, 25*(6), 13–16. https://doi.org/10.1109/MIS.2010.151