Journal of Information Systems and Informatics
Vol. 5, No. 4, December 2023 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v5i4.585 Published By DRPM-UBD

The Problem of Data Extraction in Social Media: A Theoretical Framework

Tarirai Chani[1], Oludayo O Olugbara[2], Bethel Mutanga[3]

[1,2]Information Technology Department, Durban University of Technology, Durban, South Africa
[3]Information Technology Department, Mangosuthu University of Technology, Durban, South Africa
Email: [1]tariraichani@gmail.com [2]oludayoo@dut.ac.za, [3]mutangamb@mut.ac.za

Abstract

In today's rapidly evolving digital landscape, the pervasive growth of social media platforms has resulted in an era of unprecedented data generation. These platforms are responsible for generating vast volumes of data on a daily basis, forming intricate webs of patterns and connections that harbor invaluable insights crucial for informed decision-making. Recognizing the significance of exploring social media data, researchers have increasingly turned their attention towards leveraging this data to address a wide array of social research issues. Unlike conventional data collection methods such as questionnaires, interviews, or focus groups, social media data presents unique challenges and opportunities, demanding specialized techniques for its extraction and analysis. However, the absence of a standardized and systematic approach to collect and preprocess social media data remains a gap in the field. This gap not only compromises the quality and credibility of subsequent data analysis but also hinders the realization of the full potential inherent in social media data. This paper aims to bridge this gap by presenting a comprehensive framework designed for the systematic extraction and processing of social media data. The proposed framework offers a clear, step-by-step methodology for the extraction and processing of social media data for analysis. In an era where social media data serves as a pivotal resource for understanding human behavior, sentiment, and societal dynamics, this framework offers a foundational toolset for researchers and practitioners seeking to harness the wealth of insights concealed within the vast expanse of social media data.

Keywords: Social media data, data extraction, social media data quality, theoretical framework, social network analysis

1. INTRODUCTION

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 [2]. In particular, over two hundred and thirty-seven million users are active on X (formerly Twitter) daily, posting over five hundred million tweets [3]. 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 [4] [5], 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 marketing [7], digital communication [8], and social media analytics [4] [9]. However, there are some issues that come along with dealing with social media data. Chief among these is the issue of data quality [10] [11]; data collected from social media is often not as reliable and accurate as data gathered from other sources [12-14]. 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 [15].

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 [16]. It provides researchers with access to incredibly large sample size, with the potential for access to over 414,000 tweets [17] or 1.3 million Instagram posts [18]. 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 [19].

This research, therefore, seeks to answer the following question. "What are the key components and methodological steps required to construct a systematic framework for the extraction and processing of social media data, and how can this framework be optimized to enhance the quality and reliability of data obtained from social media platforms?".

"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."

2. LITERATURE REVIEW

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.

2.1. Social Media Analysis

Social media has become an essential part of our lives and how we interact with each other. It has become increasingly crucial for businesses and organizations to understand how social media is being used and how it affects their brand [20]. Social media analysis is an important tool to help organizations gain insights about their customers and what people are saying about them. It involves examining relationships between different users or actors, analyzing content, and analyzing engagement to understand how people interact. By understanding social media trends and behaviors, organizations can better understand their customers and create better strategies to reach them [21, 22]. This section gives an overview of the different types of analyses that can be done on social media data.

2.1.1 Relationship Analysis

People or actors are linked to each other either through bi or unidirectional relationships. In some instances, these actors may be institutions, organisations or companies. These relationships form a network that can be analysed to gain insights. Studying these relationships aids in analysing and modelling relations and diffusion processes [23] among various actors in a social network, to understand how the behaviour of individuals and their interactions translate into a large-scale phenomenon [24, 25]. This type of analysis uses graph theory [26, 27] to identify the relationships between nodes or actors and how they are connected [28]. It can be used to measure aspects like the strength of ties between individuals [25, 29], the influence of an individual or group, and the overall structure of the network [30]. Influence Analysis can be used to investigate how one's behaviour or opinion impacts the behaviour or opinion of others [31, 32]. However, conducting this analysis is not a trivial task due to the complex relationships that exist among different actors. The success of such an analysis depends upon the integrity of the data—both in terms of its completeness and accuracy. Noisy or incomplete data can significantly distort results, leading to potentially misleading conclusions [33].

The precision of data extraction encompassing both the scrapping method and search criteria is paramount to ensure comprehensive and contextually relevant datasets. However, this pivotal aspect of data collection has largely remained unaddressed in prior works. Thus, it is imperative for researchers to prioritize the quality and contextual relevance of data when undertaking influence analysis in networks.

2.1.2 Content Analysis

Users post content in different formats that may be random or part of a topical discussion as a result of trending issues. Content analysis is an important method of analyzing social media data, as it allows researchers to systematically examine the content of social media posts and interactions [34-36]. This process involves identifying, coding, and categorizing data in order to identify patterns, trends, and relationships between social media posts and interactions. It is a powerful tool for gaining insight into how people engage with one another on social media and understanding how different audiences perceive and interpret messages. Content analysis can be used to study various aspects of social media interactions. For example, researchers can analyze the frequency of certain topics or keywords being discussed, the type of language used, and the sentiment expressed in the posts. This can be used to gain insight into how people form opinions and respond to content they consume on these platforms.

In order to perform content analysis, such as topic extraction, it is crucial that the data is cleaned properly. Noisy data has the potential to produce irrelevant results

[33]. In addition, it is also crucial that a complete set of data is acquired. This implies that both the scrapping method and search criteria used in the extraction of data are appropriately formulated. However, to the best of our knowledge, this issue has not been addressed. It is, therefore, crucial to ensure that data used in content analysis is complete and relevant to the context under investigation.

2.1.3 Engagement Analysis

Social media has become a major platform where users engage with one another through various types of feedback, such as likes, comments, retweets, and reactions [37]. These interactions allow users to show support for one another's posts and facilitate meaningful conversations. Engagement analysis involves analysing the interaction and reactions of users with a particular post or topic. Reactions are one-way users can express their feelings towards a post or the person who posted it. For instance, when users like a post, it allows them to express their approval and appreciation for the content. It also allows them to quickly acknowledge a post without having to type out a comment. The likes that a post receives can also be seen by others, making the poster feel more validated [38]. In addition, the number of likes could be an indication of the impact that the post has within the context [39-41].

On the other hand, comments allow users to share their thoughts and reactions to a post and can be used to ask questions and expand conversations. In addition, sharing or retweeting allows users to share a post with their own followers. This allows the content to be seen by a much wider audience. Retweets also allow users to show their support or endorsement for the content and the original poster [42]. Such interactions provide a wealth of information for insightful deductions in various areas, such as influence analysis and sentiment analysis. Consequently, research work that involves analyzing engagement activities relies heavily on data that has been carefully extracted to ensure the credibility of the results.

2.2. Data quality issues in social media data

Data quality in social media research has been an increasingly pressing topic as the use of social media platforms has grown rapidly in recent years [43, 44]. As more and more researchers have become interested in using social media data to answer research questions, there is an urgent need to understand the quality of the data being collected and used [44]. The first step in understanding data quality in social media research is understanding the sources of data being used. Social media data is often generated from a variety of sources, including users' posts and interactions, platform algorithms, and third-party data sources. Each of these sources can present unique challenges as researchers attempt to ensure the quality of the data they are collecting.

2.2.1 Accuracy

Data accuracy is a crucial concern when conducting social media research. Due to the self-reported nature of user-generated content, social media data can be subjective, incomplete, and prone to bias [45]. Additionally, the sheer volume of data generated on social media platforms can make it difficult to verify accuracy due to the lack of a centralized quality control system [46]. To ensure the accuracy of social media research data, researchers must take into account the limitations of user-generated content, employ rigorous data cleaning and verification practices, and develop reliable methods for data extraction and analysis.

2.2.2 Completeness

In research in general, data completeness refers to the extent to which all relevant data has been included in the dataset. Data completeness is a vital issue to consider when conducting research with social media data. Social media platforms often lack complete datasets due to the nature of the content being shared [47]. For example, posts that are removed from the platform can be lost forever, making it difficult to build a comprehensive dataset for research purposes. Additionally, users may opt out of sharing certain data or may not have the full range of information available, making it difficult to have a complete set of data to work with. Finally, certain algorithms and filters used by social media platforms can lead to data being incomplete or missing, reducing the reliability and validity of the data. Some social media platforms require user permission for data extraction, which can be challenging to obtain. Furthermore, many platforms limit the amount of data that can be accessed, such as X's API, which limits the number of tweets that can be collected per 15-minute window [48]. This limitation further makes it difficult to obtain a comprehensive dataset.

2.2.3 Timeliness

Data timeliness is an important factor to consider when conducting research on social media. In the context of social media research, it is vital to ensure that the data collected accurately reflects the current context of the social media platform [49]. This can be challenging due to social media's fast-paced and ever-evolving nature. As a result, researchers must ensure that the data they collect is timely by ensuring it is collected and analyzed promptly. Additionally, they should consider using data analysis techniques that allow them to quickly and accurately assess the data in order to ensure the most recent trends and conversations are accounted for.

2.2.4 Accessibility

Data accessibility issues in social media research refer to the difficulties encountered when accessing and processing data gathered from social media platforms. A variety of factors can contribute to data accessibility issues, including the nature of the data being collected, the way in which the data is formatted and stored, and the technical capabilities of the researcher. For example, social media data may not be organized or structured in a way that is conducive to analysis, or the data format may be incompatible with the software and tools used by the researcher. Additionally, the complexity of the data may present challenges to researchers without advanced technical skills [50]. As a result, data accessibility issues can make it difficult to conduct meaningful research that can be replicated in future. Furthermore, many platforms limit the amount of data that can be accessed [51, 52]. This limitation further makes it difficult to obtain a comprehensive dataset.

2.2.5 Consistency

Data consistency issues in social media research refer to the challenges of ensuring uniformity and accuracy of data collected from social media platforms. Social media data collection is complicated due to the dynamic nature of the platforms and the large volumes of data generated in real-time [53, 54]. As such, researchers must take into account the constant changes in the platforms and the data produced while designing and executing a research project. Data consistency issues can arise from a variety of sources, including the lack of standardization in the data format, the presence of duplicates or errors in the data, and the presence of multiple versions of the same data. Additionally, data consistency issues can be caused by the presence of biased data or the lack of control over the context in which the data is collected. To address these challenges, researchers must develop a comprehensive data collection strategy tailored to the specific research goals taking into account potential data consistency issues. This may include the development of data quality standards and metrics and the use of data visualization and analysis techniques to identify and address these issues.

2.2.6 Validity

Data validity refers to the extent to which the data accurately reflects what it is supposed to represent [55]. In social media research, this refers to the accuracy, consistency, and representativeness of the data collected from social media platforms. This can be impacted by a range of factors, including the selection of data sources, the interpretation of posts, the context of the posts, and the methods used to analyze the data. For example, if the data sources are limited to a particular demographic or region, the results may not reflect the opinions of a wider audience. Similarly, if the interpretation of posts is based on subjective criteria, the

accuracy of the results may be compromised. Furthermore, if the context of posts is not properly taken into account, the data may not accurately reflect the intended meaning. Finally, if the methods used to analyze the data are not sufficiently robust, the results may be unreliable. All of these potential issues can affect the validity of social media research.

2.2.7 Relevance

Data relevance is a major issue that affects research in general and social media analytics in particular. It refers to the degree to which a dataset accurately reflects the phenomenon of interest. For example, when exploring a certain topic, researchers must ensure that their data is representative of the population being studied. In essence, the collected data should be relevant and applicable to the research question. This attribute is crucial in arriving at credible conclusions. In general, data relevance can be affected by factors such as the source of the data, the number of data points, data quality, data selection, and data processing. In the context of Twitter data analytics, data relevance is especially important due to a large amount of data available and the rapid changes in the platform [56]. Twitter data analytics requires careful selection and analysis of data points to ensure that the data is relevant to the research question. This includes data cleaning techniques to remove irrelevant data points, such as bots, spam, and irrelevant tweets. Additionally, data selection should be conducted with consideration of the temporal context of the data. For example, if a study is interested in the sentiment of tweets about a certain topic, tweets that are more than one year old may no longer be relevant to the research question.

2.2.8 Interpretability

In social media analytics, interpretability issues arise from the large and heterogeneous nature of the data [57]. Data interpretability can be a challenge due to the dynamic and heterogeneous nature of the data. Social media data is often unstructured, making it difficult to identify patterns or draw meaningful insights from the data. Additionally, the presence of noise such as spam, trolls, and bots can lead to inaccurate results and obscure the true meaning of the data. As a result, researchers must take extra care when interpreting social media analytics in order to ensure that the results are reliable and valid. Additionally, researchers should consider the use of visualisations to simplify the interpretation of the data [58, 59].

3. METHODS

In this section, we present the methodological approach employed in this study to develop a comprehensive framework for extracting and processing social media data. This methodology is rooted in the idea of using existing literature and research to inform and guide the creation of a practical framework, making it a

literature-driven approach to framework development. The methodology encompassed a literature survey on various aspects of social media analysis, examining the challenges associated with this domain and developing a structured framework for data extraction rooted in the knowledge discovery process.

3.1 Literature Survey on Different Types of Social Media Analysis

To lay the foundation for our research, an extensive literature survey was conducted to explore the various types and methodologies of social media analysis. This phase aimed to comprehensively understand the diverse analytical techniques employed across different social media platforms and data types. The insights gathered from this survey were invaluable in shaping the framework, ensuring its adaptability to a wide range of data sources and analysis objectives.

3.2 Literature Survey of the Challenges of Social Media Analysis

In conjunction with understanding the methodologies, we also conducted a comprehensive literature survey focused on identifying the challenges and issues commonly associated with social media analysis. This step allowed us to pinpoint areas of concern and develop guidelines within the framework to address these challenges. By being well-versed in the obstacles, we aimed to provide practical solutions within our framework to enhance the quality and effectiveness of social media data analysis.

3.3 Development of a Framework for Extracting Social Media Data for Analysis

The core of our research involved the development of a systematic framework for extracting social media data for analysis. This framework was designed based on the knowledge discovery process. It was developed to provide a structured, step-by-step approach for researchers and practitioners, ensuring that social media data is harnessed effectively and consistently. By incorporating the insights gained from the literature surveys, the framework was tailored to address the challenges and nuances specific to social media data analysis.

4. 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.

Journal of Information Systems and Informatics
Vol. 5, No. 4, December 2023
p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

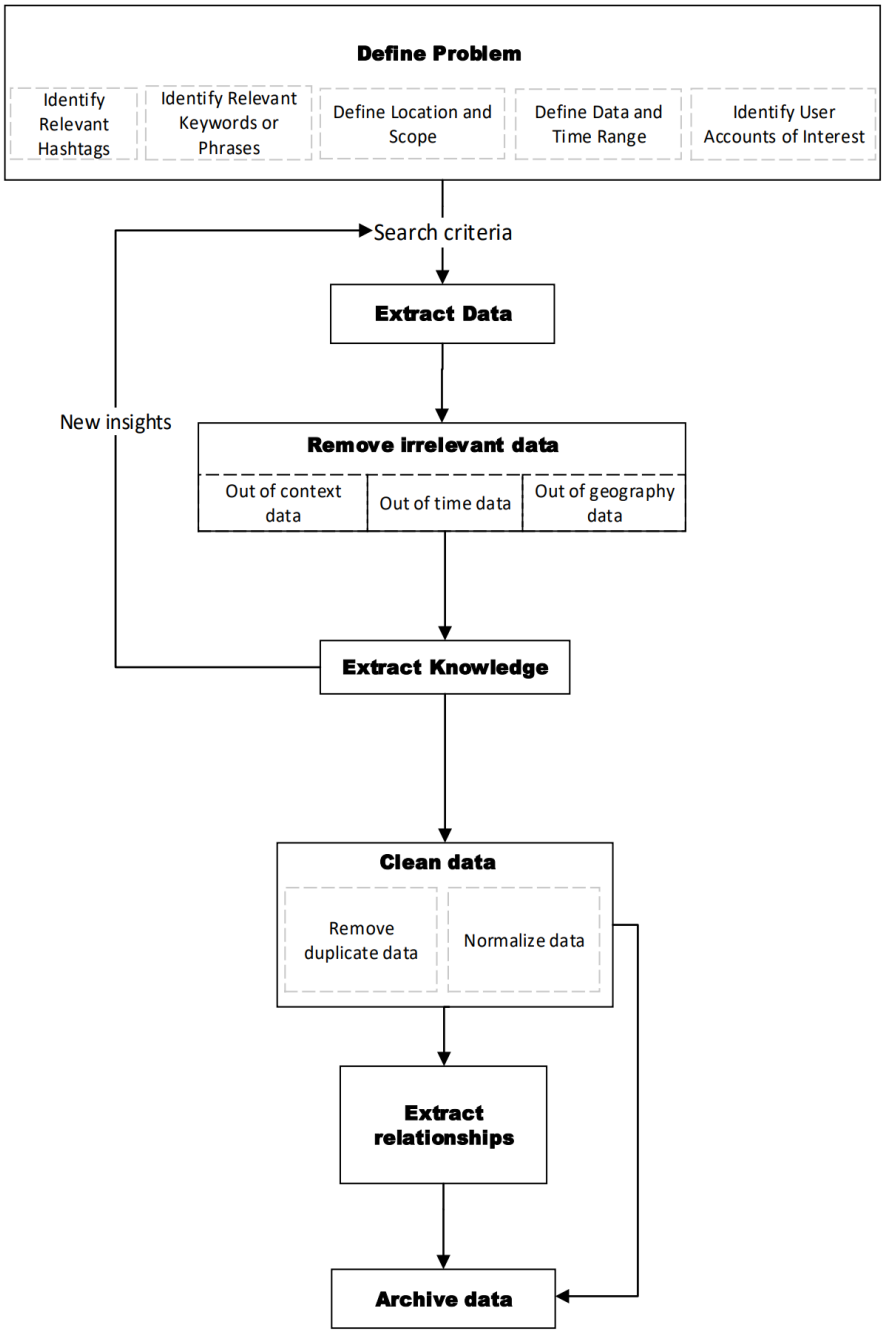

Figure 1. Framework for extracting and processing social media data

4.1 The Framework

In the following sub-sections, we will delve into each step of this framework.

4.1.1 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.

1) 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.

2) 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.

3) 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.

4) 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.

When defining time frames for social media scraping, researchers must consider several factors:

a) **Objective Alignment**: Begin by aligning the research objective with the time frame. If studying reactions to a specific event, the time frame should encapsulate the days or hours surrounding that event.
b) **Tool or API Utilization**: Many social media scraping tools and platforms' APIs allow to specify exact 'from' and 'to' dates. Ensure you correctly set these parameters to narrow down the collection to your desired range.
c) **Time Zone Adjustments**: Considering the global nature of social media platforms, ensure differences in time zones are catered for. This ensures that capturing data accurately for the desired time in different regions.
d) **Platform Restrictions Awareness**: Consider any limitations the platform or API sets, such as data quotas for a given period. This might require breaking the desired time frame into smaller chunks or adjusting the range to fit within these limits.

5) Identify User Accounts of Interest

Searching Twitter from specific accounts can be beneficial in a variety of ways. First, it allows researchers to focus on the content of a specific account, which can provide a more in-depth analysis of the content than a broad search of all Twitter users. For example, a researcher may want to explore the opinions of a particular political leader, which can be accomplished by searching Twitter exclusively from that account. Additionally, searching Twitter from specific accounts can offer insight into the account's engagement with its followers. For instance, a researcher may want to explore the level of interaction between a brand and its followers, which can be easily accomplished by searching the brand's Twitter account. Finally, searching Twitter from specific accounts can provide a more comprehensive picture of the account's content over time, as the search results

can be divided into time frames to allow for analysis of content trends. This can be especially useful for researchers interested in examining how an account's content has changed over time. Overall, searching Twitter from specific accounts can provide researchers with valuable insights into the content and engagement of an account.

There is an undeniable necessity for targeted data collection from specific accounts, particularly when these accounts hold relevance to the subject under investigation. It is crucial to identify these key user accounts for the following reasons:

a) **Contextual Relevance and Influence**: We exist in a complex social fabric where certain individuals, often termed as 'influencers' or 'key opinion leaders', exert considerable influence over their audiences. Their viewpoints, content, and interactions can set the tone for entire conversations, debates, and trends. Mining data from these specific accounts provides a more in-depth understanding of the sentiments, narratives, and potential impacts they are creating.
b) **Accuracy in Analysis**: When addressing specific research questions or exploring particular themes, broad-based data collection can sometimes result in noise and reduce the precision of analysis. On the other hand, targeting accounts that are central to the conversation ensures that the collected data is directly relevant, improving the accuracy and efficacy of subsequent analyses.
c) **Unearthing Hidden Patterns**: Influential accounts often act as nodes or hubs in social networks. By focusing on these nodes, researchers can uncover intricate patterns of information flow, discerning how ideas spread, evolve, and gain traction.

4.1.2 Extract Data

Once the problem has been clearly defined, the search terms are established as the criteria for retrieving data from Twitter. It is important to note that this initial search might not yield a complete dataset that can be considered representative. Therefore, it is imperative to ensure a thorough search has been conducted by carefully examining the extracted data for potential sources of new and insightful knowledge that could necessitate adjustments to the search terms. However, it is worth acknowledging that a considerable amount of noisy data may have been collected. Consequently, prior to analyzing this data to extract new knowledge, an initial round of data cleaning is essential, starting with the removal of irrelevant information.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

4.1.3 Remove Irrelevant Data

Social media data contains all sorts of noise as a result, the data collected needs to be cleaned to make it useful for analysis. There are many forms of cleaning the data however the initial step to cleaning the data would be to remove domain irrelevant data. Removal of irrelevant data from social media data prior to analysis involves identifying and removing erroneous, redundant, incomplete, or out-of-date data from the dataset. Irrelevant data removal is typically done by employing a combination of automated processes, such as data validation and data filtering, and manual processes, such as data scrubbing and data verification. This process involves removing outliers, correcting errors, and transforming data into a format that can be used for analysis. Once the irrelevant data is removed, the remaining data can be analyzed to draw meaningful insights. In this work we define irrelevant data as that which is "out of context", "out of time" or "out of geography".

a) **Out of context:** Irrelevant data on social media research can be defined as data that is unrelated to the research topic being studied. This can include posts, comments, and other user-generated content that is not directly related to the research topic. Examples of irrelevant data on social media research could include posts about current events, personal experiences, or topics unrelated to the research study. In order to ensure that data collected is relevant to the research topic, it is important to carefully review and filter out any posts or comments that are not directly related to the research topic.
b) **Out of time:** Involves data that may be relevant to the topic but that is outside the time scope of the current analysis.
c) **Out of geography:** This refers to data that is relevant both contextually and in terms of the time frame but not relevant due to geography. For instance, a person in Africa may comment on issues of current debate on racism in America. Although the comment and time period are relevant to the debate, the comment may be classified as irrelevant if the goal of the research is to investigate the comments made by people in America.

4.1.4 Extract Knowledge

The process of extracting knowledge from the data is essential as it can potentially lead to the discovery of new domain specific insights. These new found insights can be used to enhance the search criteria initially formulated during the problem definition phase. These refined search criteria subsequently guide the extraction of more data from Twitter, necessitating a subsequent removal of irrelevant information. To ensure data completeness this iterative cycle must continue until there is no knew knowledge to extract. In this paper, we argue that data completeness is reached when no new insights can be learnt from the extracted data. The repetitive nature of knowledge extraction presents the possibility for duplicate data entries within the extracted content.

4.1.5 Clean and Process Data

This crucial step involves normalizing and removing data duplications which is essential for improving data quality for a more credible analysis result.

1) Normalize the Data

Normalize the data to ensure that it is in a consistent format. This may include converting dates to a standardized format, converting text to a uniform character encoding scheme, or converting numbers to a common unit.

2) Identify, Document and Remove Duplicate Data

Data duplication is a common problem in social media research, as multiple users may share the same content or information. This can cause issues with data accuracy, as the same content is being used multiple times and the data may not be representative of the population as a whole. Data duplication can also lead to skewed results, as the same content is being used multiple times in the analysis. However, data duplication may point to the importance of the content. As a result, such duplications should be recorded before being removed.

4.1.6 Extract Relationships

Extracting actors' relationships within social media data is pivotal for comprehensively understanding online ecosystems. By mapping these connections, researchers can discern the structure of social networks, pinpoint key influencers, trace the flow and evolution of information, and gauge sentiment dynamics. Such insights are instrumental for predictive modeling, crafting informed marketing or crisis management strategies, and enhancing recommendation system personalisation. However, with these benefits comes the responsibility of addressing ethical and privacy considerations, ensuring that data is used respectfully and transparently.

4.2 Discussion

The proposed framework for extracting and processing social media data from Twitter offers a systematic and comprehensive approach to conducting research in the digital age. It addresses various critical aspects of data collection, cleaning, and analysis, focusing on enhancing the credibility and reliability of research outcomes. This section discusses the significance and implications of this framework in the context of social media research and its potential for broader applications.

4.2.5 Data Quality and Credibility

Data quality is a paramount concern in any research, and it is particularly challenging in the context of social media data. The framework underscores the importance of rigorous data cleaning and normalization, specifically removing irrelevant information. This is essential to ensure that the insights derived from the data are accurate and reliable. By defining "irrelevant data" as that which is "out of context," "out of time," or "out of geography," the framework provides clear guidelines for researchers to distinguish valuable data from noise. This approach contributes to the overall credibility of research findings.

4.2.6 Flexibility and Adaptability

One of the strengths of the framework lies in its flexibility. It recognizes that different research questions may demand different data collection and analysis criteria. Offering options such as hashtags, keywords, geographical filters, and time frames allows researchers to tailor their approach to the specific needs of their study. Moreover, the iterative nature of data and knowledge extraction enables adaptability, ensuring that research criteria can be refined as new insights emerge. This adaptability is especially crucial in the ever-evolving landscape of social media.

4.2.7 Beyond Twitter

While the framework is primarily tailored for Twitter, it is worth noting that many of its principles and processes can be adapted for other social media platforms. The digital age has seen the proliferation of various social media channels, each with its unique data characteristics. Researchers can leverage the framework's systematic approach as a blueprint for conducting research on different platforms, ensuring that the data they gather remains comprehensive, relevant, and reliable.

4.2.8 Future Directions

As social media continues to evolve, so do the methods and tools for data collection and analysis. Future research may explore ways to integrate emerging technologies, such as natural language processing and machine learning, to automate certain aspects of the framework, further enhancing efficiency and accuracy. Additionally, the framework's applicability to different research domains, from marketing and public opinion analysis to crisis management and sentiment monitoring, opens the door to exciting opportunities for cross-disciplinary collaboration.

5. 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

[1] D. L. Rodkey, S. Y. Nelson, A. E. Lundy, and M. D. Helgeson, "Exponential growth of social media utilization among orthopaedic surgery residency programs: a cross-sectional study," *Current Orthopaedic Practice,* vol. 32, no. 5, pp. 500-504, 2021.

[2] D. Chaffey. "Global social media statistics research summary 2023." Smart Insights. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/ (accessed 04 April 2023).

[3] S. Aslam. "Twitter by the Numbers: Stats, Demographics & Fun Facts." https://www.omnicoreagency.com (accessed 04 April 2023).

[4] S. I. Sumer and N. Parilti, *Social Media Analytics in Predicting Consumer Behavior*. CRC Press, 2023.

[5] J. Luo, J. Du, C. Tao, H. Xu, and Y. Zhang, "Exploring temporal suicidal behavior patterns on social media: Insight from Twitter analytics," *Health informatics journal,* vol. 26, no. 2, pp. 738-752, 2020.

[6] C.-w. Shen, M. Chen, and C.-c. Wang, "Analyzing the trend of O2O commerce by bilingual text mining on social media," *Computers in Human Behavior,* vol. 101, pp. 474-483, 2019.

[7] J. Ranjan and C. Foropon, "Big data analytics in building the competitive intelligence of organizations," *International Journal of Information Management,* vol. 56, pp. 1-13, 2021.
[8] F. J. Lacarcel and R. Huete, "Digital communication strategies used by private companies, entrepreneurs, and public entities to attract long-stay tourists: a review," *International Entrepreneurship and Management Journal,* pp. 1-18, 2023.
[9] I. Lee, "Social media analytics for enterprises: Typology, methods, and processes," *Business Horizons,* vol. 61, no. 2, pp. 199-210, 2018.
[10] I. Taleb, M. A. Serhani, and R. Dssouli, "Big data quality: A survey," in *2018 IEEE International Congress on Big Data (BigData Congress)*, 2018: IEEE, pp. 166-173.
[11] W. Elouataoui, I. E. Alaoui, and Y. Gahi, "Data Quality in the Era of Big Data: A Global Review," *Big Data Intelligence for Smart Applications,* pp. 1-25, 2022.
[12] R. Rawat and R. Yadav, "Big data: Big data analysis, issues and challenges and technologies," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1022, no. 1: IOP Publishing, pp. 1-9.
[13] S. Kaisler, J. A. Espinosa, W. Money, and F. Armour, "Big Data and Analytics: Issues and Challenges for the Past and Next Ten Years," pp. 805-814, 2023.
[14] M. Naeem *et al.*, "Trends and future perspective challenges in big data," in *Advances in Intelligent Data Analysis and Applications: Proceeding of the Sixth Euro-China Conference on Intelligent Data Analysis and Applications, 15–18 October 2019, Arad, Romania*, 2022: Springer, pp. 309-325.
[15] M. Henderson, K. Jiang, M. Johnson, and L. Porter, "Measuring Twitter use: validating survey-based measures," *Social Science Computer Review,* vol. 39, no. 6, pp. 1121-1141, 2021.
[16] L. Pilař, L. Kvasničková Stanislavská, R. Kvasnička, P. Bouda, and J. Pitrová, "Framework for Social Media Analysis Based on Hashtag Research," *Applied Sciences,* vol. 11, no. 8, p. 3697, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/8/3697.
[17] L. Pilař, L. Kvasničková Stanislavská, J. Pitrová, I. Krejčí, I. Tichá, and M. Chalupová, "Twitter analysis of global communication in the field of sustainability," *Sustainability,* vol. 11, no. 24, p. 6958, 2019.
[18] L. Kvasničková Stanislavská, L. Pilař, K. Margarisová, and R. Kvasnička, "Corporate social responsibility and social media: Comparison between developing and developed countries," *Sustainability,* vol. 12, no. 13, p. 5255, 2020.
[19] L. Pilař, L. Kvasničková Stanislavská, G. Gresham, J. Poláková, S. Rojík, and R. Petkov, "Questionnaire vs. social media analysis-Case study of organic food," *AGRIS on-line Papers in Economics and Informatics,* vol. 10, no. 665-2019-272, pp. 93-101, 2018.

[20] J. Yang, P. Xiu, L. Sun, L. Ying, and B. Muthu, "Social media data analytics for business decision making system to competitive analysis," *Information Processing & Management,* vol. 59, no. 1, p. 102751, 2022.
[21] M. Imran and A. Ahmad, "Enhancing data quality to mine credible patterns," *Journal of Information Science,* vol. 49, no. 2, pp. 544-564, 2023.
[22] H. Zhang, Z. Zang, H. Zhu, M. I. Uddin, and M. A. Amin, "Big data-assisted social media analytics for business model for business decision making system competitive analysis," *Information Processing & Management,* vol. 59, no. 1, p. 102762, 2022/01/01/ 2022, doi: https://doi.org/10.1016/j.ipm.2021.102762.
[23] P. Kumar and A. Sinha, "Information diffusion modeling and analysis for socially interacting networks," *Social Network Analysis and Mining,* vol. 11, pp. 1-18, 2021.
[24] I. G. García and A. Mateos, "Use of Social Network Analysis for Tax Control in Spain," *Hacienda Publica Espanola,* no. 239, pp. 159-197, 2021.
[25] D. J. Brass, "New developments in social network analysis," *Annual Review of Organizational Psychology and Organizational Behavior,* vol. 9, pp. 225-246, 2022.
[26] R. Gould, *Graph theory*. Courier Corporation, 2012.
[27] A. Majeed and I. Rauf, "Graph theory: A comprehensive survey about graph theory applications in computer science and social networks," *Inventions,* vol. 5, no. 1, p. 10, 2020.
[28] S. P. Borgatti and D. J. Brass, "Centrality: Concepts and measures," *Social networks at work,* pp. 9-22, 2019.
[29] H. Zhu, X. Yang, and J. Wei, "Path prediction of information diffusion based on a topic-oriented relationship strength network," *Information Sciences,* vol. 631, pp. 108-119, 2023.
[30] A. Tsang, B. Wilder, E. Rice, M. Tambe, and Y. Zick, "Group-fairness in influence maximization," *arXiv preprint arXiv:1903.00967,* 2019.
[31] K. Li, L. Zhang, and H. Huang, "Social influence analysis: models, methods, and evaluation," *Elsevier: Engineering*, vol. 4, no. 1, pp. 40-46. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809917308056
[32] M. Azaouzi, W. Mnasri, and L. B. Romdhane, "New trends in influence maximization models," *Computer Science Review,* vol. 40, p. 100393, 2021.
[33] T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang, "How noisy social media text, how diffrnt social media sources?," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 356-364.
[34] L. S. Lai and W. M. To, "Content analysis of social media: A grounded theory approach," *Journal of Electronic Commerce Research,* vol. 16, no. 2, p. 138, 2015.
[35] S. Myneni, N. K. Cobb, and T. Cohen, "Finding meaning in social media: content-based social network analysis of QuitNet to identify new

opportunities for health promotion," in *MEDINFO 2013*: IOS Press, 2013, pp. 807-811.

[36] N. Crossley, "Content and context in social network analysis," in *Networks in the Global World V: Proceedings of NetGloW 2020 5*, 2021: Springer, pp. 3-14.

[37] H. Purohit, Y. Ruan, A. Joshi, S. Parthasarathy, and A. Sheth, "Understanding user-community engagement by multi-faceted features: A case study on twitter," in *WWW 2011 Workshop on Social Media Engagement (SoME)*, 2011.

[38] S. Nepal, W. Sherchan, and C. Paris, "Building trust communities using social trust," in *Advances in User Modeling: UMAP 2011 Workshops, Girona, Spain, July 11-15, 2011, Revised Selected Papers 19*, 2012: Springer, pp. 243-255.

[39] C. Buntain and J. Golbeck, "Automatically identifying fake news in popular twitter threads," in *2017 IEEE international conference on smart cloud (smartCloud)*, 2017: IEEE, pp. 208-215.

[40] M. Mahdavi, M. Asadpour, and S. M. Ghavami, "A comprehensive analysis of tweet content and its impact on popularity," in *2016 8th International Symposium on Telecommunications (IST)*, 2016: IEEE, pp. 559-564.

[41] S. Kong, L. Feng, G. Sun, and K. Luo, "Predicting lifespans of popular tweets in microblog," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 1129-1130.

[42] P. Zola, G. Cola, M. Mazza, and M. Tesconi, "Interaction strength analysis to model retweet cascade graphs," *Applied Sciences,* vol. 10, no. 23, p. 8394, 2020.

[43] C. Salvatore, S. Biffignandi, and A. Bianchi, "Social media and twitter data quality for new social indicators," *Social Indicators Research,* vol. 156, pp. 601-630, 2021.

[44] R. Pozzar *et al.*, "Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire," *Journal of medical Internet research,* vol. 22, no. 10, p. e23021, 2020.

[45] F. A. Batarseh and A. Kulkarni, "Context-driven data mining through bias removal and data incompleteness mitigation," *arXiv preprint arXiv:1910.08670,* 2019.

[46] A. N. Islam, S. Laato, S. Talukder, and E. Sutinen, "Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective," *Technological forecasting and social change,* vol. 159, p. 120201, 2020.

[47] J. Li, Q. Xu, R. Cuomo, V. Purushothaman, and T. Mackey, "Data mining and content analysis of the Chinese social media platform Weibo during the early COVID-19 outbreak: retrospective observational infoveillance study," *JMIR Public Health and Surveillance,* vol. 6, no. 2, p. e18700, 2020.

[48] Twitter Inc. . "Rate limits: Standard v1.1 Twitter Developer Platform " Twitter, Inc. https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits (accessed 14 April 2023).
[49] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data,* vol. 8, no. 3, pp. 171-188, 2020.
[50] P. Koukaras and C. Tjortjis, "Social media analytics, types and methodology," *Machine Learning Paradigms: Applications of Learning and Analytics in Intelligent Systems,* pp. 401-427, 2019.
[51] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks," *Expert Systems with Applications,* vol. 164, p. 114006, 2021.
[52] D. Henry, "Twiscraper: a collaborative project to enhance twitter data collection," in *Proceedings of the 14th ACM international conference on web search and data mining*, 2021, pp. 886-889.
[53] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," *Journal of Big Data,* vol. 6, no. 1, pp. 1-16, 2019.
[54] N. A. Ghani, S. Hamid, I. A. T. Hashem, and E. Ahmed, "Social media big data analytics: A survey," *Computers in Human Behavior,* vol. 101, pp. 417-428, 2019.
[55] S. Stier, J. Breuer, P. Siegers, and K. Thorson, "Integrating survey data and digital trace data: Key issues in developing an emerging field," vol. 38, ed: SAGE Publications Sage CA: Los Angeles, CA, 2020, pp. 503-516.
[56] C. Fuchs, "Social media: A critical introduction," *Social Media,* pp. 1-440, 2021.
[57] P. Martí, L. Serrano-Estrada, and A. Nolasco-Cirugeda, "Social media data: Challenges, opportunities and limitations in urban studies," *Computers, Environment and Urban Systems,* vol. 74, pp. 161-174, 2019.
[58] A. Ghahramani and M. Prokofieva, "Visualisation for social media analytics: landscape of R packages," in *2021 25th International Conference Information Visualisation (IV)*, 2021: IEEE, pp. 218-222.
[59] J. Lowe and M. Matthee, "Requirements of data visualisation tools to analyse big data: A structured literature review," in *Responsible Design, Implementation and Use of Information and Communication Technology: 19th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2020, Skukuza, South Africa, April 6–8, 2020, Proceedings, Part I 19*, 2020: Springer, pp. 469-480.