

Assessing the quality of social media data: a systematic literature review

Oumaima Reda, Ahmed Zellou

Department of Computing Science, National Higher School of Computer Science and Systems Analysis, Mohammed V University, Rabat, Morocco

Article Info

Article history:

Received Aug 19, 2022

Revised Oct 6, 2022

Accepted Oct 24, 2022

Keywords:

Data quality

Quality assessment

Quality metrics

Social media

Systematic literature review

ABSTRACT

In recent years, social media have been at the heart of new communication technologies. They are no longer used only to facilitate interaction between family, friends, and professional relationships, but tend to become an increasingly used communication channel to address public opinion. Research involving data sources from social media are relatively a recent and expanding area of research, nevertheless, the literature remains limited regarding the complex issue of how to assess and ensure the quality of social media data significantly and adequately. Our goal in this study is to provide a clearer and deeper understanding and a comprehensive overview of the existing state of research pertaining to the assessment of data quality in social media context. We performed a systematic literature review (SLR) on the quality of social media data to collect, analyze, and discuss data on the accuracy and value of prior literature that has focused on this area, has addressed a variety of topics, and has been published between 2016 and 2021. We followed a predefined review process to cover all relevant research papers published during this period. Our results demonstrate and strengthen the significance and the importance of data quality especially in the context of social media.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Oumaima Reda

Department of Computing Science, National Higher School of Computer Science and Systems Analysis

Mohammed V University

Rabat, Morocco

Email: oumaima_reda@um5.ac.ma

1. INTRODUCTION

Every data analysis process relies heavily on data, it is a competent task which is useful to the extent that it can be quickly interpreted to reveal valuable information and extract the deep insights hidden in the data. Nowadays, social media have been at the heart of new communication technologies. They are no longer used only to facilitate the interaction between people, but used for all kinds of activities, such as exchanging experiences, posting information, expressing feelings and opinions, bringing people together and advertising a company [1]. Social media data can be analyzed to gain insights into either consumer or business behavior. Hence, it may require filtering before meaningful information can be obtained, as the data to be analyzed may contain false data or come from a spam campaign [2]. For instance, if we are looking for data published about a particular event, we may end up with irrelevant and unnecessary data, for this reason filtering all data is a required step [1]. The current literature demonstrates the use of social media in a variety of fields, including medicine, science, and economics, as well as the social sciences. During data collection, users need to be sure

about the reliability and the source of the data, discarding any suspect ones. Users therefore want to be certain that the data's quality and relevance are adequate for the particular context while processing and evaluating the data. As a consequence, the usefulness and dependability of the data improve decision-making in a variety of ways [3].

Data quality is a current topic of great attention, thus, a holistic understanding of data quality is the cornerstone of social media measurement studies. Therefore, the purpose of this study is to review the existing literature to identify the dimensions of social media data quality, the data quality issues relating to these dimensions, and the methodologies employed to analyze and measure these dimensions. By giving a clear insight into research topics related to social media data quality, our goal is to determine which aspects have been investigated and to identify any potential areas for further research. We then propose three categorizations of the work by first, identifying the manifestations of data quality issues in social media. Second, the main dimensions and metrics of data quality used in the social media context and finally summarizing the methods used to analyze and measure quality. The results of this study will also be a starting point to outline and measure the dimensions of social media data quality and to identify issues that need to be addressed to improve the quality of social media data.

The rest of this research paper is structured as follows: we start by outlining the context and preliminary, with an overview of the concepts of data quality and social media in section 2. Our research objectives are listed in section 3 after that. In section 4, we describe the research methodology and explain the systematic review's following specific steps. Section 5 summarizes and reports the results. While section 6 discusses the limits of this assessment and suggests further research avenues. Finally, section 7 presents our conclusion and future research.

2. BACKGROUND

Data quality is considered one of the most crucial parts for any decision-making, especially when such decisions have a serious influence. This section provides an overview of the main research fields related to our systematic literature review, namely data quality, and social media.

2.1. Concepts: data quality, quality dimensions and quality assessment

Generally, the notion of quality has been described as all the characteristics or attributes of a product that define its capacity to satisfy explicit or implicit needs [4]. However, several efforts were made to provide a clear definition of data quality. In the literature, some have characterized data quality as the extent to which data are useful and suitable for use in a specific task or context [5], while [6] define it as the ability to be used and processed by data consumers. As these definitions remain typically qualitative, various dimensions which represent a single aspect of quality, have been proposed by some researchers to measure and describe data quality. Hence, before selecting appropriate dimensions or categories, it is first required to specify what the desired entity is and what the data quality problem is. Then, once this has been determined, all dimensions can be quantified in the next step in order to provide information about the quality of the entity [4].

Commonly, the different dimensions of data quality are defined qualitatively, with reference to general characteristics of the data. On the other hand, the corresponding definitions usually do not provide quantitative measures, but one or more specific metrics must be assigned to the dimensions as separate and distinct properties [4]. According to Batini *et al.* [7], data quality dimensions can be gathered into clusters depending on their similarity with respect to their ability to capture an aspect of data quality. However, Wang and Strong [6] described it as a collection of data quality attributes representing a unique aspect or construct of data quality.

The focus of data quality researchers is often directed at different aspects of data quality management, as managing data quality usually includes selecting and classifying these quality dimensions in order to further categorize and develop a quality model. Which is considered as a collection of data quality dimensions or characteristics by which we can assess and evaluate the quality of our data to assess its relevance and to rate it [8]. When analyzing the quality of data with meaning, the context and intended use of the data are taken into account, this is known as data quality assessment.

2.2. Social media

Social media has been defined in multiple ways, it is not defined by a single kind of platform or data, it is a set of web-based systems that enable massive interaction, conversation and sharing messages and contents (images, videos, and articles) between members of a network [9]. Social media platforms are now playing a

bigger role in internet users' personal and professional lives, despite the fact that they were once dismissed as a fad. Indeed, people are more and more using social media to get first-hand news and information [10]. Twitter and Facebook are some examples of the most popular social media websites which serve a wide user population who are contributing and consuming content [11].

Data from social media platforms captures a wide range of information and is presented in a variety of formats, with varying access methods and levels of availability. Because of their huge amount of data and timeliness, social media platforms face multiple issues, and most importantly quality issues [10]. Social media data may be strictly textual or include audio or visual elements. Social media sites' content can be retrieved directly from the platform itself or using a number of partially or fully automated techniques [9]. Popular platforms are numerous and subject to quick change, we can classify the types of social media according to three categories, media sharing such as YouTube, DailyMotion, Instagram, Snapchat, and Flickr, professionals like LinkedIn and Viadeo, and generalists such as Facebook, Twitter, and MySpace [12].

3. RESEARCH OBJECTIVES

To the best of our knowledge, no prior work has made an attempt to provide a comprehensive review of the body of research on social media data quality using a systematic approach. In order to do this, we plan to first determine and assess the findings of all relevant and high quality papers dealing with data quality in social media context using a rigorous interpretation and evaluation approach, and then perform a comprehensive survey of all available relevant evidence on quality in social media. To this end, the aim of this study is to review the existing literature to identify the quality dimensions and metrics used in social media context, the manifestations of data quality issues in social media, and the methods used to analyze and measure the quality. Therefore, in addition to serving as a springboard for defining and measuring the quality dimensions required to assess the quality of social media data, our findings are anticipated to give us, researchers and practitioners, ideas for future research on data quality assessment. They will also provide a starting point for identifying problem areas that need to be addressed in order to improve the quality of our data.

4. RESEARCH METHOD

To conduct our systematic literature review (SLR) of data quality in social media, we follow the original SLR guidelines proposed by [13], which consists of 4 steps; i) identify the scope of the review, ii) search for the initial list of papers, iii) select all relevant papers, and iv) analysis and process of the data.

4.1. Identify the scope of the review

This step include 4 main tasks; i) specifying which period of time to take into consideration, ii) identifying the appropriate search strings, iii) selecting databases and establishing the criteria for inclusion, and iv) exclusion to be used for selecting studies.

4.1.1. Research questions

The aim of this SLR is the review of the current approaches or methods that are utilized or proposed by the research community for data quality assessment in social media, in order to obtain a state of the art in this area. Consequently, we have developed the following research questions:

- RQ1 : what are the main manifestations of data quality issues in social media context?
- RQ2: what are the most important dimensions and metrics for assessing social media data quality?
- RQ3: what methodologies have been used to measure social media data quality?

4.1.2. Range time and Search string

Our review was constrained in our study to only cover papers published between 2016 and 2021 because we had our search strategy delimited to include recent studies.

With the use of these well chosen search strings, we are able to access a comprehensive collection of articles on data quality in relation to social media, drawing from a wide range of fields and domains. Thus, the following is the search string we are using.

("data quality") AND ("quality dimensions" OR "quality assessment") AND ("social media")

4.1.3. Databases

The online search have been then applied to the list of digital libraries that is presented in Table 1. We based our selection on these digital libraries. Because they include all high-quality journals and proceedings of conferences.

Table 1. Electronic data sources

Digital libraries	URL
DBLP	http://dblp.uni-trier.de
ACM digital library	http://dl.acm.org
Science direct	https://www.sciencedirect.com/
Scopus	http://www.scopus.com

4.1.4. Inclusion/exclusion criteria

The following Table 2 provides a description of the inclusion and exclusion criteria used in our review, and accordingly we decide whether or not to select each study for further analysis.

Table 2. The inclusion and exclusion criteria used in our study

Inclusion criteria	Exclusion criteria
IC1: studies published between 2016 and 2021	EC1: studies that are duplicates
IC2: studies published in English	EC2: studies that are not accessible online
IC3: full papers	EC3: Master's and Doctoral dissertations, tutorials, editorials and magazines
IC4: studies focused on data quality in social media context	EC4: studies that were not peer-reviewed

4.2. Search for the initial list of papers

Following the initial search on the four digital libraries. The resulting collection of publications have been filtered on the basis of years of publication, abstracts, titles, and keywords. We then used the set of inclusion/exclusion criteria to determine whether or not to select a study for further processing.

4.3. Select relevant papers

In order to demonstrate and strengthen the relevance of our study, the quality of each publication was evaluated after reading the full text of the selected articles. Thus, a quality assessment checklist is provided which can be used to determine the importance of each article and whether it includes the requested information. Thus, our quality assessment checklist is presented in Table 3, we set five quality criteria (QC) whose value is either yes or no (i.e. 1 or 0) to calculate and evaluate at last the score of each paper.

Table 3. Quality assessment criteria

Quality criteria	Score
QC1: focus on data quality	Yes/No (1/0)
QC2: focus on social media	Yes/No (1/0)
QC3: DQ dimensions defined	Yes/No (1/0)
QC4: DQ metrics defined	Yes/No (1/0)
QC5: manifestations of social media DQ issues identified	Yes/No (1/0)

4.4. Analysis and process of the data

We reviewed and analyzed the findings and insights contained in each selected paper in relation to our research questions. However, not all articles answered all of our questions. Therefore, we extracted and stored the relevant answers from the papers including; i) title and authors' names, ii) publication year, iii) publication type, iv) the findings and outcomes : the contribution, and v) quality assessment.

5. RESULT

This section presents the findings of our SLR using the research methodology detailed in the aforementioned section 4. We discuss then the overview of selected papers and also the analysis of the results.

5.1. Overview of selected studies

A set of 209 papers obtained from the electronic data sources searches using the search string. A hand sort was done on the titles, abstracts, and keywords to discard duplicated studies, resulting in a set of 123 articles that qualified as relevant. Next, our inclusions and exclusion criteria were used, 10 papers were not from journals or conferences, 5 papers were written in a language other than English, 17 paper was not full paper and 19 papers are not accessible online. After full-text review, we further discarded 35 papers, as a result, 37 papers identified and are ready to be filtered by the application of our 5 quality assessment criteria presented in Table 3 in order to guarantee that the findings contained will be a useful addition to our SLR. Finally, 15 papers that scored less then 2 were, therefore, discarded and a total of 22 papers remained eligible to address our research question. The following Figure 1 shows the full overview of our selection process.

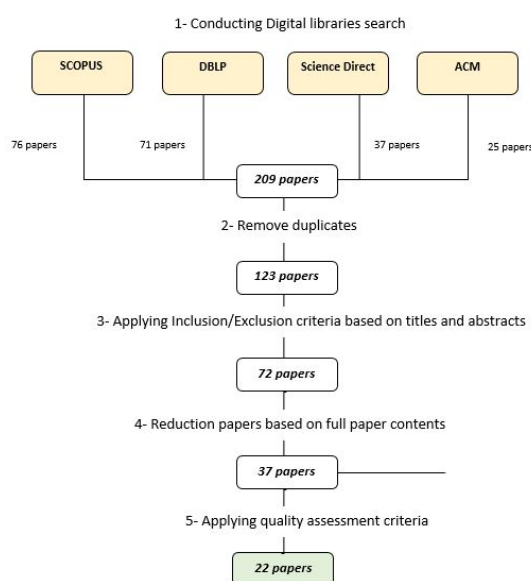


Figure 1. Summary of the selection process for the final papers

5.2. Classification of selected studies

We next analyzed and evaluated the results to generate an insight into the current state of the literature in the area of social media data quality. In this phase, we investigated the distribution of our selected studies by year of publication and data sources. Figure 2 presents the distribution of 22 selected publications between 2016 and 2021 by years in Figure 2(a) and data sources in Figure 2(b). Our findings show that the majority of the papers, i.e., 65%, are published in the last three years of the explored time range, i.e., between 2019 and 2021. However, the distribution of the papers according to data sources shows that the highest number of selected papers, with 32% were published in Scopus. The results for 2021 cannot be conclusive since the research was conducted in early 2021.

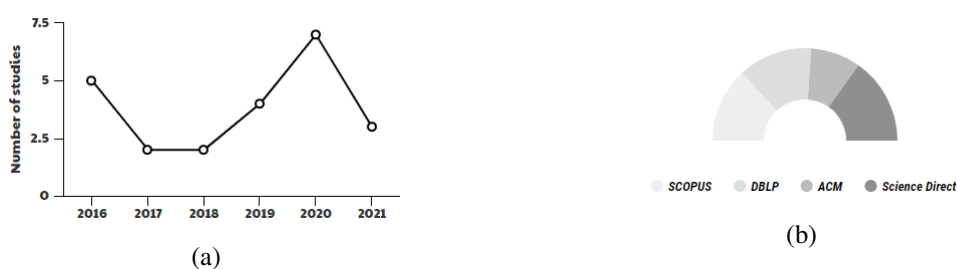


Figure 2. The distribution of the selected studies by (a) year of publication and (b) data sources of publication

5.3. Data extraction results

We highlight the research question addressed in section 4 by reporting the findings for each RQ from our systematic literature review and discussing them in this section.

5.3.1. RQ1 : What are the main manifestations of data quality issues in social media context?

With the huge popularity of social media, undesirable users are multiplying to spread content considered as spam, for example ads and phishing websites [14]. This distribution can cause major problems including; i) violating users' privacy, ii) pollute search results, and iii) degrading the accuracy of statistics obtained through information retrieval tools. The easy-to-use interfaces and the low security limits on publishing do not help to maintain a consistent level of data quality. These characteristics have made online social media vulnerable to various attacks by a certain type of malicious users [3]. According to Patone and Zhang [15], a major concern about the use of social media for research is the non-representativeness of data, and the measurement. For example, when choosing accounts to be studied based on account metadata, such as residency, one may encounter errors in case the recorded information is not updated when there has been a change in status. This type of error can directly affect the accounts selected for the study, i.e. the representation dimension of data quality. According to Aquino *et al.* [16], the first challenge of social data quality is the difference that social data represents depending on the social media from which it was created. This difference can be in the type of social interaction i.e. the type of platform or the type of data i.e. whether it is structured or not. This means that social data must be analyzed differently, depending on the social platform from which the message originated. Research by Popescu *et al.* [17] investigated a number of issues with data quality between two representations, social media and enterprise resource planning (ERP). Following that, they classified them into three potential quality deficiencies, as indicated in Table 4.

Table 4. Possible data quality deficiencies

DQ deficiencies	Possible DQ problems
Ambiguous representation	The quality data is not clear Data with diverse and imprecise meanings
Incomplete representation	Missing the reviews (or comments) on a product Missing data (description) regarding the product
Meaningless representation	Data in a wrong format A product description is related to another product

Social media is one of the volunteered geographic information (VGI) data sources that raises a number of concerns with regard to the use of specific quality measures. For instance, for usability dimension, it might be quite challenging to explore social media data to look for patterns that can be used, because it is harvested rather than crowdsourced. In spite of the potential absence of temporal metadata in the contributions, the temporal quality of social media VGI data is often good when compared to other data. Moreover, the logical consistency in the lack of precise guidelines. In addition to positional accuracy, considering that few contributions actually include a spatial reference. In terms of thematic accuracy, there is a lack of clear product standards and user characteristics, which leads to a thematic heterogeneity. Finally, completeness is also important because the location of the major users and the socioeconomic backgrounds of the contributors frequently alter the distribution of user contributions [18]. According to Hajjar *et al.* [19], there are some issues that reside in two important features in LinkedIn. For "skills and expertise" feature, the unstructured textual data like abbreviations and contextual data. The challenges related to this issue are addressed by using a standardized corpus that has been subject to some classification. As a result, this increases the correctness and reliability of the data. Regarding the other feature, "people you may know", one of the main problems with this feature is the scalability of the match data, the correspondence between a user and the other members.

5.3.2. RQ2: what are the most important dimensions and metrics to assess social media data quality?

This study examined and analyzed the quality dimensions and metrics used for assessing social media data quality and some manifestations of data quality issues disclosed in the included studies. Recently, Koumtingue [11] focused particularly on Twitter as a social media platform of reference and discussed the quality of social media data, they presented a comprehensive quality framework tailored to the analysis of Twitter data and a reformulation of the traditional dimensions of quality is provided, as well as they described new aspects of quality such as availability, reliability presentation, usability, relevance. Furthermore, Immonen *et al.* [20] addressed the issue of evaluating the quality and managing the value of social media data in each data

processing phase of the big data architecture by using quality metadata and quality policies. They provided a model of six data quality dimensions with their assessment metrics and their potential applicability in the case of Twitter. However, Pääkkönen and Jokitulppo [2] extended the previous work of Immonen *et al.* [20] and focused on the evaluation of data quality management architecture for social media data. They addressed in particular reference architecture (RA) design to facilitate design of quality management aspects into implementation architectures of big data systems. A tool has been implemented which enables assessing, filtering and querying of tweet-related quality information based on the rules defined by users. Similarly, Arroyo *et al.* [1] have designed a visual interface that could help to measure the quality of data generated from the three selected social networks by the users about a specific event. In order to develop the theoretical application, they choose to focus on the five data quality dimensions proposed by [20]: credibility, corroboration, popularity, opportunity, and relevance.

The purpose of [21] is to examine quality of short texts presented as posts, interpretations and contributions from the online communities *genius* and *stack overflow*. The trust model, based on its metadata, classified these short texts into four levels of trust: trusted, very trusted, untrusted, and extremely untrusted. Patone and Zhang [15] delineated two existing approaches, one-phase and two-phase approach to data analysis based on social media data have been identified in the literature. For data quality assessment, they demonstrated that applicable total error frameworks developed in terms of representation and measurement of general statistical data can be used. Hatimi *et al.* [18] has performed a literature review that supported the need to use social media platforms as a source of VGI in the context of risk management applications (RMA). In addition, they selected six relevant quality dimensions proposed by ISO 19157:2013 to be used in order to assess the quality of gathered VGI data which are thematic accuracy, positional accuracy, completeness, temporal quality, logical consistency and trustworthiness indicators which include reliability and reputation.

A different evaluation approach is suggested by [22], which aimed to assess the quality of data available on YouTube videos regarding the side effects of biologic therapy. The quality and reliability of the videos was assessed according to the global quality score (GQS) and discern score. Hajjar *et al.* [19] discussed the data quality issues related to data collected on social media. They choose for each social media platform, the techniques used for data capture, analysis and processing with the satisfied data quality model. The purpose of [23] is to address the issues of measurement and representativeness of data using the results of a study in which social media advertisements is used for a hard-to-reach sexual and gender minority youth population. They examine measures of data quality such as exiting the survey before completion, commonly known as "break-off", in addition to using the nonsubstantive responses to questionnaire questions such as "prefer not to answer" and "don't know". However, Kim *et al.* [24] presented a cohort study and investigated how students use and evaluate social media data to help enhancing the understanding of the changes in students' social media data use and evaluation behavior. Looking into further studies that focused specifically on assessing social media data quality. Tilly *et al.* [25] proposed a new model of data quality in social media that can explain the interplay of existing conceptualizations and provides a comparative analysis of these conceptualization. Research by Aquino *et al.* [16], focused on the issue related to data quality when dealing with social and sensor data, and present a framework for social and sensor data quality standardizing whose objective is to evaluate and control data quality aspects using two components: social and sensor DQ component, each component is composed by two subcomponents: DQ assessment and DQ enhancement sub-components. In the field of education, we have the work of [26] regarding how the human factor greatly impacts the quality of data collected from social media. They developed a social media evaluation tool, WeSQu, in order to help evaluators paying attention to the critical factors of social media service quality. The WeSQu tool includes the following dimensions: reliability, accessibility, security and privacy, supporting navigation, text readability, data presentation, and user motivation.

About credibility, Abbasi and Liu [27] focused only on one dimension; the credibility of users in social media. They start by investigating the situations in which the credibility of the content or of the user cannot be assessed based on the user's profile. Then, a CredRank algorithm was proposed to measure user credibility and analyze the online behavior of social media users. ODonovan *et al.* [28] presented an analysis of the distribution of the individual features in Twitter such as hashtags, retweets and mentions that can be used to find interesting, newsworthy and credible information. We summarize the findings of this research question in the below Table 5, in which we represent the duality dimensions and metrics used in these studies.

Table 5. A summary of DQ dimensions and metrics used in our review

DQD	Proposed metrics	Studies
Accessibility		[10], [19],[26]
Accuracy		[10], [19], [21]
Auditability		[10]
Authorization		[10]
Availability		[10], [19]
Believability	Evaluated based on creation-date, status-count, followers-count and friends-count, and is-verified	[20]
Backup		[19]
Balanced		[22], [21]
Comparability		[15]
Completeness		[10], [18]
Conformance		[25]
Correctness		[19]
Consistency		[10], [19]
Correspondence		[25]
Corroboration	Evaluated based on publication's list, all comments in all publications, the volume of data sets analyzed based on which the identified problem is recognized	[1], [20]
Coverage		[15]
Credibility	Evaluated based on registration date, social context and is a verified account?	[1], [10], [27], [28]
Definition		[10]
Efficiency		[19]
Rapidity		[19]
Fitness		[10], [25]
Identification		[15]
Integrity		[10]
Logical consistency		[18]
Mapping		[15]
Metadata		[10]
Navigation		[26]
Opportunity	Evaluated based on date of all publications	[1]
Organizational		[25]
Perceived		[25]
Performance		[19]
Popularity	Evaluated based on how many people retweet, friend, comment, like, follow, and read a tweet	[1], [2], [20]
Positional accuracy		[18]
Precision		[21]
Presentation		[10], [26]
Protection		[19]
Readability		[10], [26]
Reference		[22]
Reliability		[10], [19], [26]
Scalability		[19]
Security		[19], [26]
Semiotic		[25]
Sensitivity		[21]
Specificity		[21]
Relevancy	Evaluated based on comparing the distance cosine between the words in a tweet, the context, the list of comments and the number of occurrences of relevant keywords	[1], [2], [10], [15], [20]
Structure		[10]
Sustainability		[19]
Temporal quality		[18]
Timeliness	Evaluated based on timestamp of a tweet (creation date)	[10], [19], [20]
Thematic Accuracy		[18]
Trustworthiness		[18]
Uncertainty		[22]
Understandability		[19], [22]
Unit		[15]
Usefulness		[11], [19]
Validity		[20], [22]

5.3.3. RQ3: what methodologies have been used to measure social media data quality?

We used a similar analysis process to determine the methods and techniques conducted to measure data quality in social media. A wide range of methods for assessing and improving data quality are available in the literature. Due to the diversity and complexity of these methods, research has recently been focused on developing methodologies for choosing and implementing data quality assessment and improvement techniques. Research by Berlanga *et al.* [29], which proposed a new methodology whose goal is to determine a valid quality indicator as a metric in order to assess and analyze the overall quality of a collection of posts and user profiles from various perspectives, as well as to include the measures obtained by multiple QC used to filter the relevant posts for a social business intelligence project. Abbasi and Liu [27] proposed a method for measuring user credibility in social media by means of a CredRank algorithm which consist of detecting and clustering users who are dependent and assigning weight for each cluster based on its size. This CredRank algorithm looks for users who exhibit similar conduct and groups them together. The similarity between users can be calculated with the following, where $B(u_i, t)$ is user u_i 's behavior in timestamp t and the function $\sigma(B(u_i, t), B(u_j, t))$ that measures the similarity of two users' behavior in the given timestamp t .

$$Sim(u_i, u_j) = \frac{1}{t_n - t_0} \sum_{t=t_0}^{t_n} \sigma(B(u_i, t), B(u_j, t)) \quad (1)$$

Next, apply weights to the clusters using the following formula, where w_{C_i} represents the weight given to the cluster C_i with $|C_i|$ members. The indicated value reveals how much credibility the member has.

$$w_{C_i} = \frac{\sqrt{|C_i|}}{\sum_j \sqrt{|C_j|}} \quad (2)$$

Moreover, Hajjar *et al.* [19] discussed, evaluated and then compared the existing techniques used for data capture, analysis and processing. They presented a framework of techniques for assessing the quality of big data that is suitable for several social media according to different needs and specifications. The Table 6 introduces some analytic techniques used to identify the quality of social media data.

Tanvir *et al.* [30] suggested a model for recognizing and detecting fake news messages extracted from twitter posts. A deep learning-based model is proposed to identify whether a news is good or fake, and five different types of machine learning algorithms are used in this study to see how well the data fit into the model, such as Bayesian model, recurrent neural network (RNN), long short-term memory, logistic regression and also support vector machine. Equally, the aim of the work presented by [31] was to present and evaluate two types of classification methods namely logistic regression and fuzzy logic to assess the quality of crowdsourced social media data retrieved from the public Twitter archive during flood events in Thailand. Another kind of assessment is implemented by [22] who used questionnaires to asses the quality and the reliability of data in videos posted in youtube, where they assign scores on each question answered. The discern score presented in the following Figure 3. Last but not least, Alsyouf *et al.* [32] published a case study to assess the veracity of the most widely read articles about genitourinary malignancies on four social media sites (Facebook, Twitter, Pinterest, and Reddit) and to determine the frequency of false information that is available to patients. BuzzSumo, a social media research tool that enables users to apply keywords to search for article links on the most well-liked social media platforms, was used to find these stories.

Table 6. Quality of social media data analytic techniques [19]

SM platforms	SM techniques
Facebook	FQL, Netvizz
Twitter	DataSift, Gnip, Topsy
Flickr	NodeXL
LinkedIn	Text mining, DataFu Pig, DataFu Hourglass

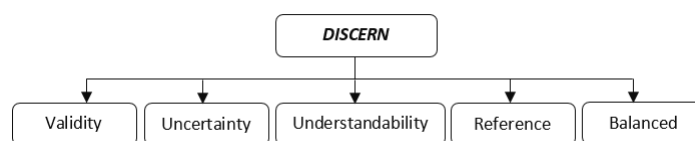


Figure 3. The discern used to assess the quality and reliability of video

6. DISCUSSION

In this section, we outline and discuss the findings and the results in RQ1, RQ2, and RQ3 to depict the current landscape of social media data quality research. Next, we highlight the limitations of our systematic review of the literature. Given an environment as open and out of control like many social media platforms, it is difficult to find the appropriate data. There is a lot of issues related to this data, that we aimed to identify in the first research question. Among 22 reviewed papers, there are 8 studies that addressed the issues and possible deficiencies of quality of data received from social media. Firstly, the issue of undesirable users called spammers, which can violate users' privacy, pollute search results and degrade the accuracy of statistics obtained through information retrieval tools. On Twitter, spammers take advantage of various services provided to launch their spam attacks via URLs, hashtags and mention services. Anti-spam mechanisms are proving to be insufficient to stop the spam problem, which raises a real concerns about the quality of data collections being vacuumed. Then, the simple usability of the interfaces and the low level of security in the publications do not contribute to keep a stable level of data quality. These characteristics have made social media vulnerable to various attacks by a certain type of malicious users. One of the main characteristics of social media is their dependence on users as the primary contributors in generating and publishing content. This dependency on user contributions could be exploited in positive ways, including understanding user needs for marketing purposes, studying user opinions, and improving information retrieval algorithms.

In this systematic review, we aimed to identify the different data quality models currently used in social media to define the quality dimensions and metrics commonly suggested by researchers. Within these different dimensions of data quality, there are a number of measures of quality that can be used to assess data quality. Therefore, the majority of research carried out between 2016 and 2021 has concentrated on the data quality model. As a result, there are almost 58 quality dimensions provided in RQ2. As can be seen in the next Figure 4, based on the reviewed papers, the most dimensions used to assess social media data quality are namely accessibility, reliability, accuracy and relevancy.

Having discussed the different methods used to analyze and measure social media data quality, several studies now look into the possibility of validating these methods. Various techniques also designed to process the quality of data such as; fuzzy logic, data mining, and machine learning. However, RQ3 addresses new methods that might be useful for assessing the quality of social media data.

Although we systematically followed a research and review strategy by using the guidelines proposed by [13], however, some research might not have been considered in our data collection for some reasons. Firstly, our final review process was restricted to 4 particular electronic sources; Scopus, Science direct, DBPL and ACM, and the use of a limited set of keywords. There may be potentially some papers, for example in different languages, which are not listed in the databases or identified using our keywords. Secondly, in our search we only focused on empirical studies on social media data quality, therefore we could have underestimated the current state of research on data quality.

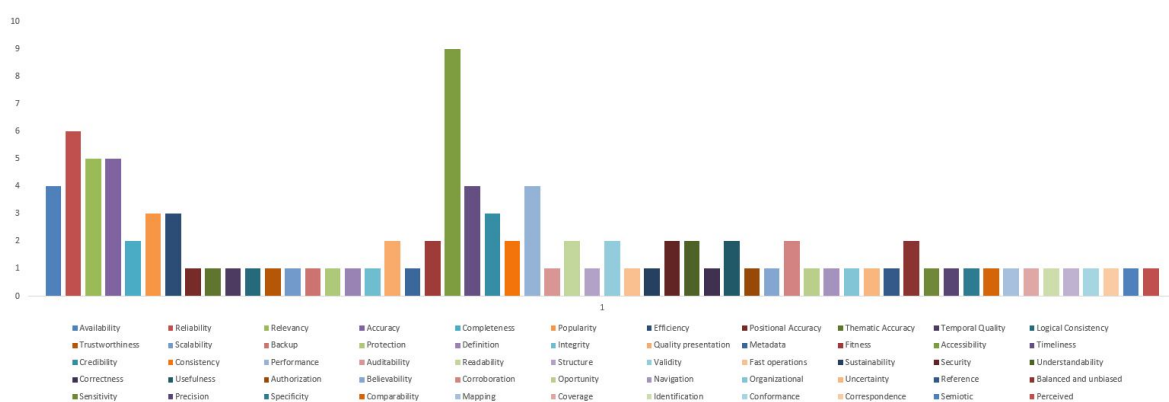


Figure 4. Data quality dimensions used in our study

7. CONCLUSION

In order to measure data quality in the context of social media, this research conducts an SLR of empirical experiments. Using data gathered from 22 research articles released between 2016 and 2021, our SLR addressed three key research questions. The current work offers a novel overview that categorizes various social media data quality issues, quality dimensions, and data analysis and measurement methods. The findings recommend future research directions including creating guidelines for defining particular dimensions of data quality and frameworks for measuring data quality in social media, as well as addressing data quality problems. As part of future efforts, it is intended to broaden quality evaluation to incorporate more data quality dimensions, as it is currently only confined to a few ones. Focusing on a particular platform or data source, creating a new quality evaluation model, then selecting and proposing metrics to evaluate them, is one of our top priorities.




REFERENCES

- [1] A. S. Arroyo, T. Onorati, and P. Diaz, "Quality assessment of social media: lessons learnt from the literature," in *2018 22nd International Conference Information Visualisation (IV)*, 2018, pp. 278–283, doi: 10.1109/IV.2018.00055.
- [2] P. Pääkkönen and J. Jokitalo, "Quality management architecture for social media data," *Journal of Big Data*, vol. 4, no. 6, pp. 1–26, 2017, doi: 10.1186/s40537-017-0066-7.
- [3] M. Washha, A. Qaroush, M. Mezghani, and F. Sedes, "Information quality in social networks: a collaborative method for detecting spam tweets in trending topics," in *International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*, 2017, pp. 211–223, doi: 10.1007/978-3-319-60045-1-24.
- [4] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, "On the meaningfulness of 'Big Data Quality' (invited paper)," *Data Science and Engineering*, vol. 1, no. 1, pp. 6–20, 2016, doi:10.1007/s41019-015-0004-7
- [5] J. M. Juran, "How to think about quality," in *Jurans Quality Handbook*, Fifth edit, J. M. Juran and A. B. Godfrey, Eds. New York, USA: McGraw-Hill, 1999, pp. 1–18, doi: 10.2307/1290459
- [6] R. Y. Wang and D. M. Strong, "Beyond accuracy: what data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, Mar. 1996, doi: 10.1080/07421222.1996.11518099.
- [7] C. Batini, M. Palmonari, and G. Viscusi, "The many faces of information and their impact on information quality," in *International Conference on Information Quality (ICIQ)*, 2012, pp. 212–228.
- [8] O. Reda, I. Sassi, A. Zellou, and S. Anter, "Towards a data quality assessment in big data," in *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*, 2020, pp. 1–6, doi: 10.1145/3419604.
- [9] J. Murphy et al., "Social media in public opinion research: executive summary of the aapor task force on emerging technologies in public opinion research," *Public Opinion Quarterly*, vol. 78, no. 4, pp. 788–794, 2014, doi: 10.1093/poq/nfu053.
- [10] C. Salvatore, S. Biffignandi, and A. Bianchi, "Social Media and twitter data quality for new social indicators," *Social Indicators Research*, vol. 156, no. 2–3, pp. 601–630, 2021, doi: 10.1007/s11205-020-02296-w.
- [11] K. Chai, V. Potdar, and T. Dillon, "Content quality assessment related frameworks for social media," in *Computational Science and Its Applications-ICCSA 2009*, Berlin, Heidelberg: Springer, 2009, pp. 791–805.
- [12] A. Koumtingue, "Exploitation of social media data for analysis of epidemiological spread" (in French: Exploitation des données des réseaux sociaux pour une analyse de la propagation épidémiologique), Université de Sherbrooke, 2017.
- [13] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and Software Technology*, vol. 55, no. 12, pp. 2049–2075, 2013, doi: 10.1016/j.infsof.2013.07.010.
- [14] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," *Collaboration, Electronic Messaging, Anti-Abuse, and Spam Conference (CEAS)*, vol. 6, pp. 1–10, 2010.
- [15] M. Patone and L. Zhang, "On Two Existing Approaches to Statistical Analysis of Social Media Data," *International Statistical Review*, vol. 89, no. 1, pp. 54–71, 2020, doi: 10.1111/insr.12404.
- [16] G. R. C. D. Aquino, C. M. D. Farias, and L. Pirmez, "Data quality assessment and enhancement on social and sensor data," *CEUR Workshop Proceedings*, vol. 2247, pp. 1–7, 2018.
- [17] M. A. -M. Popescu, M. Ge, and M. Helfert, "The Social Media Perception and Reality–Possible Data Quality Deficiencies between Social Media and ERP," in *Proceedings of the 20th International Conference on Enterprise Information Systems*, 2018, pp. 198–204, doi: 10.5220/0006788801980204.
- [18] B. E. Hatimi, H. J. Oulidi, and A. Fadil, "Quality assessment in volunteered geographic information for risk management applications," in *2020 IEEE International conference of Moroccan Geomatics (Morgeo)*, 2020, pp. 1–4, doi: 10.1109/Morgeo49228.2020.9121919.
- [19] D. A. -Hajjar, N. Jaafar, M. A. -Jadaan, and R. Alnutaifi, "Framework for social media big data quality analysis," in *New Trends in Database and Information Systems II*, Cham: Springer, 2015, pp. 301–314, doi: 10.1007/978-3-319-10518-5_23
- [20] A. Immonen, P. Paakkonen, and E. Ovaska, "Evaluating the quality of social media data in big data architecture," *IEEE Access*, vol. 3, pp. 2028–2043, 2015, doi: 10.1109/ACCESS.2015.2490723.
- [21] J. A. Qundus, A. Paschke, S. Gupta, A. M. Alzoubi, and M. Yousef, "Exploring the impact of short-text complexity and structure on its quality in social media," *Journal of Enterprise Information Management*, vol. 33, no. 6, pp. 1443–1466, 2020, doi: 10.1108/JEIM-06-2019-0156.
- [22] O. Zengin and M. E. Onder, "YouTube for information about side effects of biologic therapy: a social media analysis," *International Journal of Rheumatic Diseases*, vol. 23, no. 12, pp. 1645–1650, 2020, doi: 10.1111/1756-185X.14003.
- [23] M. J. Stern et al., "Evaluating the data quality of a national sample of young sexual and gender minorities recruited using social media: the influence of different design formats," *Social Science Computer Review*, vol. 40, no. 3, pp. 663–677, 2022, doi: 10.1177/0894439320928240.




- [24] K.-S. Kim, S.-C. J. Sin, and E. Y. -Lee, "Use and evaluation of information from social media: a longitudinal cohort study," *Library & Information Science Research*, vol. 43, no. 3, p. 101104, 2021, doi: 10.1016/j.lisr.2021.101104.
- [25] R. Tilly, O. Posegga, K. Fischbach, and D. Schoder, "Towards a conceptualization of data and information quality in social information systems," *Business & Information Systems Engineering*, vol. 59, no. 1, pp. 3–21, 2017, doi: 10.1007/s12599-016-0459-8.
- [26] K. Silius, M. Kailanto, and A.-M. Tervakari, "Evaluating the quality of social media in an educational context," in *2011 IEEE Global Engineering Education Conference (EDUCON)*, 2011, pp. 505–510, doi: 10.1109/EDUCON.2011.5773183.
- [27] M.-A. Abbasi and H. Liu, "Measuring user credibility in social media," in *Social Computing, Behavioral-Cultural Modeling and Prediction*, Berlin, Heidelberg: Springer, 2013, pp. 441–448, doi: 10.1007/978-3-642-37210-0_48.
- [28] J. O'Donovan, B. Kang, G. Meyer, T. Hollerer, and S. Adalii, "Credibility in context: an analysis of feature distributions in Twitter," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, 2012, pp. 293–301, doi: 10.1109/SocialCom-PASSAT.2012.128.
- [29] R. Berlanga, I. L. -Cruz, and M. J. Aramburu, "Quality indicators for social business intelligence," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2019, pp. 229–236, doi: 10.1109/SNAMS.2019.8931862.
- [30] A. -A. -Tanvir, E. M. Mahir, S. Akhter, and M. R. Huq, "Detecting fake news using machine learning and deep learning algorithms," in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, 2019, pp. 1–5, doi: 10.1109/ICSCC.2019.8843612.
- [31] C. Songchon, G. Wright, and L. Beevers, "Quality assessment of crowdsourced social media data for urban flood management," *Computers, Environment and Urban Systems*, vol. 90, p. 101690, 2021, doi: 10.1016/j.compenvurbsys.2021.101690.
- [32] M. Alsyouf, P. Stokes, D. Hur, A. Amasyali, H. Ruckle, and B. Hu, "Fake news' in urology: evaluating the accuracy of articles shared on social media in genitourinary malignancies," *BJU International*, vol. 124, no. 4, pp. 701–706, 2019, doi: 10.1111/bju.14787.

BIOGRAPHIES OF AUTHORS



Oumaima Reda    is pursuing Ph.D. in Computer Science from National School of Computer Science and System Analysis (ENSIAS), Mohammed V University in Rabat, Morocco. She obtained a master degree in internet of things and services mobile (IOSM) from National School of Computer Science and System Analysis, in 2019. Her research interest includes data science, data quality, and social media. She can be contacted at email: oumaima_reda@um5.ac.ma.



Ahmed Zellou    received his Ph.D. in Computer Science at the Mohammedia School of Engineers, Mohammed V University, Rabat, Morocco 2008, his habilitation to supervise research work in 2014. He becomes full professor in 2021. His research interests include interoperability, mediation systems, distributed computing, data quality and semantic web where he is the author/co-author of over 100 research publications. He can be contacted at email: ahmed.zellou@um5.ac.ma.