Bus Inf Syst Eng
https://doi.org/10.1007/s12599-023-00840-9

RESEARCH PAPER

A Data Quality Multidimensional Model for Social Media Analysis

María José Aramburu · Rafael Berlanga · Indira Lanza-Cruz

Received: 28 November 2022 / Accepted: 12 September 2023

Abstract Social media platforms have become a new source of useful information for companies. Ensuring the business value of social media first requires an analysis of the quality of the relevant data and then the development of practical business intelligence solutions. This paper aims at building high-quality datasets for social business intelligence (SoBI). The proposed method offers an integrated and dynamic approach to identify the relevant quality metrics for each analysis domain. This method employs a novel multidimensional data model for the construction of cubes with impact measures for various quality metrics. In this model, quality metrics and indicators are organized in two main axes. The first one concerns the kind of facts to be extracted, namely: posts, users, and topics. The second axis refers to the quality perspectives to be assessed, namely: credibility, reputation, usefulness, and completeness. Additionally, quality cubes include a user-role dimension so that quality metrics can be evaluated in terms of the user business roles. To demonstrate the usefulness of this approach, the authors have applied their method to two separate domains: automotive business and natural disasters management. Results show that the trade-off between quantity and quality for social media data is focused on a small percentage of relevant users. Thus, data filtering can be easily performed by simply ranking the posts according to the quality metrics identified with the proposed method. As far as the authors know, this is the first approach that integrates both the extraction of analytical facts and the assessment of social media data quality in the same framework.

Keywords Data quality · Social media data · Business intelligence · Text analytics

1 Introduction

Social media has emerged as a valuable source of information for companies, enabling them to understand customer opinions, analyze market trends, and uncover new business opportunities, among other benefits (Ruhi 2014). Although social media contents are highly heterogeneous and difficult to manage, they can produce meaningful business information for decision-makers. The research presented here focuses on data quality management of social media data collections for Business Intelligence applications.

Business intelligence (BI) is the process of collecting, storing and analyzing data from business operations to assist organizations in becoming data-driven (Sabherwal and Becerra-Fernandez 2013). Although BI tools are primarily powered by operational data sources (i.e., OLTP data), they also allow business users to access heterogeneous types of data from historical/current, structured/unstructured, internal/external sources. BI user practices range from analytics and reporting to data mining and predictive analytics. BI platforms rely on data warehouses

Accepted after one revision by Óscar Pastor.

M. J. Aramburu (✉)
Department de'Enginyeria i Ciència dels Computadors, Universitat Jaume I, 12071 Castelló de la Plana, Spain
e-mail: aramburu@uji.es

R. Berlanga · I. Lanza-Cruz
Department de Llenguatges i Sistemes Informàtics, Universitat Jaume I, 12071 Castelló de la Plana, Spain
e-mail: berlanga@uji.es

I. Lanza-Cruz
e-mail: lanza@uji.es

Published online: 10 November 2023

for storing their reference information. More specifically, a traditional data warehouse aggregates operational data into multidimensional data structures applied by online analytical processing (OLAP) engines to execute data-intensive queries. Although BI tools are primarily descriptive in summarizing historical and present data evolution, data analytics modules can also become part of modern BI systems, enhancing them with statistical tools as well as artificial intelligence and machine learning capabilities. These tools enable deeper business insights producing business information ranging from descriptive to predictive, prescriptive, and self-explanatory (Gröger 2021).

Social Business Intelligence (SoBI) is defined in Gallinucci et al. (2015) as the discipline that aims at combining business with social media data to form a corporate data warehouse which lets decision-makers enhance their business needs based on the trends and moods perceived from the environment. Until now, companies have used social networks mainly for marketing purposes (Păvăloaia et al. 2020). In fact, SoBI tools are often applied by marketing departments to monitor the performance of their social media activities with metrics like number of likes, followers, or replies (Keegan and Rowley 2017; Lee 2018) as well as the feelings and concerns of their customers (Choi et al. 2020). However, social media data have many more uses and applications for businesses and industries (Gioti 2018), and integrating social media metrics with corporate data can help to produce better strategic indicators to drive companies forward (García-Moya et al. 2013; Ruhi 2014; Stieglitz et al. 2014).

Social media analytics (SMA) is another large family of related applications that can be defined as the ability to gather and find some meaning in social media data to aid business decisions and measure the performance of social media actions based on those decisions (Ruhi 2014). Dealing with social media data, these data analytics tools also form part of modern SoBI platforms as they can be applied to help business decisions (Holsapple et al. 2018). In a general business setting, SMA is focused on statistical and machine learning tools that apply correlation, regression, and classification, together with sentiment extraction, to transform social media data into meaningful information for business purposes (Stieglitz et al. 2014). SMA has many real-world applications and has been widely applied by the research community to solve different types of problems related to business management (Stieglitz et al. 2014; Zachlod et al. 2022).

In this paper, we use SoBI and SMA as a unified term and treat social media big data analytics as a related field. However, to emphasize our perspective of integrating social media data into a BI environment, we mainly use the term SoBI for the remainder of the paper. All these systems have in common that, to produce valuable insights, they require feeding with collections of social media data of good quality with respect to their analysis objectives. However, building good quality collections is difficult because social media posts consist of unstructured texts with a high level of semantic heterogeneity. Fake posts, jokes, bots, and misinformation are often mixed with serious user-generated contents. In addition, the range of users participating in social networks is also diverse and their posts serve very different purposes. In a business environment, it is possible to find anonymous customers who publish opinions about a brand (e.g., offers, products or services), employees of the company who generate ads for marketing purposes, and other professionals posting contents somehow related to the brand (e.g., journalists, professional or customer associations, influencers, etc.).

Current approaches build collections of social media data by translating a subject of analysis into a set of retrieval keywords (i.e., topics, usernames, and hashtags). These keywords are then applied to filter one or several social networks (e.g., Twitter, Facebook, etc.), generating in this way a stream of potentially relevant posts with different degrees of quality for the analysis objectives (Holsapple et al. 2018; Arolfo et al. 2022). Most times, in the large volume of data retrieved, there are many posts apparently related to the subject of analysis but turn out because of their origin, intention or specific contents, to be useless and to produce noise or misunderstandings (Aramburu et al. 2021). These posts do not add any value to the analysis tasks and may even be counterproductive, due to the misinformation and the noise they produce. For example, in our experiments, when attempting to gather customer opinions about Ford car models, it became challenging to prevent the retrieval of numerous irrelevant memes related to the actor Harrison Ford, as well as the words “fiesta”, “escort”, or “focus”, which also correspond to certain Ford car models. Therefore, before exploiting a collection of social media posts, it is necessary to perform some additional quality management operations to assess its overall quality and to filter the posts that are relevant for the specific analysis task (Tilly et al. 2017).

Previous frameworks for SoBI and SMA have not paid the required attention to data quality, and more research is needed (Alrubaian et al. 2019). As noted by Stieglitz et al. (2018), in the papers that already document the data tracking and preparation steps of their social media analysis projects, these steps are often dealt with superficially and never with as much extension as data analysis tasks. The authors conclude that the phases of data discovery, collection and preparation of social media data projects require more research. Most approaches just apply a series of ad-hoc rules to posts (e.g., tweets with more than three retweets, users with more than one hundred followers, and so on) to filter out those to be analyzed (Choi, J., et al.

2020; Arolfo et al. 2022). Authors do not explain how the concrete filtering rules have been identified nor how to measure their efficiency to increase the quality of the collection. In general, in the task of building collections of social media posts for analytical applications, two important quality management issues are missing: general purpose data quality models and systematic methods for identifying the best quality metrics for the posts.

1.1 Our Approach and Contributions

To clarify our approach, we must first define the key concepts of our quality model. A *quality attribute* is a qualitative property of the data that expresses some aspect to improve from the analyst's perspective. As we will see, credibility is the most frequent quality attribute for social media data, followed by trustworthiness, reliability, credibility, veracity, relevance, and validity, among others. The concept of *quality metric* refers to any method or function that serves to estimate the level of achievement of a quality attribute for a collection of data. These metrics are typically quantitative, generating numerical values that we denote as *quality measures*. These measures may be presented to analysts in different formats, or they contribute to the calculation of *quality indicators* that combine several metrics to implement more complex metrics.

Selecting the best quality criteria for social media data is a complex task that requires a deep understanding of both the context and objectives of analysis. In the literature, data provenance is the main quality dimension considered for social network analytics applications (i.e., the credibility of the author of the post), yet it has been superficially treated with ad-hoc combinations of aggregated metrics such as the number of likes, mentions or followers, and without considering any contextual circumstances. However, the credibility of a post depends largely on its intrinsic properties and the role the poster plays in the subject of analysis. More specifically, we believe that the credibility of most social media users can be well understood by measuring several aspects in their account definitions, in their metadata and profile descriptions.

In Berlanga et al. (2019), we presented a method to build indicators to assess the overall quality of collections of social media data by integrating the measures obtained by several quality criteria. By considering the peculiarities of each SoBI project (e.g., its context, objectives, topics, and participants), this method helps to find the quality criteria that best suit both the participants and the available posts data, and then integrate them to form a valid quality indicator. This approach relies on the selection of a ranking of relevant users associated with the different categories of posters and taking this ranking as reference, the method automatically calculates the impact of each quality metric. This method was included as a complementary component of the SoBI workflows in Aramburu (2021).

In this paper, we propose a new integrated approach for data quality assessment in SoBI projects, where quality is assessed at the same time that analytical facts are extracted from social media data. In this paper, we extend our previous work by providing a new formal framework that allows the definition of quality indicators adapted to the specific analysis tasks of a SoBI project. More specifically, our approach contributes to the current state of the art in the following aspects:

1. It provides a novel and formal method to measure data quality in social media data according to the main aspects identified in the literature, namely: credibility, reputation, usefulness and completeness. Our method relies on metadata-based metrics as well as content-based metrics derived from language models.
2. Our approach defines a novel multidimensional model, Q-cubes, to capture and profile the quality metrics. This model's main feature is the dimension of user roles, which allow a better understanding of the data quality with respect to their authors. The multidimensional model covers the three main kinds of factual data handled in SoBI, namely: posted contents, involved users, and mentioned topics.
3. The combination of quality indicators and Q-cubes allows analysts to define in a straightforward way the necessary data filters to obtain high quality collections for analysis. Moreover, Q-cubes provide an overall picture of the main features of the extracted facts in terms of their contents and the users that generate or interact with them.
4. The proposed approach integrates both fact extraction and data quality assessment from the beginning of the process (i.e., data sources) to the final output for analysis. The method provides analysts with global quality indicators, as well as partial quality indicators per topics and/or user roles, together with the necessary quality metrics thresholds for filtering the data.

1.2 Organisation of the Paper

The paper is organized as follows. Section 2 reviews the main related work with respect to quality assessment in SoBI projects. Section 3 describes the main quality aspects considered in this paper. Section 4 presents the proposed approach for quality assessment. Section 5 is devoted to the experiments carried out over two long-term data streams and their results. Finally, Sect. 6 provides conclusions, limitations and presents the future work.

2 Related Work

From the point of view of quality management, building collections of social media data for analysis applications involves several things. On the one hand, it is necessary to identify the best group of topics for retrieving the posts from social networks. Here, the purpose is to obtain a set of posts as homogeneous and complete as possible with respect to the objectives of analysis, without biases or missing information. On the other hand, it is also necessary to perform some cleaning operations to select those from the retrieved posts that are really related to the subject of analysis as well as to validate the overall quality of the final collection. To this end, it is necessary to determine the best quality metrics to be applied by means of filtering operations as well as to assess the quality of the overall collection. In both cases, the goals and circumstances of the analysis operation at hand determine a context that is of primary importance when defining these quality operations (Arolfo et al. 2022).

In this section, we review previous methodologies for the construction of social media data collections from the point of view of quality management. Next, we summarize the different approaches to measuring credibility, which is the most frequent quality attribute for social media data. Finally, we review and categorize the remaining quality attributes for social media data.

2.1 Previous Approaches

As explained in the introduction, most papers on social media data analysis work with ad hoc constructed collections. Although many of them recognize the importance of data cleaning operations, they do not explain how to ensure data quality during the preparation of a data collection for real-world scenarios. The framework for enterprise social media analytics proposed by Holsapple et al. (2018) is complete in all aspects of SoBI and SMA systems, but its processing method does not consider any data cleansing and quality assessment tasks. Next, we review a set of papers that have been selected because they illustrate the different ways in which previous approaches to the construction of social media data collections for SoBI and SMA applications have incorporated data quality management operations into their methodologies. Briefly, these approaches range from such that do not provide sufficient support to data quality, via those based on manual or black box methods, through to those focused on social media data streams, and finally up to recent approaches that allow for the definition of some parameters in order to adapt their behavior to the application context.

The uniform data management approach of Goonetilleke et al. (2014) reviews three main groups of research challenges to address when building a Twitter data analytics platform. For data collection, the main issue is the specification of the best set of retrieval keywords and hashtags. For data pre-processing, they demand specific text processing and information extraction strategies for Twitter data. Finally, for data management, they explain that quality management is a major issue, and quality metrics such as trust in authority or authenticity should be included in user languages to query social networks. Consequently, although this paper identifies all these limitations of the available technology, it does not provide a methodology to address them in the proposed framework.

The methodology for SoBI of Abu-Salih et al. (2015) proposes to execute cleaning operations to remove dirty data and ensure data consistency at the data acquisition stage prior to data storage. Later on, during data analysis, the collected data is processed to infer a domain-based value of trust for the relevant data based on the credibility of the data producers. Trustworthiness is estimated by means of a set of key credibility metrics (i.e., number of likes, retweets, replies, …) whose measures feed various machine learning modules proposed to predict high-influential users in a domain (Abu-Salih et al. 2020). In this way, the exploited social media data acquires a minimum level of trust with respect to its domain. This methodology does not include any tools to assist with the selection of the best quality criteria for cleaning operations nor any credibility measures for trustworthiness estimation.

A second methodology for SoBI (Francia et al. 2016) recognizes that crawling design can be one of the most complex and time-consuming activities and aims at retrieving in-topic clips by filtering off-topic clips. They also explain that filtering off-topic clips at crawling time could be difficult due to the limitations of the crawling languages and propose to filter them at a later stage by using the search features of a document's database. The authors note that manually labelling a sample of the retrieved clips enables the team to trigger a new iteration where the crawling queries are redefined to remove off-topic clips more effectively. However, this work does not consider the quality of data as a main objective, and it does not deal with the question how to obtain a good set of quality measures.

In the quality management architecture for social media data presented in (Pääkkönen and Jokitulppo 2017), the data acquisition, data processing & analysis, and decision-making phases can include functionalities for quality control and monitoring. In this approach, data quality management consists of assigning values to a predefined set of quality attributes that depend on the purpose of the data set at hand. In the following, data quality can be evaluated from the point of view of the data source (i.e., data provenance), the data (i.e., data quality) and the user (i.e.,

trustworthiness). The quality, organizational and decision-making policies of the organization define the criteria to filter the quality data. Although the proposed architecture can represent all these data quality elements, the authors do not propose a methodology for defining and applying them.

Over the last few years, several software architectures have emerged to process social media posts in near real-time for analytical purposes. For example, the work in Hammou et al. (2020) is a distributed intelligent system for real-time social big data analytics. This system takes advantage of distributed machine learning and deep learning techniques for enhancing decision-making processes. After data ingestion and storage, and before the text embedding translations, some cleaning operations can be executed. However, these operations only serve to perform a series of pre-processing actions such as removing the numbers, URLs, and hashtags.

Alternatively, Podhorany (2021) proposes an advanced architecture and workflow based on Apache Hadoop and Apache Spark Big Data platforms for collecting, storing, processing, and analyzing intensive data from social media streams. It uses text analysis methods and location estimation techniques to analyze the reported situation by using the information included in the processed posts. Although during the experiments, a cleaning phase executes various filters and text adjustment techniques, data cleaning operations were not included in the architecture proposed in the paper.

Finally, in a recent work, Arolfo et al. (2022) demonstrate the reliability of Twitter data for decision-making processes by means of a software tool that processes streams of tweets for presenting several graphics with quality measures. Its quality model considers four dimensions (i.e., the reliability, completeness, usefulness, and trustworthiness quality attributes) and measures them via a set of basic metrics whose measures are available in the tweets. The user can dynamically adjust the weights of the four dimensions to fit different contexts or interests. This approach constitutes a first attempt to define a context-aware quality model for social media data, but it is still quite limited because it relies on a fixed and non-validated set of metrics. For example, it measures usefulness in terms of sentiment expressions, which is only valid for a concrete type of applications (i.e., sentiment analysis). Similarly, measuring trustworthiness according to being a verified user or having many followers is also a way of restricting the interpretation of this quality dimension. In general, a context-aware data quality model should, first, allow users to define their own application-specific set of metrics to measure quality attributes and, second, provide them with formal tools to validate and choose the best metrics for each concrete analysis task.

2.2 Credibility and Reputation Metrics for Social Media Data

Credibility is the most frequent quality attribute for social media, and many different approaches have been proposed to measure it (Viviani and Pasi 2017; Alrubaian et al. 2019). The literature review clearly reveals that many SMA projects aim at analyzing concrete events such as a catastrophe or a terrorist attack where the main issue is to evaluate posts' credibility (Gupta et al. 2014; Kaufhold and Christian 2020; Saroj and Pal 2022). Customer review analytics is another large application field of social media data and also here credibility is the main issue (Hu et al. 2020; Zheng 2021). It is important to clarify that, for all these works, credibility is a broad concept that intersects with other semantically related quality attributes such as trust, reliability, believability, veracity, relevance, validity and, in some cases, even understandability and reputation.

Among the numerous metrics that feed into these algorithms, some are derived from processing post content, primarily focusing on textual attributes, writing styles, linguistic expressions, sentiments, and additional elements such as URLs or images. A second set of metrics is based on social parameters extracted from post metadata, including information about each post and its author. Lastly, there is a category of metrics that provides insights into the behavior and actions of users within the social network. Table 1 shows a sample of state-of the-art metrics used to measure credibility (Sikdar et al. 2013; Gupta et al. 2014; Viviani and Pasi 2017; Alrubaian et al. 2019) which, in many cases, could also be applied to assess other quality attributes. The broad spectrum of metrics demonstrates that credibility can be interpreted in diverse ways. It is the responsibility of the user to select the most suitable metrics for each project, considering its domain, available data, and the applied technologies (Aramburu et al. 2021).

The review of Alrubaian et al. (2019) found that most related work on Twitter content credibility assessment was performed at four levels of feature extraction: post, user, topic/event (computed as a numerical score for each tweet regarding that topic/event), and hybrid levels. Most approaches use automated and semi-automated techniques, including supervised and unsupervised machine learning algorithms, weighted algorithms, and graph-based methods. Data-driven models classify social media data as credible and not credible, which makes their results difficult to understand for users as they do not receive feedback on the quality features of credible posts. Alternative approaches based on various criteria are emerging, which focus on aggregation schemes to assess an overall credibility estimate (Pasi et al. 2019). Finally, graph-based approaches exploiting the social structure of connected

Table 1 Sample of metrics to measure credibility in social media data found in the literature (Aramburu et al. 2021)

Posts Contents	Posts and Posters Metadata	Users Behavior
# Chars/words	Account age	# Retweets
# Punctuation symbols	Listed count	# Tweets
# Pronouns	Status count	# Tweets favorited
# Swear words	Favorites count	# Mentions
# Uppercases	# Friends	# Tweets are a reply/retweet
# Emoticons	# Followers	Mean time between tweets
#URLs/images	# Followings	# Likes received
# Hashtags	Ratio of followers to friends	# Directed tweets
# Misspelled words	Mean text length in tweets	# Users that propagate the user
# Sentences	Mean hashtags in tweets	# Users the user propagates from
Average length of sentences	Mean # URLs/ mentions in tweets	# Tweets propagated by other users
# Product mentions	Verified account	# Users that converse with the user
# Product features mentioned	User image in user profile	Mean number of conversations
# Opinion sentences	Tweet geographical coordinates	Average length of chain-like behavior

entities analyze credibility propagation in social networks (Viviani and Pasi 2017).

The work of Pasi et al. (2019) proposes a multi-criteria decision-making approach aimed at assessing the credibility of user-generated contents. It considers features connected to the contents, the information sources and the relationships established in social media platforms. Then, the users are asked to manually evaluate all these features in terms of their impact on veracity. By considering different aggregation schema for the partial performance scores and their impact, the authors calculate an overall score of veracity. With respect to data-driven approaches based on machine learning techniques (Crawford et al. 2015), their approach enhances user awareness of the data features influencing the proposed decision, thereby reducing the problem's reliance on specific data. Furthermore, they also argue that making a binary decision on the credibility of a tweet is difficult in most contexts, and it would be better to provide users with both a binary classification and a ranking of credibility.

Finally, Abu-Salih et al. (2019) consider that adding a user-domain dimension to credibility assessment enhances understanding users' interest, but the literature shows a lack of approaches for measuring user-based trust. In particular, the accurate classification of the users' interest assists in providing a better understanding of posts contents. Previous work frequently considers simple measures such as the number of followers to calculate indicators of users' credibility, i.e., when users are in many Twitter lists and have many followers it is because the contents they generate satisfy many users. These approaches ignore that in a domain, users' interests can be diverse and evolve and change over time. To account for this, Abu-Salih et al. (2020) propose to consider this quality attribute as a time and domain-dependent parameter.

Regarding reputation, most work has been focused on identifying the influential users in a specific domain (Amigó et al. 2014). Existing approaches mainly rely on metrics similar to those presented in Table 1, plus combinations of the "followers" and "friends" metrics (Cresci et al. 2015) and vocabulary-based signals (Rodríguez-Vidal et al. 2019). More recent works showed that influential users can be effectively identified by their language models (Nebot et al. 2018; Rodríguez-Vidal et al. 2019).

2.3 Quality Attributes of Social Media Posts for SoBI Applications

While we have previously discussed how credibility can be assessed by combining user and post metrics, we note that the remaining quality attributes can be defined based on post content, user characteristics, and topic dimensions. Below, we review the most significant attributes according to this classification.

The work of Salvatore et al. (2021) also defines a set of quality dimensions and indicators for Twitter data, building upon the framework proposed by Cai and Zhu (2015) for Big Data. In this work, quality was represented into five dimensions: availability, usability, reliability, relevance, and presentation. Authors noted that quality categories are not independent of each other, as changes in a quality dimension impact other dimensions as well, for example, improving data completeness may lead to a loss of data accuracy. The resulting framework was oriented towards the identification of the main sources of error by means of a set of indicators and a collection of good practices that

should be undertaken when using social media data. Although this work is far from being a method to help to find the best quality criteria for a particular social media data collection and analysis task, the complete list of quality attributes that it contemplates are also considered in the following review.

2.3.1 Post Contents Quality Attributes

Two important attributes for measuring post quality are the legibility and clarity of post contents. Previous research has assessed these attributes by using metrics like the readability features proposed by Duan et al. (2012), as well as sets of facets related to linguistic quality from Berardi et al. (2011) and Gupta et al. (2014). The main purpose of these approaches was to discard the posts that were difficult to understand or not very credible because of their linguistic deficiencies. However, to filter quality posts and represent them in a format that is easy for analysis applications to process, the preparation phase should go beyond the linguistic properties of the texts and should try to extract their meaning, which would help users to recognize the semantic elements that are useful for analysis tasks (Kolajo et al. 2020).

Accuracy is an important quality attribute that ensures that the data is free of error. In the case of social media data, this attribute is difficult to measure due to the lack of a comparison baseline (Shankaranarayanan and Blake 2017). In the case of social media posts, the accuracy can be analyzed at two levels: posts contents and user metadata and are in both cases very difficult to measure.

Another important dimension of data quality is timeliness. Among the quality attributes of a domain, it may be useful to define the period during which the posts will add value. These time properties will depend on the objectives and circumstances of the analysis tasks. For example, a review of a car model could be valid for much longer than its promotion at a fare, since it could last for many months, until the manufacturer launches a new edition of the model or it disappears from the market.

The value of posts lies in the usefulness of the data contained in them, in the sense that it should be possible to extract from their contents the values that analysis tasks require (Berkani et al. 2019). Here lies an important source of risk which is the availability of metadata. For example, some analysis tasks, such as segmenting the market opinions with gender, age, location, or profession attributes, require metadata. Social media users do not always provide their real profiles so the available metadata may lack key attributes for the analysis. In some cases, it can be helpful to infer some of this data by semantically processing the content of all the posts, although this is difficult to keep updated for every user (Hernandez et al. 2013).

2.3.2 Users Quality Attributes

Social media users are followers of other users, so considering them as a source of business information makes the author's reputation a quality attribute of utmost importance. Valuable posts come from users with good reputation, because this fact conveys the credibility and accuracy of the contents that they post. The literature frequently considers the number of followers, likes and retweets as indicators of good reputation, e.g., the users that appear in many Twitter lists and have many followers generate posts that satisfy many users. However, reputation is a quality attribute that depends greatly on the business domain. Therefore, measuring the quality of posts cannot be as easy as checking the number of followers of their posters, it also requires considering further domain dependent conditions.

In social media platforms, a user account is verified if it proves to be a public interest account. Users with professional purposes will obtain better results by using verified accounts. Verification standards are clear and, among other strict conditions, the user account definition must contain serious information including a profile description, header photos, name, biography, and location. In general, account profile descriptions delimit the role of the relevant users in a business domain or application context. Therefore, the quality of verified users' accounts should always be considered together with their profile descriptions.

2.3.3 Quality Attributes for Topics

In social media platforms, there are mechanisms to retrieve posts by means of keywords, usernames, and hashtags (Goonetilleke et al. 2014). However, data completeness is not ensured due to the following causes:

- Using keywords there is no certainty of retrieving all the posts that deal with the subject, therefore bias and data loss may occur (Plachouras et al. 2013).
- It is difficult to find the set of hashtags that must be part of an analysis subject (Bansal et al. 2015).
- It is almost impossible to identify all the representative users of a topic of analysis, and a percentage of representative voices will be lost due to the simple fact that they have not participated in social media (Czernek 2018).

Rather than completeness, topic coverage is a quality attribute for social media that indicates whether the query used to retrieve a collection's posts is complete in the sense that it includes all relevant topics related to the objectives of the analysis task, considering keywords, hashtags and usernames as the representative elements of a topic of analysis.

2.4 Main Conclusion and Methodology of Work

In this section, we have reviewed relevant methodologies to build collections of social media data from the point of view of quality management. The main conclusion is that while most approaches to social media analysis for decision support apply different quality criteria during data preparation, it is not clear at this point how to define a general-purpose quality management method. Previous work has proposed many different quality metrics for multiple purposes, which depend mainly on the respective application, but whose effectiveness and validity is unproven. The experience demonstrates that, whatever method applied to assess the quality of data, a good combination of different types of metrics is part of the solution. However, there is no systematic methodology for identifying a valid set of quality metrics to build a reliable collection of social data for analysis tasks in a domain of application.

Our work methodology can be classified as design science research (DSR) (Johannesson and Perjons 2014). Initially, we identify the overarching issue of data quality in social network data, a topic extensively studied in the literature. Furthermore, we observe the absence of a well-founded methodology for assessing data quality in social network analysis. Consequently, in the subsequent sections, we propose the utilization of two grounded theories to address the data quality problem: multidimensional data modeling, and information retrieval (IR). The former explains the collection and summation of metrics to form quality indicators, considering various perspectives of the data quality issue. The latter deals with the relevance ranking of quality metrics. Our primary hypothesis starts from the assumption that data quality is significantly influenced by both the relevance of its posters for and the coherence of their posts in relation to the application domain. As a result, the solution development is consistently supported by the chosen theories and premises. Finally, the proposed method provides essential information to measure the quality of analytical data and make decisions about data filtering and/or parameter updating. Consequently, the evaluation of the resulting dataset by analysts may imply redefining some of the parameters of the entire extraction process, such as the keywords used to retrieve the data and the set of reference users. Subsequent iterations can then be carried out to further enhance the dataset. These iterations should always be guided by the automatically derived quality indicators, which demonstrate whether the actions taken have improved the results.

3 Data Quality Management Dimensions

Nowadays, data quality management is considered to be one of the main factors that guarantee a successful adoption of AI technologies by modern business and organizations (Jöhnk et al. 2021). As explained in Sadiq and Indulska (2017) and Zhang et al. (2019), traditional methods for managing data quality follow a top-down user-centric approach: the analysist specifies some quality rules that serve to govern data, to assess data quality, and to execute cleaning operations. This approach is suitable for managing the quality of data generated internally by an organization. However, when the organization does not control the external processes that generate the available data, as in the case of social media, quality assessment requires prior knowledge about the data features. To gain this knowledge, data quality management follows a bottom-up approach that starts with submitting the source data to some exploratory tasks (Zhang et al. 2019). These tasks help to find data quality rules and requirements that will drive the data collection process. To execute the preliminary exploration of the available data, interactive, statistical and data mining techniques are applied (Stieglitz et al. 2018).

Social media posts present many distinct aspects that could serve to filter them, with posts contents and users' attributes and interactions being the main contributors to quality metrics (see Table 1). However, the selection of the best quality metrics for a specific SoBI project requires a deep understanding of its business context, strategy, and objectives of analysis, as well as of the relevant social media data (i.e., posts and users) to be managed (Immonen et al. 2015; Berlanga et al. 2019). Thus, we consider three different dimensions for data quality: the social media users, the posts they generate, and the topics they write about. As Table 1 shows, these dimensions have been widely adopted in most of the approaches of social media analysis. They provide different quality metrics whose convenience, in the case of users and posts, will depend on the types of users that participate in the business domain.

In our work, we propose performing *global quality analyses* over long-term data streams. This is because quality problems, such as redundancy, bias and noise are often difficult to detect by means of local analysis (i.e., directly over the streamed data). The other strategy we propose for data quality management is *profiling the long-term data stream* according to a series of quality dimensions. Basically, as we will explain in following sections, profiling is performed by analyzing the language models of the users' profiles and their posts according to the intended quality analysis dimensions.

3.1 User Dimensions for Data Quality

Social media data profiling allows the analyst to have a better understanding of the real market of the business domain. For example, finding the most frequent topics in a collection of car reviews can help us to identify the range of aspects that should be part of the product features analysis dimension, as they are the hot topics in the market. Furthermore, profiling the range of users that post on the domain along with their metadata is also important for determining the measures and dimension attributes available to take part of the analysis multidimensional data structures (i.e., cubes). For example, demographic data in user descriptions can help defining the attributes of the customers' analysis dimension. Similarly, classifying the range of users who post about the car models of a brand into the different stakeholder groups can help the analysis' purposes in many different ways.

One main novelty of the proposed model is that we profile social media users according to business-related classes. For example, most verified user accounts have a strong relation to professional purposes, and their definitions contain useful information including a profile description, header photos, name, bio, and location. In this paper, we profile the users of a generic business domain according to the following main categories:

- **"Domain Business Users",** which have on-domain professional/business profiles and apply social media accounts to promote their products and services by posting high-quality contents regularly. They are often verified users.
- **"Domain Influential Users",** which can be identified by their profile descriptions and their large number of followers and retweets. In case they are unverified users, other users give them authority, and as experts they often publish quality posts for that domain. Influential users are followers of business users.
- **"Domain Interested Users",** which are relevant because of the high level of similarity between the domain and their profile description. Usually, interested users are followers of business and influential users.

Figure 1 shows some examples of this classification applied to the automotive domain. This classification allows analysts to distinguish users with clear roles from those whose relationship is more sporadic or irrelevant. In general, the credibility of the users with a clear role in a business domain and the quality of their posts is higher than that of the rest of out-of-business users.

3.2 Social Media Data Quality Perspectives

In this work, we define four perspectives for social media data quality, namely: credibility, reputation, usefulness, and completeness. These perspectives are derived from the discussion presented in Sect. 2. They serve to classify the chosen quality metrics and facilitate their combination into specific quality indicators to estimate the degree of achievement of each quality perspective.

Credibility indicators must reflect how reliable the user accounts are. Reliability means that the users are real and relevant to the analytical goals, and that they post information that can be trusted when performing an analysis of these data. Measures related to credibility are primarily associated with the activity of the users, the coherence of the contents they generate, and other evidence that characterize good posters. Users whose intentions differ significantly from the expected ones should be assigned an extremely low value for credibility. For example, spammers and jokers should be categorized as of low credibility.

Reputation indicators should consider the factors that contribute to user influence, and, therefore, the impact of the content they generate. Usually, high quality is associated with reputed accounts. However, in some domains, highly influential users are not aligned with the analytical goals, being the contents, and generate useless posts for the analytical goals at hand. In this case, although it is always desirable to have a suitable number of reputed accounts, there must be a trade-off with respect to other quality perspectives, such as usefulness.

Usefulness indicators are of primary importance as they give us the clues of the potential impact of data on the analytical tasks. These indicators mainly measure the relevance and readability of the extracted data to derive useful facts for analysis. In this paper, we introduce the concept of coherence, which aims at measuring how well the languages of the data stream and the analytical goals are aligned. We will define the usefulness indicators over these kinds of measures.

Lastly, **completeness** should be viewed as a measure of how well the data covers a specific analytical topic. In this case, data has already been transformed into facts and we can directly measure how well facts cover the desired dimensions of analysis.

In the following section, we propose a new multidimensional model that integrates the elements defined in this section (i.e., user categories and quality perspectives) as a way to improve the analysis of social media data quality from the point of view of the different types of users participating in the application domain.

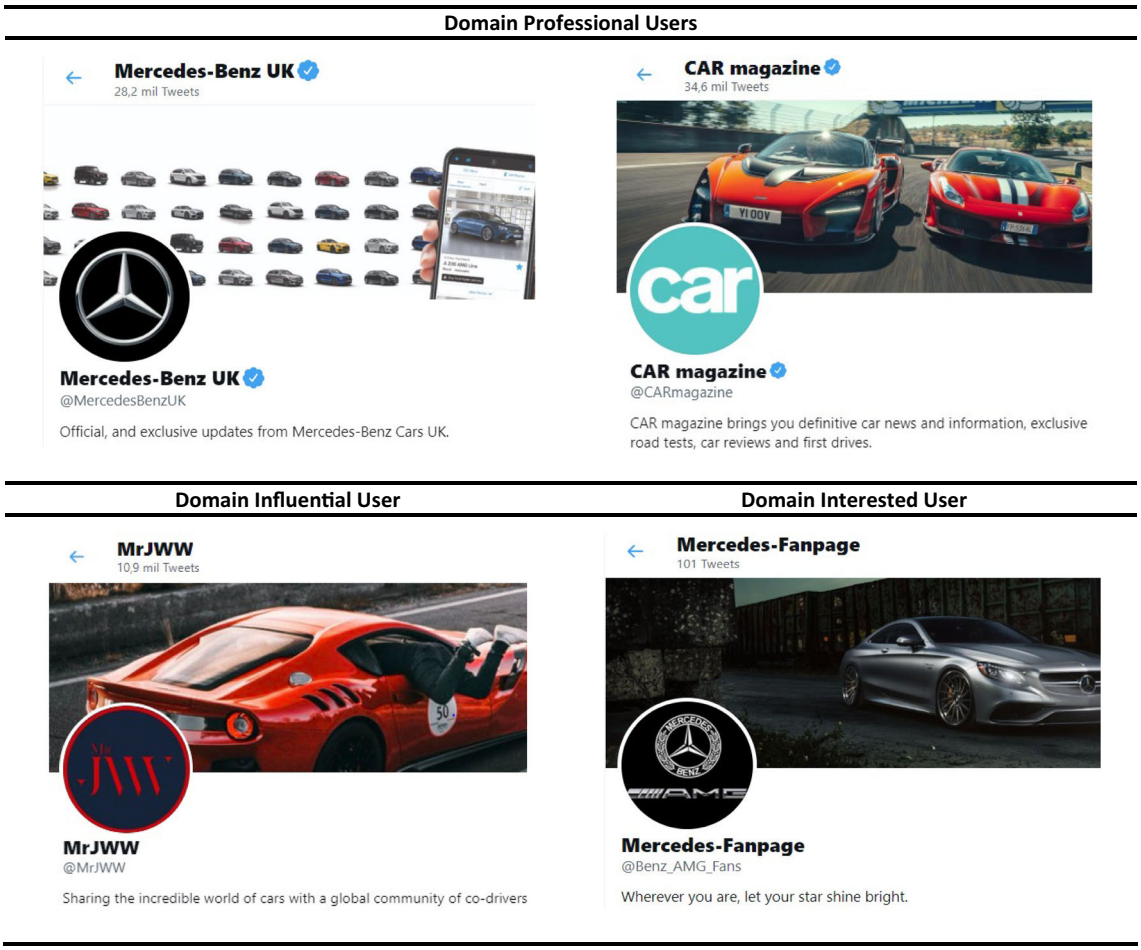

Fig. 1 Examples of Twitter accounts in the automotive domain for each category of relevant user

4 A Multidimensional Model for Quality Assessment

Following the multidimensional data model for Business Intelligence systems (Kimball and Ross 2013), we define several data structures (i.e., cubes) for representing the different quality perspectives as required by analysts. In these cubes, we store measures of the quality metrics' impact to filter out high quality posts for each type of user. Thus, we define two separate dimensions that represent the different post quality metrics and the user roles that can help to decide on the appropriate quality criteria for each analysis task. In this section, we develop the elements of this quality model whose main purpose it is to assess its metrics for a domain-related data stream.

4.1 Facts Extraction and Analysis

The process of social media analysis starts with the definition of a posts data stream using the social network API. The data stream is configured with a series of keywords that are directly related to the goals of analysis. Quality analysis can serve users as a guide to assess the effectiveness of the chosen keywords and the potential lack of data for the intended analysis goals. Figure 2 summarizes the process of social media fact extraction and the subsequent process for quality assessment, which is described in turn.

At this point, it is important to note that Fig. 2 consists of two parts. The lower part (shaded in grey) corresponds to the extraction of facts from the data sources and is not treated in this paper because it is part of our previous work on the SLOD-BI infrastructure (Berlanga et al. 2015). The upper part of the figure includes the Aggregation and Quality Assessment phases and constitutes the central contribution of this work, namely, a new data processing method for quality assessment for social media analysis. In the following paragraphs, we will briefly explain the main components of Fig. 2.

Analysts design their goals by choosing the topics of interest and associating them to a series of analysis dimensions and measures. For example, the topic "car recalls" will have associated dimensions like "location",

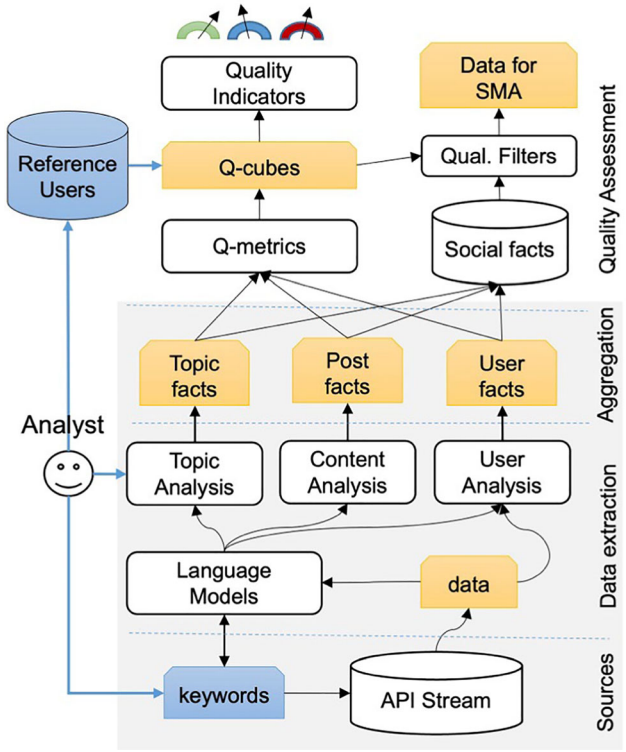

Fig. 2 Proposed data processing method for quality assessment for social media analysis

"detected failure" and "car model" and measures like "reported_cases" and "social media impact". Thus, we assume that the analyst has defined the set of dimensions D that are of interest for their analysis. For the sake of simplicity, we define an *analysis dimension* $d_i \in D$ as a set of values where each value is associated with a description and the lexical elements that enable recognizing the value in the post's contents and metadata.

Once data are retrieved from the social network API, we process them from three different perspectives: topic analysis, content analysis and user analysis.

Topic analysis concerns the discovering and organization of themes of interest from a large collection of posts. Topic analysis is one of the central tasks of any social media analysis as it serves to gain insight into the main concerns of social media users. Its main challenges are the high dynamicity and semantic drift of the user-generated contents, which make it necessary to continuously track the stream. Unsupervised machine learning has usually been adopted for topic analysis, mainly clustering and statistical methods like n-gram analysis and Latent Dirichlet Allocation (Chauhan and Shah 2021).

Content analysis is primarily focused on extracting implicit information from user-generated contents such as sentiments (i.e., polarity), emotions and entity mentions relevant to the analysis goals. Both supervised and unsupervised machine learning methods have been proposed in the literature for this purpose (Birjali et al. 2021). The output of the content analysis module usually are the facts that end-users are supposed to analyze. Content analysis is directly guided by the analysis dimensions and measures defined by the analyst. The facts extracted through content analysis will be inserted into the fact tables for performing the integrated BI tasks.

The User analysis component assigns a profile to each user account according to the analytical goals. Author profiling is a related task that aims at identifying user attributes from their generated content and biographies. Approaches in the literature have mainly focused on demographic attributes such as age, race, and gender. Some works have also treated more interesting attributes for BI such as professional profiles and influence degree (Amigó et al. 2014; Han et al. 2017; Nebot et al. 2018; Rodríguez-Vidal et al. 2019).

These components produce three types of facts as output: topic, post, and user facts. These facts represent all the explicit and implicit data that are useful for analysis. Therefore, our aim is to measure the quality of these facts and propose methods to improve their quality for analysis tasks.

In this paper, we assume that these components are dealing with a collection of posts C, from which a series of facts are extracted, denoted as *facts*(C), which can be further distinguished to be topic facts (*t-facts*), user facts (*u-facts*) and post facts (*p-facts*) when necessary. Finally, we can filter the extracted facts by applying the quality criteria derived from the quality cubes. In the following sections, we discuss how to measure the quality of social media data in terms of these facts and the set of *Reference Users* whose posts are recognized to possess of good quality.

4.2 Q-cubes: Multidimensional Analysis of Quality Metrics

Our approach defines a novel multidimensional model, consisting of three quality cubes (Q-cubes), to capture and profile quality metrics. Specifically, we propose two Q-cubes for analyzing the quality metrics of a domain-related data stream, namely: the Posts Quality Cube (PQC) and the Users Quality Cube (UQC). In addition, to assess the quality of the posts for each specific analysis topic, we also define a third cube called the Topic Quality Cube (TQC).

The PQC aims at measuring the impact of quality metrics derived from the contents and metadata of the posts. Table 2 summarizes the main aspects regarded for the PQC cube. Similarly, the UQC aims at measuring the impact of quality metrics associated to different aspects

Table 2 Metrics categories for the Posts Quality Cube

Group Id	Post Attributes	Quality Metrics
P1	Metadata	Metrics provided along with the posts (e.g., #retweets, #likes, etc.)
P2	Contents	Metrics derived from the contents of the posts, like text and images
P3	References	Metrics involving the quality of links, mentions and hashtags included in the posts

Table 3 Metrics categories for the Users Quality Cube

Group Id	User Attributes	Quality Metrics
U1	User's Posts	Aggregate metrics over the posts written by the user (e.g., #tweets on-domain, stylistic-related metrics)
U2	Description	Metrics derived from the description in the users' profile accounts
U3	Metadata	Aggregate metrics of the posters (e.g., #followers, #friends, etc.)
U4	Interactions	Metrics derived from the interactions of the posters and towards the posters (e.g., #performed actions, #received actions, etc.)

related to the users. Table 3 summarizes the main aspects of metrics included in the UQC.

The TQC cube will provide a summary of the quality aspects associated with each analysis topic, as well as the necessary information for selecting the suitable quality criteria for data filtering. Table 4 shows the two groups of metrics we consider for topics. In addition to these metrics, as topics are subsets of posts, all quality metrics in Table 2 can be also applied to topics by aggregating them accordingly.

The Q-cubes are built with the impact values derived from processing the long-term data stream. We use Q-facts tables to store the extracted facts that will serve to fill the Q-cubes, that is, each Q-fact table includes all the observations of the quality metrics for each post/user/topic of the long-term data stream. More details about how PQC and UQC facts are processed and aggregated in stream can be found in Lanza-Cruz et al. (2018).

The three Q-cubes share the User-Role dimension. This dimension regards the classes of reference users for the domain at hand (see Sect. 3), and it is a clear indicator of the credibility of the social media users. Therefore, the role of users becomes the main dimension for assessing the quality metrics. As Fig. 3 shows, the User-Role dimension includes the corresponding Domain Business Users, Domain Influential Users and Domain Interested Users hierarchies. Analysts can further refine these conceptual categories into more specific user profiles for a business domain. For example, in the case study of this paper, we define the following sub-categories: Employees (E), Professionals (P), Public Services (PS), Journalists (J) and Lovers & Fans (L&F). These categories are inspired in the RepLab 2014 dataset and designed according to the experts' criteria involved in the project (Amigó et al. 2014).

The User-Role dimension is populated with a list of Reference Users who are supposed to produce high-quality posts contents. These users of reference must be present in the long-term data stream to compare them with the rest of users. Before we start building the Q-cubes, we need to attach the User-Role dimension to the Q-facts tables. For this purpose, we label as relevant all the facts involving the reference users, and we add the user category labels associated to them. The last step is to build the Q-cubes by measuring the impact of all the included quality metrics. This step is performed as follows:

Table 4 Metrics categories for the Topics Quality Cube

Group Id	User Attributes	Quality Metrics
T1	Audience	Percentage of users deemed relevant to the topic, ratio of users per relevant business roles, etc.
T2	Contents	Number of tweets, contents diversity (e.g., Yule's metric), temporal distribution, coverage of dimensions, etc.

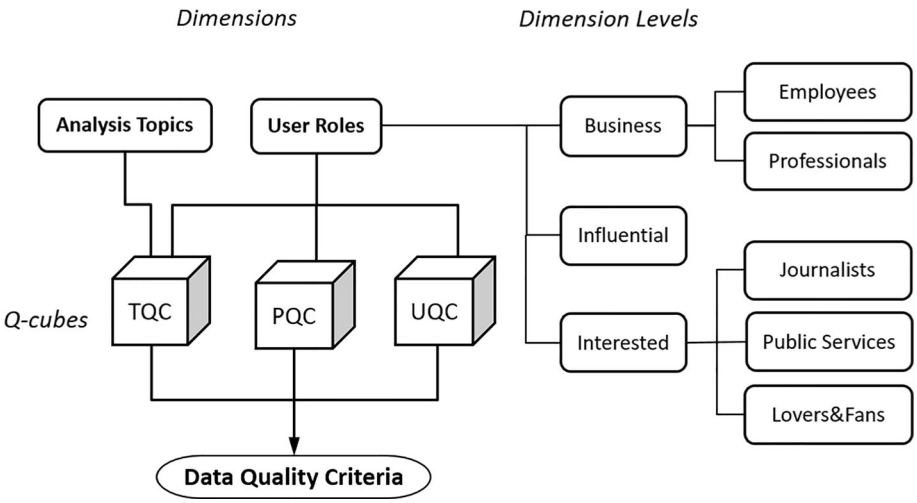

Fig. 3 Quality cubes for social media data

1. For each quality metric, we arrange the Q-fact table based on its values. Being a quality metric, the default ordering should be descendant, that is, the greater the value the higher the quality.
2. We then calculate the overall impact of the quality metric using an evaluation metric applied to the obtained ranking of facts (in the way explained in Sect. 4.3).
3. Finally, we measure the impact of the quality metric for all the categories associated to the User-Role dimension to populate the corresponding cube.

As a result, the final Q-cubes show the impact of each included quality metric broken down by user categories. From these Q-cubes, we can finally derive the quality criteria that allow us to refine the final dataset for analysis purposes.

4.3 Impact of Quality Metrics

To assess the impact of quality metrics, we apply the average precision (AP), which has been widely applied in information retrieval (Baeza-Yates and Ribeiro-Neto 1999). This metric is both easy to implement and efficient to compute. Moreover, recent work has shown how this metric can be approximated with a differentiable function, allowing it to be included in deep learning models (Cakir et al. 2019). Given a list of ordered items (users or posts), the metric AP is defined as follows:

$$AP = \frac{\sum_{k=1}^{N} P_k \cdot rel(k)}{R}$$

where R is the number of relevant items in the collection, N is the size (i.e., number of items) of the complete collection, P_k is the precision at position k, and $rel(k)$ is a binary number indicating whether the element at position k is relevant or not. Notice that if relevant items are uniformly distributed in the ranking, then the value of AP is $AP_{unif} = \frac{R}{N}$

In order to compare impact metrics from different rankings and perspectives, we define a normalized metric that takes into account the relative change with respect to AP_{unif}, namely:

$$AP_{rel} = \frac{(AP - AP_{unif})}{AP}$$

Notice that rankings with AP_{rel} near zero are not useful for quality assessment since they are not able to promote high-quality items. Negative scores indicate poorer quality metrics since their corresponding ranking demotes qualified data.

Figure 4 illustrates the relationship between AP_{rel} and the ranking power of two different metrics. The graphic above represents a good metric, in this case, the number of tweets in the domain, as it promotes reference users to the top positions when ranked by the metric. The graphic below represents a worse metric, in this case, the number of followers, as reference users have been positioned randomly when ranking by the metric.

In this way, we define the impact of a quality metric M, denoted $impact(M)$, as the AP_{rel} measure when ordering the data with M.

It should be noted that the usefulness of a metric is intrinsically defined by how effectively it distinguishes relevant users from irrelevant ones. The analyst might discover some justification for the metric's behavior a posteriori by analyzing its results; for example, the metric "#followers" does not perform well in the automotive domain because many relevant users have a discrete value for this metric. Being familiar with the domain of an application can help analysts identify the set of relevant users, understand the behavior of different quality metrics, and enhance the overall efficiency of the method.

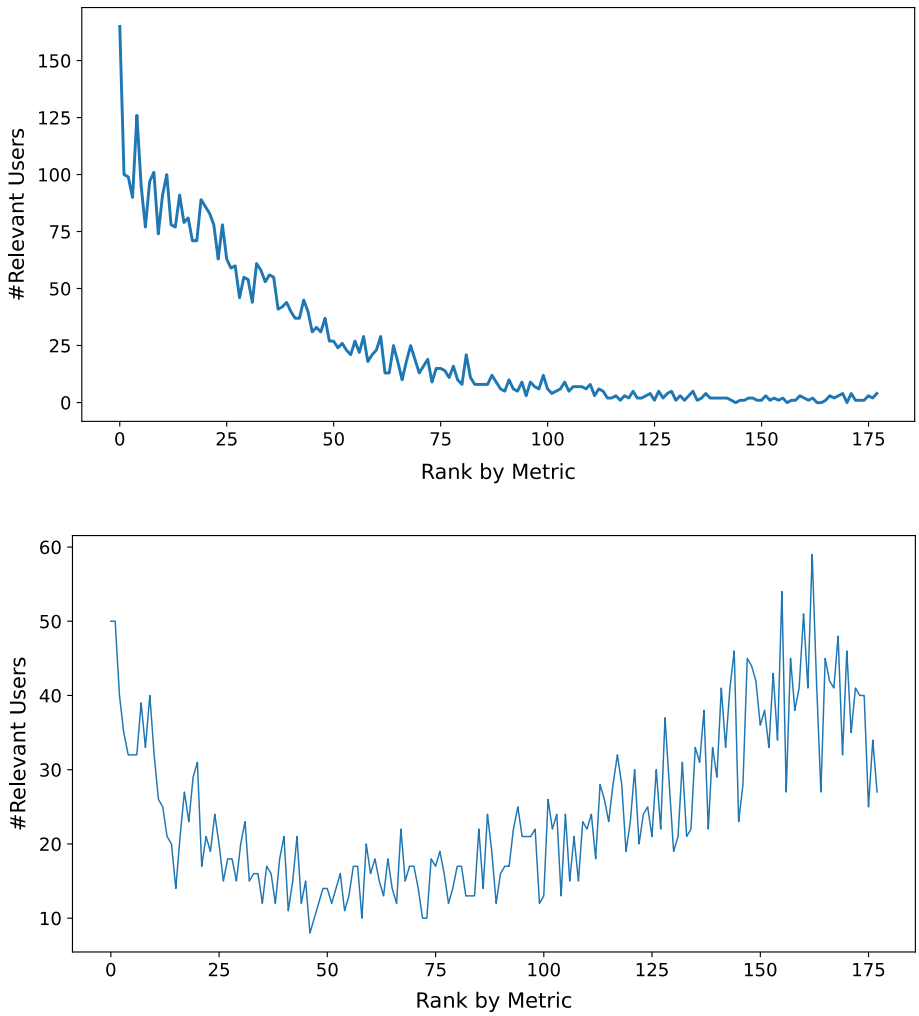

Fig. 4 Examples of the impact of two metrics in the ranking of reference users in the automotive domain (top: #on-domain tweets, bottom: #followers)

4.4 Definition of Quality Indicators

In this section, we explain how Q-Cubes can be applied to define quality indicators for the four quality perspectives adopted in Sect. 3.2. We define a quality indicator as having the following elements:

- A normalized value, usually between 0 and 1, where 1 indicates the maximum quality level and 0 the lowest one; sometimes for convenience we will use percentages.
- A series of quality metrics taken from the available data, which provide the support for the quality indicator.
- A formula to derive the indicator value from the selected quality metrics.

Thus, Q-cubes provide the quality metrics along with their impact for the different analysis facts (i.e., posts, users, and topics). Table 5 summarizes the proposed methods to measure the quality in the different perspectives.

Table 5 Quality perspectives and involved methods

Quality perspective	Methods for quality indicators
Credibility	UQC
Reputation	UQC
Usefulness	Language models + PQC
Completeness	Language models + PQC + TQC

We can define credibility and reputation indicators directly from the Q-cubes metrics, whereas usefulness and completeness depend on the topics and analytical goals at hand. Notice that usefulness measures the relevant facts that we can extract from the data, and completeness depends on the specific dimension values involved in a specific analysis (e.g., mentioned organizations, places, etc.). It is also worth mentioning that completeness depends on how posts are arranged to form the final analytical facts. The common approach is to treat posts

individually (Arolfo et al. 2022), which implies that only a few dimensions can be extracted from each post. To achieve a higher level of completeness, we need to group posts around entities and valid times to capture the different dimensions of the intended analytical facts.

A straightforward method to define quality indicators consists in estimating the percentage of users or posts that fulfil some property (Arolfo et al. 2022). For example, this could be the number of posts mentioning at least one brand name, or the number of posts sent by users with a certain number of followers. However, these properties are difficult to define, and they strongly depend on each specific domain, the analytical goals, and the community of users that generates the contents. In this work, by means of the Q-cubes, we aim at identifying the relevant quality metrics that allow us to distinguish relevant users and contents. As these measures promote reference users at the top positions, we can derive an indicator directly from the distribution induced by them.

We define a quality indicator from a quality metric as follows:

$$\begin{aligned}\text{score}(M) = \ & \text{impact}(M) \\ & \cdot \max_{x \in M}(2 \cdot \text{cov}(x) \cdot q(x)/(\text{cov}(x) \ + q(x)))\end{aligned}$$

This score combines the impact of the metric M according to the corresponding Q-cube, and the maximum of the harmonic mean between the ratio of covered posts and the quantile at a given cut point x in the metric M. In other words, we try to maximize the coverage of posts and the ratio of good posters. A high quantile value combined with a high impact implies high quality, because we are selecting a small number of very relevant users. This value is combined with the covered posts by these users so that we can find a good trade-off between them. The cut point of the metric could be directly used to filter out the data to increase the quality of the dataset. However, this should be only performed when the impact of the metric is high enough.

Credibility and reputation quality dimensions are directly associated with these scores. More specifically, we will take as reference the maximum score of the metrics that are related to these perspectives. For example, metrics like "#followers" are usually associated with reputation, whereas the meta-attribute "verified account" refer to credibility. Notice that these perspectives can take metrics from both posts and users Q-cubes, however, our experiments have demonstrated that user metrics (UQC) obtain much better results.

4.4.1 Language Models for Posts Facts Quality

To account for the usefulness and completeness of the dataset, we propose to use language models. A language model is a probability distribution assigned to each word or term of a vocabulary, which can be further conditioned by a series of contextual parameters (Baeza-Yates and Ribeiro-Neto 1999). As we aim at identifying entities and mentions of dimension values in the dataset, language models are a useful and well-grounded tool for estimating how well the domain is covered by the collection of posts and how well analytical goals are aligned to their contents. It is worth mentioning that previous work on language models for social networks have been shown effective in profiling users by their posts (Nebot et al. 2018; Rodríguez-Vidal et al. 2019).

We define two quality measures based on language models, namely: profile coherence and post coherence. These measures directly calculate the log likelihood of the profile and post contents with respect to a language model L representing the intended analysis goals for a particular domain. Coherence for a fact f is defined as follows:

$$coherence(f, L) = -\sum_{v \in f} \log P(v|L)$$

It is worth mentioning that the lower the metric the higher the coherence. Thus, we must filter out high values to ensure a high coherence. This metric is equivalent to the perplexity of the language model L, which has been also proposed for post quality in (Lin and Morgan 2011).

Usefulness of a post collection C can be defined as the ratio of facts extracted from C that are coherent enough to the language model of intended analytical goals L. This can be formally stated as follows:

$$usefulnes(C, L) \quad = \frac{\#\{f | f \in facts(C) \wedge coherence(f, L) < \delta_U\}}{|facts(C)|}$$

The parameter δ_U can be empirically set from the set of reference users specified for building the Q-cubes.

Completeness can be also expressed in terms of language models as follows. In this case, we need to measure the overlap between the vocabulary and the analysis problem, which consists of a set of dimensions $\{D_i\}$, and the facts extracted from the user-generated contents. The completeness of a fact f is simply defined as:

$$completeness(f, L) \quad = \frac{\sum_{i=1}^{N} \#\{d | \exists v \in f, d \in D_i \wedge P(d|v, L) > \delta_c\}}{\sum_{i=1}^{N} |D_i|}$$

In other words, we measure the ratio of dimension values that can be entailed by the fact values. The entailment relationships between terms are established by using either traditional statistical techniques or via modern word/

sentence neural embeddings (Lauriola et al. 2022). With this proposal, we also account for potential alignments that are not explicitly established between the dimension values and the extracted facts. As an example, the location attribute of a post can be inferred from the user or post metadata. In many cases, these locations are not specified in a systematic way, and it is very unlikely that they match our dimension values for locations. This way to express completeness, will account for hidden associations that otherwise will be lost because of issues like lexical mismatching. The threshold for entailments δ_c can be empirically set up by analyzing historical data.

Completeness can be easily defined for the whole collection just estimating the average over all the facts extracted from the intended collection.

$$completeness(C, L) = \frac{\sum_{f \in facts(C)} completeness(f, L)}{|facts(C)|}$$

Notice that all these measures can be applied to any arbitrary collection of facts extracted from user generated data. Moreover, these facts can be defined at user, post, or topic levels. Thus, for topic analysis we can restrict the collection to the facts relevant for some topic and estimate the previous indicators over the selected facts. In this case, the reference language model L must be also adjusted to the dimensions of interests for that topic.

5 Results

For demonstrating the usefulness of the proposed approach, we have chosen two long-term streams of tweets related to the automotive domain and to natural disasters. The automotive domain stream has been active since 2015 until now and it has served as basis of several studies about SoBI (Berlanga et al. 2015, 2019; Lanza-Cruz et al. 2018). We have generated this stream by specifying a series of keywords related to different car models and brands. It contains around 1,930,617 tweets, written in both Spanish (456,059) and English (1,474,558). The total number of involved users in this stream is 318,469 (up to November 2022).

For the second stream, the theme "impact of natural disasters and migration in the tourism sector" has been defined, which has been created by simply picking up the keywords of the theme, namely: "natural disaster", "migrants" and "tourism" (both in English and Spanish). It is important to notice that these keywords can introduce a lot of noise because Twitter searches for each word individually, not only in posts but also in screen names. In this work, we will deal with the data generated during the period 2019–2022, which contains around 26 million tweets involving around 21 million users. This period includes very popular topics like Brexit and the COVID-19 pandemic.

5.1 Users-Role Dimensions

For the first domain, we make use of two data sources for identifying relevant users. The first one is the RepLab 2014 dataset (Amigó et al. 2014), which contains a track for the automotive domain. We have selected only influencers from this dataset to ensure high-quality users. Thus, we get 480 influential users from RepLab (RL). For the second source, we have analyzed the bigrams language of the user descriptions and we have selected a representative set of bigrams for each user category. Table 6 summarizes the number of selected users along with some bigram examples used to build each category. The total number of reference users from this second source is 28,165.

For the "natural disasters and tourism" data stream, we have mainly focused on three main categories of reference users, namely: people and organizations working in aid/recovery, tourist destinations officers, and journalists. In this domain, the total number of reference users is 7386, from which 1290 are recovery-involved users, 3531 are destination officers, and 2565 are journalists. For this data stream, we have no influential users as reference. Since the automotive domain is much richer in terms of user categories, and for the sake of space, we will show only the quality cubes associated to the automotive domain in the next section. Global quality indicators are shown in Sect. 5.3 for both domains.

5.2 Quality Cubes

Table 7 presents the resulting PQC for the automotive domain. This table only includes quality metrics with a significant impact in some of the user categories. Shadowed cells represent near zero (< 0.1) and negative scores for AP_{rel}, (i.e., they are not relevant metrics for quality assessment). It is worth mentioning that AP_{rel} scores can only be compared within the same column, as the number of references varies by category, resulting in different scales.

In the PQC (Table 7), we have included the metric P1.1 as a fake quality metric so that we can check that effectively it has no impact in the quality assessment. We can see that many quality metrics for tweets have little impact in many user categories. In general, PQC metrics are less relevant than UQC metrics (Table 8). Tweet quality metrics mainly contributed to influencers, especially the metrics favorites (P1.4) and re-tweets (P1.5). Text coherence only contributed to the professional category, indicating that this is the main voice of the stream. The most relevant metric when regarding all users are the use of punctuation

Table 6 Examples of bigrams used for categorizing users in the automotive domain

Employees (E)	Professionals (P)	Public Services (PS)	Journalists (J)	Lovers & Fans (L&F)
2,389	8,705	739	8,611	7,721
·community manager	·used cars	·call emergency	·latest news	·love family
·project manager	·cars service	·crime call	·news information	·sports fanatic
·writer photographer	·car dealership	·report call	·motoring news	·love cars

Table 7 Results for the Post Quality Cube of the automotive-related Twitter stream (in bold face max value per role)

Quality Metrics	User Categories						
	All Categories	Influential Users	Business Users		Interested Users		
			E	P	PS	J	L&F
P1.1 Tweet date							
P1.2 Tweet is reply					0.198		0.311
P1.3 #Tweet replies		0.115	0.316		0.109		0.170
P1.4 #Tweet favorites	0.135	**0.737**	0.414		0.566	0.295	0.192
P1.5 #Re-tweets	0.190	0.716	0.386		**0.858**	**0.339**	0.128
P2.1 Coherence				0.288			
P2.2 #Numeric tokens	0.147			**0.379**			
P2.3 Tweet polarity	0.141		0.197	0.260			0.336
P2.4 Tweet repeats	0.172	0.247	0.124	0.145	0.463	0.215	0.153
P2.5 #Punctuation	**0.205**	0.131		0.157	0.302	0.123	0.318
P2.6 #Emoticons			0.104		0.204		
P2.7 #Mentions		0.122	**0.444**				**0.337**
P3.1 #Links		0.250					
P3.2 Question marks	0.167		0.164	0.233			0.176

Table 8 Results for the User Quality Cube of the automotive Twitter stream (in bold face max value per role)

Quality metrics	All categories	Influential users	User categories				
			Business users		Interested users		
			E	P	PS	J	L&F
U1.1 #Tweets on domain	0.729	0.962	0.834	0.708	0.772	**0.874**	0.624
U2.1 Coherence*	**0.760**	0.663	0.454	**0.859**	**0.780**	0.855	0.604
U2.2 Description length	0.406	0.439	0.595	0.563	0.721	0.350	0.587
U3.1 Account age		0.748	0.139			0.190	0.477
U3.2 #Statuses		0.617					0.118
U3.4 #Followers		0.924	0.367			0.535	0.285
U3.5 #Friends		0.766	0.408			0.367	0.284
U3.6 #Listed count		0.933	0.570		0.594	0.756	0.278
U3.6 Has location		0.329	0.144		0.167	0.158	
U3.7 Verified account		0.651	0.486			0.172	
U4.1 #Performed actions	0.481	0.933	**0.958**	0.254	0.344	0.660	**0.734**
U4.2 #Received actions	0.594	**0.975**	0.867	0.400	0.491	0.857	0.506
U4.3 #Total interactions	0.596	0.974	0.922	0.380	0.463	0.530	0.648

symbols for formatting the message (P2.5), followed by other stylistic-related metrics. Another relevant metric is the repetition of the tweet text: the more a tweet is repeated (not re-tweeted) in the stream the more relevant. Notice that this is valid for the automotive domain where intensive marketing campaigns are frequently performed. In many other domains, this metric would indicate instead low quality because it implies data redundancy.

The most relevant quality metric associated to users (Table 8) is the number of tweets related to the domain (U1.1), which have a high impact in both all users and influencers. The coherence of the description with respect to the domain has also a high impact in quality, being the best metric for the professional category. We can see that the different user categories present different quality metrics profiles, showing thus different behaviors in the data stream. Consequently, the assessment of user quality cannot only rely on fixed quality metrics; instead, it should consider metrics that align with each specific user role.

Regarding the influencers, we observe that most quality metrics have a high impact. The interaction metrics (U.4.2 and U.4.3) obtain the maximum impact, being also quite high the reputation metrics (U.3.4 and U.3.6) and the posting activity in the stream (U.1.1). The most similar profile to influencers is that of employee (E), which mainly comprises community managers. However, this profile shows much lower quality in reputation metrics than influencers.

Finally, the TQC assesses the quality of the specific analytical tasks (topics) chosen by the experts. This cube captures the quality indicators that can be associated to the subset of facts represented by each topic. As an example, we have selected five analytical topics of the two domains at hand. We have included some fake or non-relevant topics that mainly correspond to memes or noisy expressions. Table 9 shows the statistics of the corresponding TQCs for the two domains. Notice that non-relevant topics always poorly cover the reference users in contrast to true topics. At this point, we should reject all topics with low coverage for reference users (shadow boxes in Table 9). The topic "Vendo Opel Corsa" is an example of noisy expression that have two different meanings: the literal sense "Opel Corsa for sale" and the ironic expression "it matters little to me". In this case, the latter usage prevails, resulting in low coverage of reference users. Notice also that the "natural disaster & tourism" data stream contains the false topic "Gran Turismo", which corresponds both to a car model and a videogame. This topic ranks second in terms of the number of tweets in this stream, contributing to a high level of noisy data. Considering both domains, we can see that the threshold to consider a topic as of low coverage for reference users is very different in each case. For the automotive domain it can be set at 1% (shadow boxes in the left part of Table 9), whereas for the second domain the threshold is in 0.1%, Notice that the coverage values for the reference users in the second domain are lower because the number of users of reference is smaller (7386 vs. 28,165).

Notice that TQC also gives us clues about the coverage of interesting topics of the data stream. Regarding the "natural disasters & tourism" domain, we can see that the topic "Cyclone Idai" has a low coverage in this data stream. This is due to how the data stream has been defined, which do not include any keyword related to specific disasters or events. To increase the coverage of these topics we need to redefine the set of keywords in a similar way than in the automotive domain.

Table 10 shows the results of the TQC for these topics taking as main quality criteria the coherence of the user profiles and posts (U2.1 and P2.1). That is, we rank Q-facts according to these two criteria and then evaluate the corresponding AP_{rel} scores. As expected, all topics deemed as relevant have as main relevant voices the professional and the journalist categories.

Table 9 Statistics of the topic examples used for the TQC (* *fake/non-relevant topics*)

Automotive Domain				Natural Disasters & Tourism Domain			
Topics	#Tweets	#Users	%Refer. Users	Topics	#Tweets	#Users	%Refer. Users
Car recalls	10,603	5,523	7.9%	Migrants	176,798	104,709	0.2%
Car repair	16,771	2,834	5.6%	*Gran Turismo**	106,282	59,152	0.03%
Sell accessories	108,537	4,820	2.3%	Tourism industry	85,966	55,777	1.2%
Stolen cars	2,381	1,787	2.3%	Disaster relief	83,835	59,978	0.8%
New models	722	486	22%	*Recipe for disaster**	76,102	68,818	0.08%
*Vendo Opel Corsa**	3,674	2,604	0.3%	Disaster management	44,403	29,256	1.3%
*Harrison Ford**	455	248	0.4%	Disaster response	25,558	17,471	2%
*ADHD vs. focus joke**	253	246	0.4%	Cyclone Idai	4,139	3,309	1.7%

Table 10 TQC for the relevant topic examples (%Relevant > 1) in the automotive domain

Analysis Topics (U2.1 and P2.1)	All categories	Influential users	User categories				
			Business users		Interested users		
			E	P	PS	J	L&F
Car recalls	0.613	0.633		0.598	0.190	0.516	
Car repair	0.866	0.224	0.166	0.875	0.264	0.156	
Sell accessories	0.673	0.877	0.103	0.768		0.398	0.542
Stolen cars	0.911	0.680		0.926		0.684	
New models	0.527	0.562	0.189	0.516		0.468	

5.3 Quality Indicators

Tables 11 and 12 present the results of the values of the quality indicators for users and posts respectively. Examining Table 11, we can deduce that a few users post most of the information (metric "on domain"), with a good degree of interactions and an acceptable coherence of their profiles. Regarding the reputation perspective, we can conclude that posting users are not influential but receive a good number of interactions (metric U4.2).

Table 12 reports the scores for usefulness and completeness of the automotive domain. It is noteworthy that only 31% of posts in the stream involve business dimensions. Most of these posts express associations between sentiment words and dimension values (29% of total posts). To measure completeness, we focus on a particular topic (e.g., car recalls) and measure the coverage of the different cuboids of interest for this topic. Table 12 shows that most of the posts in the topic at least mention the model or the brand. Facts involving all dimensions are covered by 30% of the posts in the topic.

Tables 13 and 14 report the quality indicators obtained for the "natural disasters" domain. In this case, the stream exhibits slightly lower user credibility compared to the previous domain but significantly higher user reputation. It is worth mentioning that this stream involves the main disaster relief organizations and tourist destinations. Regarding the Table 13, we can see much lower scores than in the automotive domain, indicating that most posts in this stream are not relevant to the analysis goals. For measuring completeness, we chose the topic "natural disaster management", which covers posts related to damage, location and main organizations involved in the recovery. Table 14 reveals that even when narrowing down the dataset to the specific topic of interest, the results are substantially worse than in the automotive scenario, indicating lower data quality. In this scenario, it makes no sense to measure the coverage of facts with polarity since many words involved in the dimensions have a negative polarity (e.g., disaster, victims, damage, etc.)

5.4 Filtering by Ranking

As a final step, we need to filter out low-quality data from the selected topics of analysis. In this case, we aim at identifying the main voices of the topic and then apply the best quality criteria to them. Let us illustrate this process with the analytical topic "car recalls" from the automotive domain. In this topic, we want to analyze car brands and models affected by recalls due to known manufacturing defects. The TQC in Table 10 shows us that the main business roles for this topic are Professional and Journalist. Tables 7 and 8 show us that the best criteria for these roles are P1.5 and U4.2 for journalist, and P2.2 and U2.1 for professional. Moreover, in the column Cut of Table 11, we can get the thresholds for some of these metrics.

Once the thresholds are applied, we can then rank the remaining facts according to the previous chosen criteria. In this case, we apply the criteria in sequential order from highest to lowest relevance to obtain the final classification. Finally, we set up a cut-off point to reject low-quality facts for this topic. As an example, Table 15 shows the 4-top and

Table 11 User quality indicators for the automotive domain

Perspective	Metric	Impact	Q Value	Coverage	Cut	Score
Credibility	U1.1. #Tweets in domain	0.749	0.871	0.776	> 6	**0.61**
	U2.1 Coherence	0.759	0.763	0.747	< 17	0,57
	U4.3 Total interactions	0.596	0.790	0.47	> 11	0.35
Reputation	U3.4 #Followers	–	–	–	–	–
	U4.2. #Received actions	0.595	0.897	0.41	> 9	**0.33**

Table 12 Post quality indicators for the automotive domain

Perspective	Dimensions	Coverage
Usefulness	All dimensions	31%
	All dims. & Polarity	29%
Completeness (Topic "car recalls")	Model or Brand	**97%**
	Model & Part	49%
	Model & Defect	44%
	Model & Part & Defect	30%

Table 13 User quality indicators for the "natural disasters" domain

Perspective	Metric	Impact	Q Value	Coverage	Cut	Score
Credibility	U1.1. #Tweets in domain	0.655	0.827	0.782	> 8	0.53
	U2.1 Coherence	0.733	0.791	0.628	< 26	0.51
	U4.3 Total interactions	0.692	0.814	0.777	$>$, 20	**0.55**
Reputation	U3.4 #Followers	0.807	0.63	0.61	> 320	0.50
	U4.2. #Received actions	0.822	0.78	0.541	> 1	**0.53**

Table 14 Post quality indicators for the "natural disasters" domain

Perspective	Dimensions	Coverage (%)
Usefulness	All relevant dimensions	23%
Completeness	Natural Disaster	21%
	Damage	25%
	Natural Disaster & Damage	5.4%
	Natural Disaster & Organization	3.6%
	Natural Disaster & Location	12%

4-bottom ranked posts. Top positions feature news reports on recalls of various brands and models by qualified users, while bottom positions include comments and opinions related to recalls but not reporting them.

6 Conclusions

The research presented here focuses on assessing data quality in social business intelligence (SoBI) applications. In this paper, we aimed at defining an integrated multidimensional view to capture the impact of quality metrics for the different types of facts extracted from social networks. This integrated model encompasses the main aspects pointed out in the literature, namely: credibility, reputation, usefulness, and completeness. The main components of the proposed model cover the main facts included in SoBI applications, namely: posts, users, and topics. As a main novelty, we claim that users must be the keystone in quality assessment. Therefore, we introduce the user role dimension to better understand the origin of the data, and to measure its impact on the quality metrics. This claim is in accordance with previous work that recommends adding a

Table 15 Top and bottom ranked posts for the topic "car recall"

Posts in Top Positions
ford recalls transit vans for air bags … {lnk} {lnk}
nissan recalls nearly 640,000 u.s. cars: nissan pathfinder, rogue, infiniti jx35, qx60: nissan north america has issued two separate {dots}
ford escape, transit connect recalled for dimwitted dash: -ford is recalling certain 2014 and 2015 escape suvs and transit connect va {dots}
ford escape, transit connect recalled for dimwitted dash: -ford is recalling certain 2014 and 2015 escape suvs and transit connect va {dots}
Posts in Bottom Positions
i5gornascimento pll @i5gornascimento mitsubishi faz recall do pajero full para trocar {qmark} airbag mortal {qmark} no brasi {dots} {lnk}
oligarcs like mitsubishi distributor wil not get away w thr excuses on thr montero.duterte wil make thm pay,recall wil b mandatory
stp revenue theft wt nigeria recalls yaris,hilux ova faulty airbags.shell faces risks 4rm $1.1bn nigrian oil scndl
{lnk} more recalls {punct} #sellcar #cardealer #buymycar #dealerbid #usedcars #cardeals #e4drive

user-domain dimension to credibility assessment to enhance the understanding of users' interest (Abu-Salih et al. 2019; Arenas-Márquez et al. 2021).

We have carried out experiments over two different long-term data streams for two separate domains. These experiments show the usefulness of the approach in revealing the main quality aspects that characterize each of these data streams. The primary distinctions between these streams are attributed to their data collection methodologies. Whereas in the automotive domain we used a large set of keywords for each car model, in the "natural disaster & tourism" we just used a few abstract keywords. As expected, quality indicators show better scores for the automotive domain, but not in all of the aspects. Users in the automotive domain exhibit lower scores in reputation metrics compared to those in the other data stream. Finally, the Q-cubes of the multidimensional model allowed us to better understand the quality features of our datasets and propose effective ways to select topics and filter out low-quality data.

The proposed methodology allows us to address any other analysis domain following the steps designed for it. Firstly, analysts must define a reference set of users and the keywords for retrieving the posts of interest. Since this step is exploratory, it assumes some knowledge of the application domain to identify an initial set of relevant users and a good choice of retrieval keywords. Once these two elements are defined, the proposed method automatically constructs all the analytical facts and calculates the quality indicators along with their impact. By selecting the top impact quality indicators, low-quality data can be filtered out. The entire process can be then refined by updating the retrieval query's keywords and/or the reference set of users. These updates can be suggested after the analysts have inspected the resulting dataset in the previous iteration, as well as by applying previous domain knowledge. Thus, new relevant users can be identified, and incomplete dimensions may require further keywords in the retrieval query.

The main practical implication of this study is that the proposed method allows analysts to measure the quality of the processed social media data from different perspectives and considering the profiles of the users that generate the contents. We demonstrate that data quality heavily depends on the domain and topics at hand, requiring the combination of different metrics according to their impact, different thresholds, and different filtering criteria. Our analysis of metrics in relation to user categories has provided us with valuable insights into the characteristics of the generated data, enabling us to formulate more effective strategies for filtering high-quality data.

This proposal has several limitations that need further research. The first is related to the language models used to measure coherence, usefulness, and completeness. In this paper, we used a simple approach by just taking the word distributions of the dimension values (e.g., car models, defects, natural disasters, etc.) However, this approach will depend heavily on the richness of the available metadata for analysis. Therefore, in complex domains with scarce linguistic resources, we will need new methods to achieve accurate results. In future research, we intend to explore advanced methods that leverage semantic annotations (Berlanga et al. 2015; Lanza-Cruz et al. 2018) and apply NLP sentence encoders (Reimers and Gurevych 2019) to enhance the precision and reliability of quality assessment.

Another limitation of our approach is the reliance on ad-hoc methods to construct reference collections of relevant users for the selected domains. These methods, which predominantly depend on predefined rules applied to screen names, profile descriptions, and expert-curated external resources, can be resource-intensive and do not provide the scalability and reliability required. To overcome this limitation, we need to develop more automated and robust techniques. Our initial approach involves a user classification into business roles, a process that can be further refined and automated through the implementation of advanced NLP text classifiers (Nebot et al. 2018; Lanza-Cruz et al. 2023). This will facilitate the identification of relevant users aligned with specific analytical goals.

Finally, our future research agenda also includes exploring innovative approaches to combine quality metrics effectively, with the aim of maximizing the AP (average precision) metric. Methodologies such as fastAP, developed for image retrieval (Cakir et al. 2019), can be adapted to our domain. However, it is essential to note that fastAP requires a pool of negative examples, a challenge we intend to address by using the provided set of reference users.

Acknowledgements This research has been partially funded by the Spanish Ministry of Science under grants PID2021-123152OB-C22 and PDC2021-121097-I00 both funded by the MCIN/AEI/ https://doi.org/10.13039/501100011033 and by the European Union and FEDER/ERDF (European Regional Development Funds). We also would like to thank valgrAI (Valencian Graduate School and Research Network of Artificial Intelligence) foundation for their support.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

Abu-Salih B, Wongthongtham P, Beheshti S, Beheshti B (2015) Towards a methodology for social business intelligence in the era of big social data incorporating trust and semantic analysis. In: 2nd International conference on advanced data and information engineering. Springer, Heidelberg

Abu-Salih B, Bremie B, Wongthongtham P, Duan K, Issa T, Chan KY, Alhabashneh M, Albtoush T, Alqahtani S, Alqahtani A, Alahmari M, Alshareef N, Albahlal A (2019) Social credibility incorporating semantic analysis and machine learning: a survey of the state-of-the-art and future research directions. In: Barolli L et al (eds) Web, artificial intelligence and network applications. Springer, Cham, pp. 87–100. https://doi.org/10.1007/978-3-030-15035-8_87

Abu-Salih B, Chan K. Y, Al-Kadi O, Al-Tawil M, Wongthongtham P, Issa T, Saadeh H, Al-Hassan M, Bremie B, Albahlal A (2020) Time-aware domain-based social influence prediction. Int J Big Data 7, Article 10. https://doi.org/10.1186/s40537-020-0283-3

Alrubaian M, Al-Qurishi M, Alamri A, Al-Rakhami M, Hassan M, Fortino G (2019) Credibility in online social networks: a survey. IEEE Access 7:2828–2855

Amigó E, Carrillo-de-Albornoz J, Chugur I, Corujo A, Gonzalo J, Meij E, de Rijke M, Spina D (2014) Overview of RepLab: author profiling and reputation dimensions for online reputation management. In: Kanoulas E et al (eds) Information access evaluation. Multilinguality, multimodality, and interaction. https://doi.org/10.1007/978-3-319-11382-1_24

Aramburu MJ, Berlanga R, Lanza I (2021) Quality management in social business intelligence projects. In: Proceedings of the 23rd International Conference on Enterprise Information Systems, pp 320–327. https://doi.org/10.5220/0010495703200327. https://www.scitepress.org/Papers/2021/104957/104957.pdf

Arenas-Márquez F, Martinez-Torres R, Toral S (2021) Convolutional neural encoding of online reviews for the identification of travel group type topics on TripAdvisor. Inf Proc Manag 58(5). https://doi.org/10.1016/j.ipm.2021.102645

Arolfo F, Cortés-Rodriguez K, Vaisman A (2022) Analyzing the quality of Twitter data streams. Inf Syst Front 24(1):349–369

Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval. Addison-Wesley

Bansal P, Bansal R, Varma V (2015) Towards deep semantic analysis of hashtags. ECIR. https://doi.org/10.1007/978-3-319-16354-3_50

Berardi G, Esuli A, Marcheggiani D, Sebastiani F (2011) ISTI@-TREC Microblog Track: Exploring the use of hashtag segmentation and text quality ranking. https://trec.nist.gov/pubs/trec21/papers/NEMIS_ISTI_CNR.microblog.final.pdf. Accessed 15 Jul 2022

Berkani N, Bellatreche L, Khouri S, Ordonez C (2019) Value-driven approach for designing extended data warehouses. DOLAP. http://ceur-ws.org/Vol-2324/Paper25-NBerkani.pdf. Accessed 15 Jul 2022

Berlanga R, García-Moya L, Nebot V, Aramburu MJ, Sanz I, Llidó DM (2015) SLOD-BI: An open data infrastructure for enabling social business intelligence. Int J Data Wareh Min 11(4):1–28. https://doi.org/10.4018/ijdwm.2015100101

Berlanga R, Lanza-Cruz I, Aramburu MJ (2019) Quality indicators for social business intelligence. In: 6th International Conference on Social Networks Analysis, Management and Security, Granada, pp 229–236. https://doi.org/10.1109/SNAMS.2019.8931862

Birjali M, Kasri M, Beni-Hssane B (2021) A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl-based Syst 226

Cai L, Zhu Y (2015) The challenges of data quality and data quality assessment in the big data era. Data Sci J 14, Article 2

Cakir F, He K, Xia X, Kulis B, Sclaroff S (2019) Deep metric learning to rank In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1861–1870, https://doi.org/10.1109/CVPR.2019.00196

Chauhan U, Shah A (2021) Topic modeling using latent dirichlet allocation: a survey. ACM Comput Surv 54(7)

Choi J, Yoon J, Chung J, Coh B-Y, Lee J-M (2020) Social media analytics and business intelligence research: A systematic review. Inf Proc Manag 57(6). https://doi.org/10.1016/j.ipm.2020.102279

Crawford M, Khoshgoftaar TM, Prusa JD, Richter AN, Al Najada H (2015) Survey of review spam detection using machine learning techniques. J Big Data 2(23). https://doi.org/10.1186/s40537-015-0029-9

Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M (2015) Fame for sale: Efficient detection of fake Twitter followers. Decis Support Syst 80:56–71

Czernek A (2018) Social measurement depends on data quantity and quality. Millward Brown Dynamic Logic. https://cupdf.com/document/social-measurement-depends-on-data-quantity-and-2014-07-17-social-measurement.html. Accessed 15 Nov 2022

Duan Y, Zhimin C, Furu W, Ming Z, Shum H (2012) Twitter topic summarization by ranking tweets using social influence and content quality. In: Proceedings of the 24th International Conference on Computational Linguistics, pp 763–780. https://www.aclweb.org/anthology/C12-1047

Francia M, Gallinucci E, Golfarelli M, Rizzi S (2016) Social business intelligence in action. In: Nurcan S et al (eds) Advanced information systems engineering. Lecture Notes in Computer Science, vol 9694. Springer, Cham

Gallinucci E, Golfarelli M, Rizzi S (2015) Advanced topic modeling for social business intelligence. Inf Syst 53:87–106

García-Moya L, Kudama S, Aramburu MJ, Berlanga R (2013) Storing and analysing voice of the market data in the corporate data warehouse. Inf Syst Front 15:331–349. https://doi.org/10.1007/s10796-012-9400-y

Gioti H, Ponis S, Panayiotou N (2018) Social business intelligence: review and research directions. J Intell Stud Bus 8:23–42. https://doi.org/10.37380/jisib.v8i2.320

Goonetilleke O, Sellis T, Zhang X, Sathe S (2014) Twitter analytics: a big data management perspective. ACM SIGKDD Explor Newsl 16(1):11–20

Gröger C (2021) There is no AI without data. Commun ACM 64(11):98–108. https://doi.org/10.1145/3448247

Gupta A, Kumaraguru P, Castillo C, Meier P (2014) TweetCred: real-time credibility assessment of content on Twitter. In: Proceedings of the 6th International Conference on Social Informatics, pp 228–243. https://doi.org/10.1007/978-3-319-13734-6_16

Hammou B, Lahcen A, Mouline S (2020) Towards a real-time processing framework based on improved distributed recurrent neural network variants with fastText for social big data analytics. Inf Proc Manag 57(1). https://doi.org/10.1016/j.ipm.2019.102122

Han X, Wang L, Liu G, Zhao D, Xu S (2017) Occupation profiling with user-generated geolocation data. In: 2nd International

Conference on Knowledge Engineering and Applications, pp 93–97. https://doi.org/10.1109/ICKEA.2017.8169908
Hernandez M, Hildrum K, Jain P, Wagle R, Alexe B, Krishnamurthy R, Stanoi IR, Venkatramani C (2013) Constructing consumer profiles from social media data. In: IEEE International Conference on Big Data, pp 710–716. https://doi.org/10.1109/BigData.2013.6691641
Holsapple C, Hsiao S, Pakath R (2018) Business social media analytics: characterization and conceptual framework. Decis Support Syst 110:32–45. https://doi.org/10.1016/j.dss.2018.03.004
Hu S, Kumar A, Al-Turjman F, Gupta S, Seth S, Shubham, (2020) Reviewer credibility and sentiment analysis based user profile modelling for online product recommendation. IEEE Access 8:26172–26189. https://doi.org/10.1109/ACCESS.2020.2971087
Immonen A, Pääkkönen P, Ovaska E (2015) Evaluating the Quality of Social Media Data in Big Data Architecture. IEEE Access 3:1–1. https://doi.org/10.1109/ACCESS.2015.2490723
Johannesson P, Perjons E (2014) An introduction to design science. Springer, ISBN: 978–3–319–10632–8
Jöhnk J, Weißert M, Wyrtki K (2021) Ready or not, AI comes – an interview study of organizational AI readiness factors. Bus Inf Syst Eng 63:5–20. https://doi.org/10.1007/s12599-020-00676-7
Kaufhold M-A, Christian M (2020) Rapid relevance classification of social media posts in disasters and emergencies: a system and evaluation featuring active, incremental and online learning. Inf Proc Manag 57(1). https://doi.org/10.1016/j.ipm.2019.102132
Keegan B, Rowley J (2017) Evaluation and decision-making in social media marketing. Manag Decis 55:15–31. https://doi.org/10.1108/MD-10-2015-0450
Kimball R, Ross M (2013) The data warehouse toolkit, 3rd edn. Wiley, p 48. ISBN 978–1–118–53080–1
Kolajo T, Daramola O, Adebiyi A, Seth A (2020) A framework for pre-processing of social media feeds based on integrated local knowledge base. Inf Proc Manag 57(6). https://doi.org/10.1016/j.ipm.2020.102348
Lanza-Cruz I, Berlanga R, Aramburu MJ (2023) Multidimensional author profiling for social business intelligence. Inf Syst Front. https://doi.org/10.1007/s10796-023-10370-0
Lanza-Cruz I, Berlanga R, Aramburu MJ (2018) Modeling analytical streams for social business intelligence. Inform 5:33. https://doi.org/10.3390/informatics5030033
Lauriola I, Lavelli A, Aiolli F (2022) An introduction to deep learning in natural language processing: models, techniques, and tools. Neurocomput 470:443–456
Lee I (2018) Social media analytics for enterprises: Typology, methods, and processes. Bus Horiz 61(2):199–210. https://doi.org/10.1016/j.bushor.2017.11.002
Lin J, Snow R, Morgan W (2011) Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 422–429. https://doi.org/10.1145/2020408.2020476
Nebot V, Rangel F, Berlanga R, Rosso P (2018) Identifying and classifying influencers in Twitter only with textual information. In: Nat Lang Proc Inf Syst 28–39. https://doi.org/10.1007/978-3-319-91947-8_3
Pääkkönen P, Jokitulppo J (2017) Quality management architecture for social media data. J Big Data 4(6). https://doi.org/10.1186/s40537-017-0066-7
Pasi G, Viviani M, Carton A (2019) A multi-criteria decision making approach based on the Choquet integral for assessing the credibility of user-generated content. Inf Sci 503:574–588. https://doi.org/10.1016/j.ins.2019.07.037
Păvăloaia V, Anastasiei I, Fotache D (2020) Social media and e-mail marketing campaigns: symmetry versus convergence. Symmetry 12(12):1940. https://doi.org/10.3390/sym12121940
Plachouras V, Stavrakas Y, Andreou A (2013) Assessing the coverage of data collection campaigns on Twitter: a case study. In: Demey Y, Panetto H (eds) On the move to meaningful internet systems. OTM 2013 Workshops. Lecture Notes in Computer Science vol 8186. https://doi.org/10.1007/978-3-642-41033-8_76
Podhoranyi M (2021) A comprehensive social media data processing and analytics architecture by using big data platforms: a case study of Twitter flood-risk messages. Earth Sci Inform 14. https://doi.org/10.1007/s12145-021-00601-w
Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks, In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
Rodríguez-Vidal J, Gonzalo J, Plaza L, Anaya-Sánchez H (2019) Automatic detection of influencers in social networks: authority versus domain signals. J Assoc Inf Sci Technol 70:675–684. https://doi.org/10.1002/asi.24156
Ruhi U (2014) Social media analytics as a BI practice: current landscape & future prospects. J Internet Soc Netw Virtual Commun. https://doi.org/10.5171/2014.920553
Sabherwal R, Becerra-Fernandez I (2013) Business intelligence: Practices, technologies, and management. Wiley
Sadiq S, Indulska M (2017) Open data: Quality over quantity. Int J Inf Manag 37:150–154. https://doi.org/10.1016/j.ijinfomgt.2017.01.003
Salvatore C, Biffignandi S, Bianchi A (2021) Social media and Twitter data quality for new social indicators. Soc Indic Res. https://doi.org/10.1007/s11205-020-02296-w
Saroj A, Pal S (2022) Use of social media in crisis management: a survey. Int J Disaster Reduct 48. https://doi.org/10.1016/j.ijdrr.2020.101584
Shankaranarayanan G, Blake R (2017) From content to context: the evolution and growth of data quality research. J Data Inf Qual 8:1–28. https://doi.org/10.1145/2996198
Sikdar S, Kang B, ODonovan J, Höllerer T, Adah S (2013) Understanding information credibility on Twitter. In: International Conference on Social Computing, Alexandria, pp 19–24. https://doi.org/10.1109/SocialCom.2013.9
Stieglitz S, Dang-Xuan L, Bruns A, Neuberger C (2014) Social media analytics. Bus Inf Syst Eng 6:89–96. https://doi.org/10.1007/s12599-014-0315-7
Stieglitz S, Mirbabaie M, Ross B, Neuberger C (2018) Social media analytics – Challenges in topic discovery, data collection, and data preparation. Int J Inf Manag 39:156–168
Tilly R, Posegga O, Fischbach K, Schoder D (2017) Towards a conceptualization of data and information quality in social information systems. Bus Inf Syst Eng 59:3–21. https://doi.org/10.1007/s12599-016-0459-8
Viviani M, Pasi G (2017) Credibility in social media: opinions, news, and health information – A survey. WIREs Data Mining Knowl Discov 7(5). https://doi.org/10.1002/widm.1209
Zachlod C, Samuel O, Ochsner A, Werthmüller S (2022) Analytics of social media data – State of characteristics and application. J Bus Res 144:1064–1076. https://doi.org/10.1016/j.jbusres.2022.02.016
Zhang R, Indulska M, Sadiq S (2019) Discovering data quality problems. Bus Inf Syst Eng 61:575–593. https://doi.org/10.1007/s12599-019-00608-0
Zheng L (2021) The classification of online consumer reviews: a systematic literature review and integrative framework. J Bus Res 135. https://doi.org/10.1016/j.jbusres.2021.06.038