Social Network Analysis and Mining (2023) 13:150
https://doi.org/10.1007/s13278-023-01148-y

ORIGINAL ARTICLE

Fulmqa: a fuzzy logic-based model for social media data quality assessment

Oumaima Reda[1] · Ahmed Zellou[1]

Received: 7 June 2023 / Revised: 26 September 2023 / Accepted: 27 September 2023 / Published online: 8 November 2023

Abstract

Nowadays, with the advancement of information technology and the growing importance of social media, social media platforms such as Twitter and Facebook have become deeply entrenched in our lives and are growing rapidly around the world. On these platforms, users have the freedom to share and publish whatever they want without control, leading to faster generation and dissemination of information, as a result, rumors and false information can then be spread and seen by a larger number of people. Consequently, assessing the quality of these data proves to be of major importance. In this paper, we aim to present a new and efficient quality assessment model including the quality metrics needed for an accurate assessment. Our work presents an extension of the SMDQM (Social Media Data Quality Model) (Reda and Zellou, in: International conference on innovative research in applied science, engineering and technology, IEEE, 2022), which is used to assess the quality of data provided by social media platforms. We suggest using fuzzy logic to describe data quality metrics in order to overcome imprecision and subjectivity. Next, we perform extensive experiments to evaluate the performance of our model using a real-world implementation by performing an evaluation on two separate Twitter data sets. The findings indicate that the model can successfully evaluate all tweets with high performance.

Keywords Data quality · Quality assessment · Data model · Social media · Fuzzy logic

1 Introduction

In an increasingly linked world where information moves swiftly and reaches a vast number of people, social media has developed dramatically in recent years. Currently, social media data are growing explosively, they are becoming more and more complicated and diversified, and this is due to the vast population of social media users, as well as the sheer volume of user-generated content (Gabr et al. 2021).

Social media has allowed any individual with access to the internet to publicly express their opinion, to be free to share and publish whatever they want (El Alaoui et al. 2019). This kind of liberty creates various data issues that can lead to a degraded quality of data, and it seems plausible that any poor quality of data could have a negative impact. Typically, it is widely known that data quality problems, such as incompleteness, redundancy, and inconsistency, make useless the overall process of using and processing this data. Therefore, getting a high level of data quality is deemed to be one of the most significant challenges, this level of quality is assessed using various dimensions and quality metrics. According to (Alizamini et al. 2010), data will be of high quality, if it is suitable for decision making, applications, and planning. Inspecting data to assess its present quality and the possible scope of any data quality issues is the goal of analyzing the quality of social media data (Reda and Zellou 2022; Woodall and Parlikad 2010). This paper's goal is to outline how to measure and assess the quality of social media data using a set of quality metrics already defined during the construction of the SMDQM model presented in Reda and Zellou (2022). For this, a thorough review of the literature is conducted to collect existing quality models in order to identify the most commonly used dimensions needed to assess the quality of social media data in a reliable way. In this study, we introduce a novel quality model for the assessment of social media data quality, in which we select eight quality metrics which are credibility, timeliness, popularity, relevancy,

Oumaima Reda
oumaima_reda@um5.ac.ma

Ahmed Zellou
ahmed.zellou@um5.ac.ma

[1] Software Project Management Team (SPM), ENSIAS, Mohammed V University in Rabat, Rabat, Morocco

accessibility, reliability, presentation and accuracy. We suggest modeling data quality metrics with fuzzy logic to address imprecision and subjectivity, allowing us to describe data quality requirements using a set of linguistic expressions on quality measurements. Then, we perform extensive experiments to evaluate the SMDQM model; to do that, we chose Twitter because it is one of the most widely used social media platforms for consumers to share their opinions, complaints, concerns and compliments about a product or service, as well as because of its attributes that are significant for data analysis.

This paper is structured according to the following: Section 2 defines the basic terms used in this work, and the state-of-the-art of data quality, quality dimensions and metrics, quality model and quality assessment. Section 3 presents the background of fuzzy logic. A detailed description of the implementation of the proposed model SMDQM is presented in Sect. 5. Section 6 presents experimental results. Finally, Sect. 7 concludes the paper by mentioning our future directions to approach them in the coming up works.

2 Preliminaries

2.1 Data

In the literature, various views exist regarding the definition of data, such as the raw materials needed for information, or as a set of facts. Elmasri and Navathe (2000) defined data as facts with implicit meaning that may be measured, recorded, and documented. According to Lee et al. (2006), data can be considered as a valuable asset. Typically, an asset can be either tangible, such as place or location, or intangible, such as knowledge and methodologies. Therefore, a data is an elementary description of a reality or an abstract notion, which has not yet been interpreted and put into context. In general, there are three types of data to consider: structured, semi-structured and unstructured data.

- Structured data are based on a predefined schema and conform to particular specifications (Sint et al. 2009). It can be stored in relational database system and thus fit into a clearly defined data model or structured files (i.e., csv).
- Semi-structured data are a kind of structured data but does not have the strict structure of the data model. They are frequently described as terms without schema or self-describing, without a distinct description of the data type or structure such as web page and XML file (Sint et al. 2009).
- Unstructured data lack a standardized identifiable form. Photographs and graphic images, videos, streaming data, and files, for example, cannot be stored in rows and columns in a relational database (Sint et al. 2009).

2.2 Quality

The quality is defined as the set of characteristics, properties that make something correspond well or poorly to its nature [9]. The International Organization for Standardization (ISO) defines it as all of an entity's attributes that confer the capacity to meet explicit and implicit standards [10].

2.3 Data quality

In data quality literature, there is no precise definition of the term data quality. However, a widely accepted way of defining data quality is as *fitness for use* (Tayi and Ballou 1998). it is considered as the degree to which the data meets the user's needs or is suitable for a specific process and depends hence on the context of use and the requirements of the users (Berti-Equille 1999).

Data quality is defined by the International Organization for Standardization (ISO) as the degree to which a certain collection of intrinsic characteristics satisfies a specified requirement or expectation, often inferred or required (Hoyle 2006), while (Deming 1982) and Juran (2003) defined it as the ability to be used in the context of decision making, activities, and planning, and it is considered to be directly related to context.

Generally, the quality of data is often defined as the adequacy of a given data and its properties for a specific use case, depending on the data consumer who uses it (Nikiforova 2020). It is conceived as a multidimensional concept and, over time, researchers have proposed various definitions, such as "*user satisfaction*" (Wayne 1983) and as "*conformity to requirements*" (Crosby 1979). According to Olson (2003), the concept of data quality can refer to the adequacy of data to fulfill the intended goal. It depends on the requirements that the user expects to execute or the data value that the user expects to obtain. This concept involves several dimensions, as different aspects have to be taken into account. These aspects are modeled using data quality dimensions that are assessed by certain metrics (Ehrlinger and Woss 2018).

2.4 Quality dimensions and metrics

The quality dimensions are an important part for assessing data quality, as these data quality can be evaluated using several dimensions and metrics. A quality dimension refers to a property of data quality that represents an aspect of these data and can be used to guide the process of understanding quality (Laranjeiro et al. 2015). According to Wang and Strong (1996), data quality dimension is a set of data quality

Table 1 Examples of data quality dimensions

QD	Description
Accuracy	Indicates that the data must correctly depict reality and originate from a credible source (Ardagna et al. 2018)
Timeliness	Refers to the degree to which the age of data is appropriate for the job at hand (Wang and Strong 1996)
Completeness	Defined as the extent to which data comprise all expected data values and are sufficiently large for the task at hand (Scannapieco 2006)
Consistency	Describes the extent to which data is delivered in the same manner and is consistent with earlier data (Wang and Strong 1996)

attributes that represent a unique aspect of data quality, i.e., each dimension concerns a specific aspect.

Managing data quality involves selecting a set of dimensions within a particular context, and then classifying them in order to establish a quality model with which we can evaluate the quality of our data and assess their relevance (Reda et al. 2020). Several dimensions of data quality are addressed in the literature, but the most commonly used are accuracy, timeliness, completeness, and consistency. Examples of quality dimensions are illustrated in Table 1.

A data quality metric is a quantifiable resource that defines how a dimension is measured (Reda and Zellou 2023). Within each quality model, a set of specific quality dimensions, quality sub-dimensions, quality measures, and the relationships between these dimensions and measures are described. The quality metrics defined in a quality model provide information about the dimensions and sub-dimensions of quality, providing relevant details around quality measures, such as definitions, formulas or scales, and generally a classification of various quality metrics is included in the quality model (Radulovic et al. 2018). In other words, quality metrics may be viewed as formulas for quantifying the fulfillment of dimensions in numerical values. While a quality dimension is a wide idea in general, a related metric helps to establish a specific meaning for the dimension. As a result, many metrics may be associated with the same quality dimension, and their application will assess a variety of aspects of the dimension (Arolfo et al. 2020). Table 2 shows several examples of quality measures.

Table 2 Examples of data quality metrics

QD	Metrics
Accuracy	Number of correct data/total number of data
Completeness	Number of non-missing data/total number of data
Consistency	Number of data that respects constraints/total number of data

2.5 Quality model

The goal of a data quality model is to provide a definition of the various data quality characteristics within a hierarchy, in order to know the level of quality of the data included in different systems, as well as to help the user create his own quality profile. Hence, a quality model is used as a reference to the quality measures to be evaluated (Reda and Zellou 2022). By specifying detailed quality measures, such as formulas, definitions, or scales, quality models indicate which measures are useful for evaluation and how they should be measured (Hitzler et al. 2016). Various researchers described data quality models as invaluable resources for quality assessment and, accordingly, they serve as a reference for the quality measures to be evaluated (Even and Shankaranarayanan 2009).

2.6 Data quality assessment

According to ISO definition, quality assessment describes the overall evaluation of data quality using established quality criteria. It might be the outcome of a quality scoring or grading procedure, either quantitative or qualitative scoring may be used [31]. Data quality assessment refers to the quality evaluation of meaningful data, where the context and intended use of the data are taken into account (Caballero et al. 2009). Typically, it entails the development of metrics to test and identify the present level of data quality; this evaluation is regarded as a component of data quality management (Reda and Zellou 2022). The first step toward any quality assessment is to define the quality requirements, which are based on data users and their needs. It then consists of measuring the relevant dimensions or quality criteria and comparing the results of this evaluation with the user's quality requirements (Radulovic et al. 2018).

3 Fuzzy logic background

It is frequently challenging to give precise definitions or descriptions of certain concepts and the relationships between them, especially in the context of data quality. Fuzzy sets are used in our study to reflect the quality of the data in order to tackle this issue.

The concept of fuzzy logic introduced by ZADEH (1965) has been implemented in various scientific and technological sectors. Fuzzy sets have been developed to represent and operate imperfect data while tolerating different kinds of uncertainty. Typically, there are three major processes involved: Fuzzification, fuzzy rules, and defuzzification are the first three steps. Figure 1 visualizes the overview of operation of fuzzy systems.

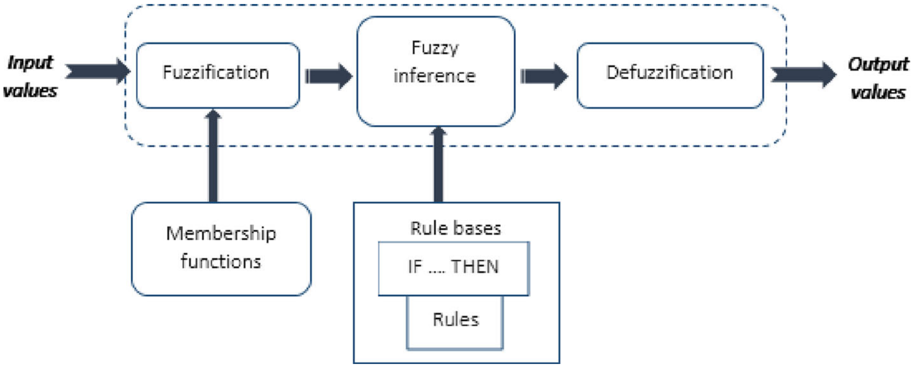


Fig. 1 Overview of the operation of fuzzy system

A fuzzy set is characterized by a membership function defined on the universe of discourse. This function specifies the degree to which a given element belongs to the fuzzy set by mapping it to a range spanning the interval [0,1]. When the membership function has a value of 0, it signifies that the associated element is not a part of the fuzzy set. When it has a value of 1, it fully belongs to the set. Values between 0 and 1 represent fuzzy members, which only partially belong to the fuzzy set (ROSS 2012). The formal definition of a fuzzy set A is:

$$A = \{(x, \mu_A(x)) | x \in X\} \tag{1}$$

where X is the discourse universe, μ_A is the membership function representing a fuzzy set A, and $\mu_A(x)$ is the membership degree of the element x in the fuzzy set A ($\mu_A(x)$ specifies how much x is compatible with the set A).

The fuzzification process is constructed using a variety of membership functions, both linear and nonlinear. This step quantifies the degree to which each input belongs to different categories (ROSS 2012).

The rule base, often referred to as fuzzy if-then rules, is included into language terms that are created by experts or that are based on data set extractions. Every rule is made up of a mathematical procedure that converts professional knowledge into murky if-then rules. The antecedent (the if section) and the consequent (the then part) are the two components of fuzzy rules (WANG et al. 2009).

In defuzzification step, the fuzzy output sets are converted back into crisp values or decisions. Various defuzzification methods can be used, such as centroid, height, or bisector, depending on the specific application.

4 SMDQM

In our previous study, as described in Reda and Zellou (2022), our objective was to develop a social media data quality model aimed at assessing the quality of data collected from various social media platforms. To achieve this, we proposed a comprehensive five-step methodology, as depicted in Fig. 2. The proposed model encompasses eight key quality dimensions, which were derived from a compilation of 60 dimensions collected from 13 different research sources. These eight dimensions were identified as the most prominent and significant factors based on their prevalence and relevance in the existing literature.

By establishing this social media data quality model, we sought to offer researchers and practitioners a robust framework to evaluate and ensure the reliability, accuracy, and usability of social media data. The model's effectiveness was validated through a rigorous evaluation process, and it holds great potential in advancing research in various domains reliant on social media data analysis.

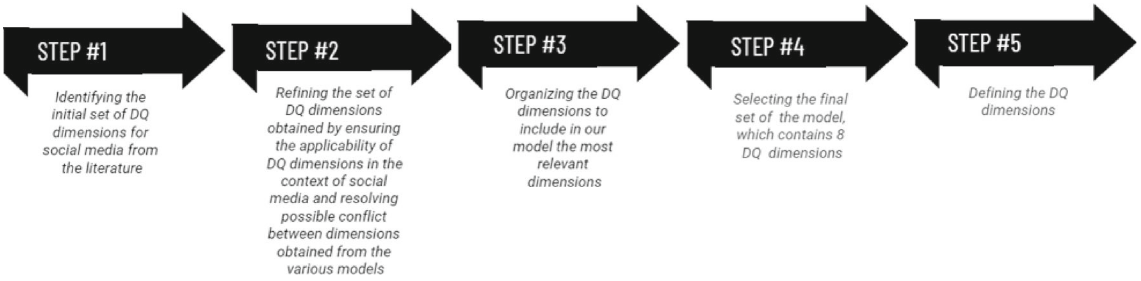

STEP #1
Identifying the initial set of DQ dimensions for social media from the literature
STEP #2
Refining the set of DQ dimensions obtained by ensuring the applicability of DQ dimensions in the context of social media and resolving possible conflict between dimensions obtained from the various models
STEP #3
Organizing the DQ dimensions to include in our model the most relevant dimensions
STEP #4
Selecting the final set of the model, which contains 8 DQ dimensions
STEP #5
Defining the DQ dimensions

Fig. 2 SMDQM model process

5 FULMQA: fuzzy Logic-based model for quality assessment

In the following sections, we describe in greater depth and provide quality metrics for each dimension of the novel data quality model, which includes credibility, timeliness, popularity, relevancy, accessibility, reliability, presentability, and accuracy.

5.1 Timeliness

Over time, correct data may become incorrect due to changing circumstances. Timeliness dimensions is defined in Cai and Zhu (2015) as the difference in time between the generation and acquisition of data and its use. It refers to the freshness of the data, and represents the measurement of whether the data is up-to-date enough for the given task (Chai et al. 2009; Immonen et al. 2015). A post closer to the time of the search, the more certain we are that the data are relevant (Reuter et al. 2015). The timeliness can be assessed by calculating the age metric which measures the age of a post; it is calculated by the difference between the current timestamp T_{now} and the time when the post was created T_{post}, with the following equation:

$$Age_{post} = T_{now} - T_{post} \tag{2}$$

Given T_0 which represents the age of the most recent post, and T_t the age tolerable, the timeliness of post p is calculated by:

$$T(p) = \begin{cases} 1, & \text{if } Age_{post} \leq T_0 \\ \frac{T_t - Age_{post}}{T_t - T_0}, & \text{if } T_0 < Age_{post} < T_t \\ 0, & \text{if } Age_{post} \geq T_t \end{cases} \tag{3}$$

When we have a recent post, we are more assured that the timeliness is high, i.e., the more recent the post, the more timeliness is high. Figure 3 shows the presentation of timeliness dimension.

5.2 Popularity

Popularity dimension can be defined by the use and reuse of a resource; this extensive use tends to lead to greater trust. When an author posts accurate data, that author has a particular number of followers, or the data are liked and then shared or circulated by other people (Immonen et al. 2015). According to the social context, the popularity can be assessed using

Fig. 3 Presentation of timeliness dimension

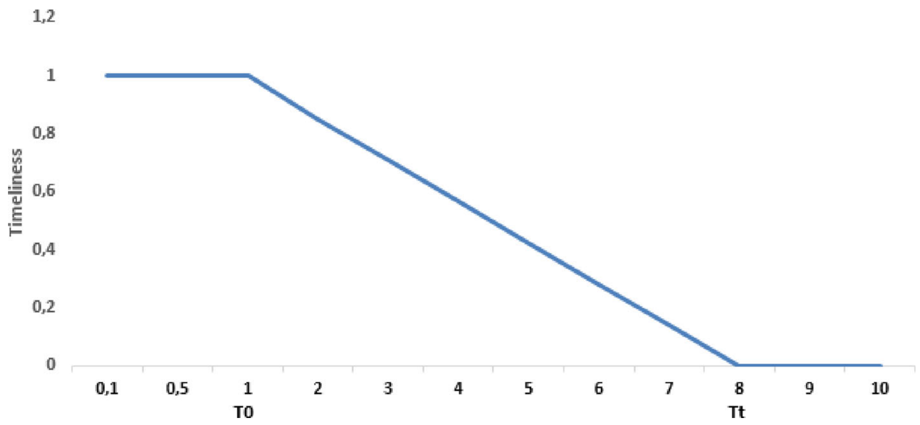
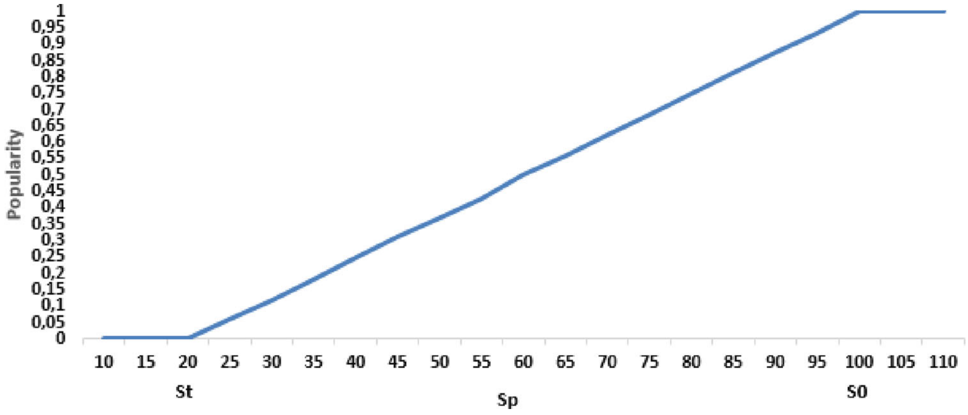


Fig. 4 Presentation of popularity dimension



a lot of criteria such as the number of followers, number of friends, number of comments, number of likes, number of retweets, and number of share. Based on several studies (Alrubaian et al. 2017), the criteria number of shares/retweets is considered to be one of the best indicators used to measure the popularity, because it can be very attractive for the user to use a post which has been shared or retweeted many times. To measure popularity, we use S_p to denote the number of shares or retweets of a post P, S_0 is the number of shares/retweets of the post that has the highest number of shares/retweets, and S_t the tolerable number of shares/retweets. Popularity can be calculated as follows:

$$P(p) = \begin{cases} 1, & \text{if } S_p \geq S_0 \\ 1 - \frac{S_0 - S_p}{S_0 - S_t}, & \text{if } S_t < S_p < S_0 \\ 0, & \text{if } S_p \leq S_t \end{cases} \quad (4)$$

The degree of popularity increases with the increasing number of the criteria given to a particular post or author, i.e., with the increasing number of shares/retweets of the post. Figure 4 shows the presentation of popularity dimension.

5.3 Credibility

Evaluating the credibility of social media data is a topic of broad and current interest. Whenever we refer to the credibility of data, we are directly asking whether this data is supposed to be believed or not, which means, if the data are credible, so we can believe it, whether it is real or not. According to Salvatore et al. (2020), it is about the different objective and subjective parts of the believability of a source or a message. Several research propose the assessment of credibility at user, and content levels (Gupta et al. 2019; Yang et al. 2019; Verma et al. 2019). However, in our study, we suppose that a credible users provide credible contents, so users with high-credibility are more likely to share believable contents.

Credibility of users can be measured based on the number of followers/friends of a user. For instance, a lot of users are likely to believe a Twitter user with many followers. We use F_p to denote the number of followers/friends of the user created the post P, and F_0 the number of followers/friends given to the user with the highest number of followers/friends and F_t the number of followers/friends tolerable. Credibility can be measured as follows:

$$C(p) = \begin{cases} 1, & \text{if } F_p \geq F_0 \\ 1 - \frac{F_0 - F_p}{F_0 - F_t}, & \text{if } F_t < F_p < F_0 \\ 0, & \text{if } F_p \leq F_t \end{cases} \quad (5)$$

The degree of credibility increases with the increasing number of followers/friends given to a particular author. Figure 5 shows the presentation of credibility dimension.

5.4 Reliability

In a social media context, all people can publish whatever they like and any kind of information, whether truthful or not. When we talk about data reliability, we logically imply whether the data are trustworthy or not; this means, if the data is reliable, we can trust it. For example, people can rely on the opinions of others through Twitter posts to know the quality of a movie; similarly, if a YouTube video has received a lot of views, many people may trust it. According to Chai et al. (2009), the reliability describes the degree to which the data is trustworthy and correct. There are some authors who frequently link reliability with four other dimensions like verifiability, credibility, popularity and reputation (Firmani et al. 2015; Arolfo et al. 2020). However, no social media platform, including Twitter, has the ability and want to verify all their users, so it is expected that the majority of users in social media are unverified. Based on this, we consider the following dimensions to define the reliability:

$$Reliability = Credibility \vee Popularity \vee Reputation \quad (6)$$

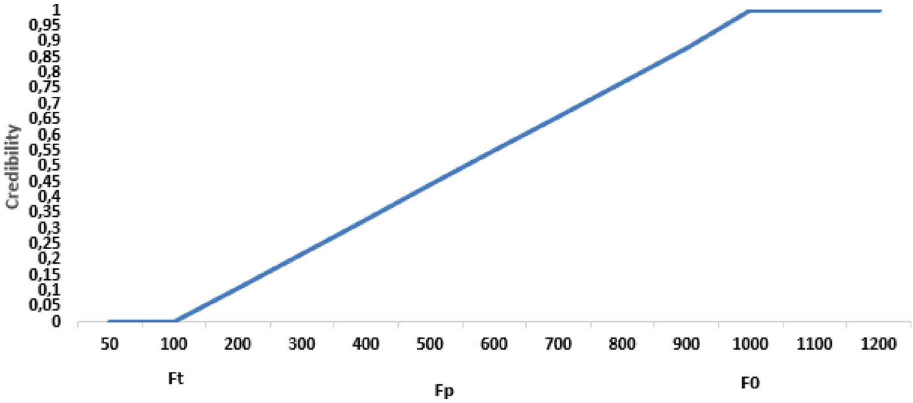


Fig. 5 Presentation of credibility dimension

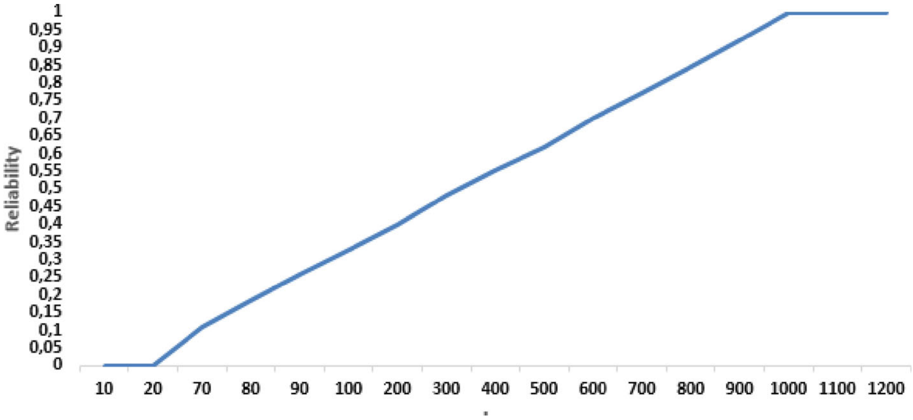


Fig. 6 Presentation of reliability dimension

Since we already evaluated the popularity and credibility in our study, we will focus on the reputation which is defined as a judgment done by a user to decide the Trustworthiness of a source (Arolfo et al. 2020). The reputation can be evaluated for example based on the number of followers/friends of an account, number of likes, registration date, or number of posts from the same author. We use L_p to indicate the number of likes given to the post P, and L_0 the number of likes of the post that has the highest number of likes, and L_t is the number of likes tolerable. The reputation can be calculated as follows:

$$Rep(p) = \begin{cases} 1, & \text{if } L_p \geq L_0 \\ 1 - \frac{L_0 - L_p}{L_0 - L_t}, & \text{if } L_t < L_p < L_0 \\ 0, & \text{if } L_p \leq L_t \end{cases} \tag{7}$$

Using the method proposed by (Larsen 1980), the final output of the reliability Rel(p) is the weighted average of all outputs of the 3 metrics (aggregation) as follows:

$$Rel(p) = \frac{m_1 C(p) + m_2 P(p) + m_3 Rep(p)}{3} \tag{8}$$

where m_1, m_2, m_3 are the coefficients. The degree of reliability increases with the increasing number of followers/friends, likes, and shares/retweets given to a particular post or author. Figure 6 shows the presentation of reliability dimension.

5.5 Relevancy

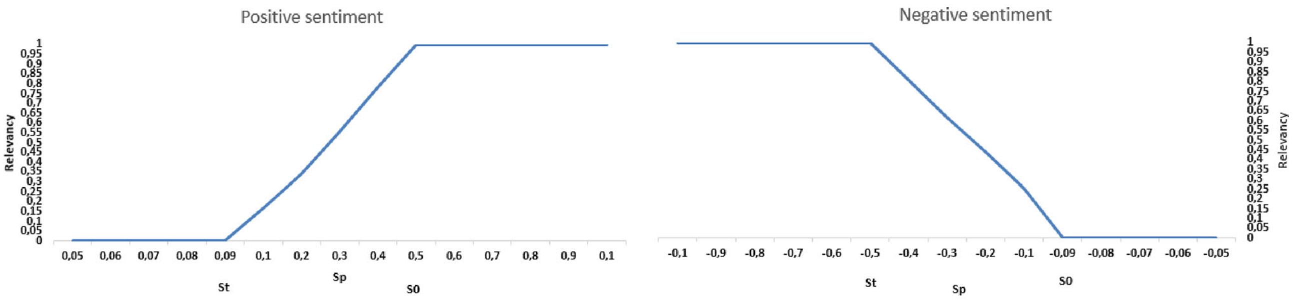
Generally, the relevancy dimension is related to the gains that a user can achieve by using the data; it is used to describe the extent to which these data are useful and applicable for the purpose (Chai et al. 2009; Immonen et al. 2015). In other words, it is the degree to which the data produced match the needs of the users (Salvatore et al. 2020). The relevancy could be measured by detecting a sentiment from the post, whether positive, negative or neutral. Typically, there are different methods for calculating the sentiment of a text. A sentiment analysis, for example, is used to derive meaning from text. It is one of the most widely used text classification techniques for analyzing and categorizing the sentiment underlying a communication, whether positive, negative, or neutral (Salvatore et al. 2020). In this study, we will calculate sentiment score by categorizing and counting the amount of negative and positive words from the supplied post, and then taking the ratio of the difference between positive and negative word counts and total word count. The following equation is used to determine sentiment score:

$$Sentiment = \frac{Positive_{words} - Negative_{words}}{Total_{words}} \tag{9}$$

If Sentiment score > 0, it is a positive post, if Sentiment score < 0, it is a negative post, and if Sentiment score $= 0$, it is a neutral post. Therefore, if a post expresses a positive or negative feeling, it will be considered as a relevant post, and if the sentiment is neutral, i.e., no sentiment could be computed by a Natural Language Processing (NLP) tool, it will be considered not relevant. Let S_p be the score of post P, and S_0 be the score associated with the post that has the highest score positive/negative, S_t is the tolerated score for the most relevant post positive/negative, and the relevancy can be calculated as follows:

$$R(p) = \begin{cases} 1, & \text{if } (S_p > 0 \text{ and } S_p \geq S_0) \text{ OR } (S_p < 0 \text{ and } S_p \leq S_0) \\ 1 - \frac{S_0 - S_p}{S_0 - S_t}, & \text{if } S_p > 0 \text{ and } S_t \leq S_p \leq S_0 \\ \frac{S_t + S_p}{S_t + S_0}, & \text{if } S_p < 0 \text{ and } S_t \geq S_p \geq S_0 \\ 0, & \text{if } (S_p > 0 \text{ and } S_t \geq S_p) \text{ OR } (S_p < 0 \text{ and } S_t \leq S_p) \end{cases} \tag{10}$$

The degree of relevancy increases with the increasing number of positive and negative words. Figure 7 shows the presentation of relevancy for negative and positive sentiment.



(a) (b)

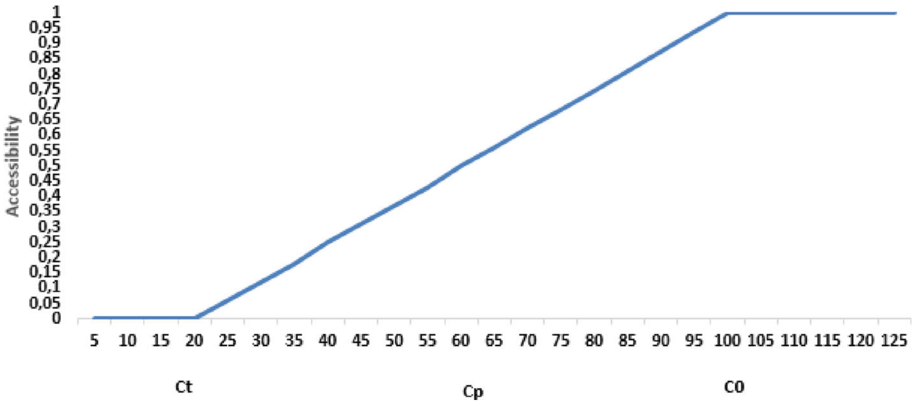
Fig. 7 Presentation of relevancy for negative **a** and positive **b** sentiment

5.6 Accessibility

The accessibility dimension is about the ability to retrieve the data, i.e., the ease of access. It could be closely correlated with how challenging users find it to access the data. In other words, it indicates how quickly and simply the data may be retrieved (Salvatore et al. 2020), and it also deals with the additional requirements to access the data, i.e., registrations that are required before accessing any social media platform. To determine the accessibility of a post, we must first determine whether or not users can find it, which can be accomplished by examining for instance the number of shares/retweets, likes, or comments. Specifically, if the post has been shared or commented on by several users, the post is now accessible to users. In our study, we use the numbers of comments to measure accessibility, we use C_p to indicate the number of comments given to the post p, and C_0 the number of comments of the post that has the highest number of comments, and C_t is the number of comments tolerable. The accessibility of p can be calculated as follows:

$$Acc(p) = \begin{cases} 1, & \text{if } C_p \geq C_0 \\ 1 - \frac{C_0 - C_p}{C_0 - C_t}, & \text{if } C_t < C_p < C_0 \\ 0, & \text{if } C_p \leq C_t \end{cases} \tag{11}$$

Fig. 8 Presentation of accessibility dimension



The degree of accessibility increases with the increasing number of comments given to a particular post. Figure 8 shows the presentation of accessibility dimension.

5.7 Presentability

The presentability of data specifies how correct and error-free data are, and it is strictly linked to the concept of '*errors*' which can affect the results of the analysis. According to Salvatore et al. (2020), presentability refers to the right interpretation of data that enables users to completely comprehend the information; it describes the level of correctness that assures the data is error-free. For instance, suppose when a manually written text is processed, and some of the characters are not clearly defined. In this regard, it is suggested to use the percentage of misspelled words as a gauge of the presentability of data. Given a post p, with W a collection of correct and non-correct words in the post p, we use Earley parser algorithm (Earley 1970) which receives as input two parameters: a grammar noted G and a list of words representing the post to check correct words.

$$Score = \frac{\text{Number of misspelled words in P}}{\text{Total words in P}} \tag{12}$$

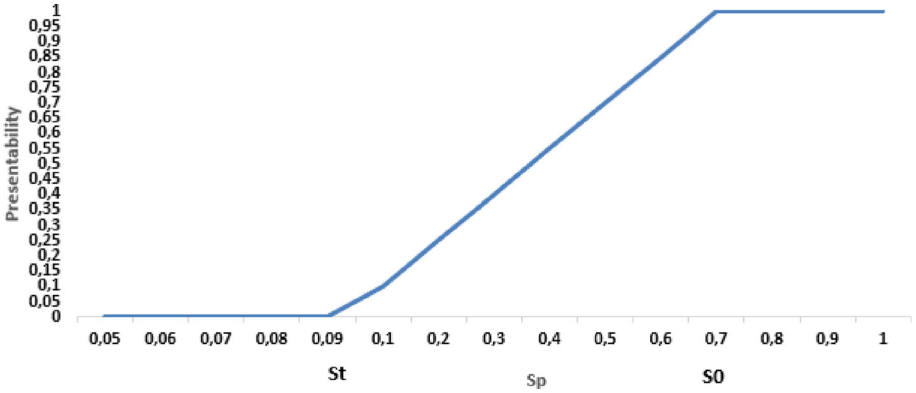


Fig. 9 Presentation of presentability dimension

Let S_p be the score of the post p, S_0 is the score of the post that has the highest score, and S_t is the score tolerable. The presentability of p, denoted Pre(p), can be measured as follows:

$$Pre(p) = \begin{cases} 1, & \text{if } S_p \geq S_0 \\ 1 - \frac{S_0 - S_p}{S_0 - S_t}, & \text{if } S_t < S_p < S_0 \\ 0, & \text{if } S_p \leq S_t \end{cases} \tag{13}$$

The degree of presentability increases with the increasing score given to each post. Figure 9 shows the presentation of presentability dimension.

5.8 Accuracy

Accuracy dimension concerns the degree of validity of data; it refers to the facility and the process of extracting insight from the data. In other words, it accuracy describes the closeness between two values, the correct representation of the real fact that the other aims to represent. According to Wang and Strong (1996), accurate data must be correct, relevant, and come from credible and reputable sources. In our study, we primarily focus on these three metrics to measure accuracy, we have previously evaluated the presentability to ensure that the data correct, relevancy and reliability are also evaluated for credibility and reputation. Based on this, we consider the following dimensions to define the accuracy:

$$Accuracy = Presentability \vee Relevancy \vee Reliability \tag{14}$$

Using the method proposed by Larsen (1980), the final output of the accuracy A(p) is the weighted average of all outputs of the 3 metrics (aggregation) as follows:

$$A(p) = \frac{c_1 Pre(p) + c_2 R(p) + c_3 Rel(p)}{3} \tag{15}$$

where c_1, c_2, c_3 are the coefficients. The degree of accuracy increases with the increasing number of presentability, relevancy, and reliability given to a particular post. Figure 10 shows the presentation of accuracy dimension.

6 Experiments and results

This section serves as a comprehensive exploration of the implementation of our new model, FULMQA, along with the details of the experiments conducted to assess its performance utilizing the metrics previously introduced. Additionally, it encompasses the reporting of the experiment outcomes and engages in an in-depth discussion of the findings. Twitter was selected as the social media platform for this evaluation, representing a pertinent and dynamic environment for

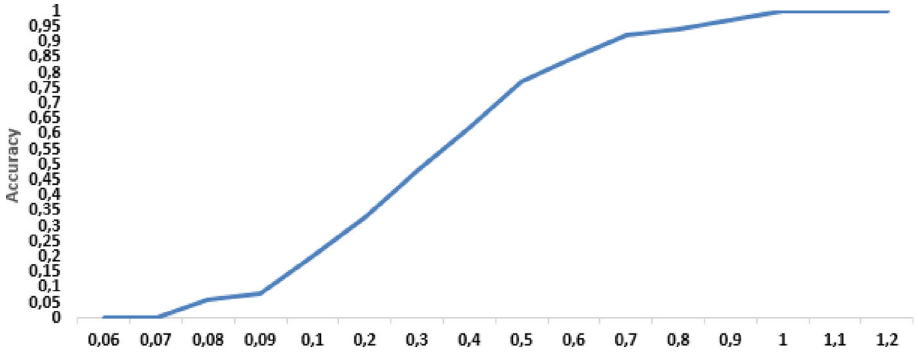


Fig. 10 Presentation of accuracy dimension

Table 3 Features for metrics computation

Features	Description
ID_str	The ID of the tweet
Entities	The hashtags
Created_at	Date creation of the tweet
Followers_count	The number of followers of the user
Favorite_count	The number of likes of the post
Retweet_count	The number of Retweets of the post
Reply_count	The number of comments of the post
Misspelled_count	The number of misspelled words in the post
Sentiment_polarity	The sentiment expressed in the post

evaluating tweet quality. The overarching objective of these experiments is to not only assess tweet quality, but also to provide a demonstrative and illustrative framework for evaluating the quality of tweets on social media platforms. Through these experiments, we aim to shed light on the methodologies and metrics employed to gauge the quality of tweets in a transparent and informative manner.

6.1 Datasets

We utilized the Python Twitter API,[1] known as Tweepy, to access publicly available tweets and extract user-friendly content and user information. Our data collection process involved creating two distinct datasets, each centered around a different hashtag: #coronavirus and #cancer. One dataset comprised approximately 30 tweets. Regarding hashtag selection, the choice of hashtags was indeed a crucial component of our experimental design. These hashtags were selected due to their relevance and significance in contemporary social media discussions. We aimed to evaluate the impact of different hashtags on tweet quality, and this selection was a key variable in our methodology.

It is important to note that Twitter's API imposes certain limitations on our queries. We are confined to a query length of 1000 characters, and we can retrieve historical data spanning roughly 6 to 10 days. Our experiment encompassed all eight metrics. To gain a clearer picture, Table 3 provides a breakdown of the specific features we utilized in the computation of these metrics.

Now, let us delve into the specifics of our experimental approach. Figure 11 visually outlines the step-by-step implementation of our evaluation model.

6.2 Implementation

The step-wise procedure of each metric is presented in the following algorithms. The algorithms 1, 2, 3, 4, 5, 6, 7, and 8 describe our assessment procedure for respectively timeliness, presentability, credibility, popularity, relevancy, reliability, accessibility, and accuracy. These algorithms serve as comprehensive guides to the systematic evaluation of each metric, ensuring transparency and reproducibility in our approach to assessing tweet quality.

6.3 Results and discussions

Once the experiment has been executed, the overall results are depicted in Fig. 12. This presentation encapsulates the overarching results, offering a concise and accessible summary of our research outcomes.

[1] Twitter API. https://developer.twitter.com/en/docs/tweets/search/overview.

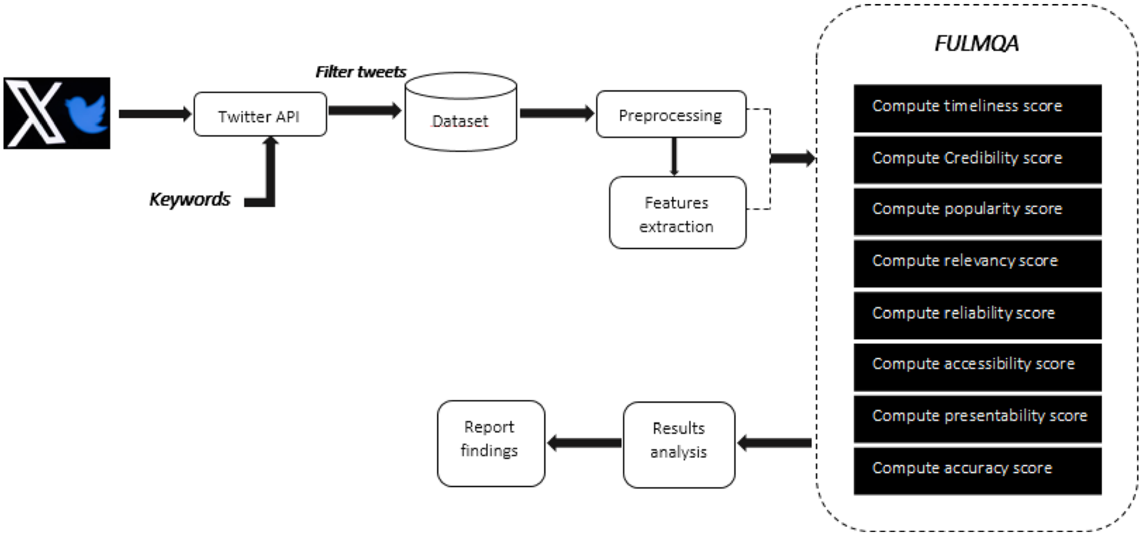


Fig. 11 A comprehensive design of the experiments

Algorithm 1: TIMELINESS ASSESSMENT ALGORITHM

```
Input: Tweet ID
Output: Timeliness score
Initialize the age of the recent post : $Age_0$
Initialize the age tolerable : $Age_t$
Initialize the current timestamp : $T_{now}$
Initialize the creation time of the post : $T_{post}$
for each tweet do
    $Age_{post} \leftarrow T_{now} - T_{post}$
    if $Age_{post} \leq Age_0$ then
        $Timeliness \leftarrow 1$
    else
        if $Age_{post} < Age_t$ AND $Age_{post} > Age_0$ then
            $Timeliness \leftarrow \frac{Age_t - Age_{post}}{Age_t - Age_0}$
        else
            if $Age_{post} \geq Age_T$ then
                $Timeliness \leftarrow 0$
            end if
        end if
    end if
end for
return $Timeliness$
```

Algorithm 2: PRESENTABILITY ASSESSMENT ALGORITHM

```
Input: Tweet ID
Output: Presentability score
Initialize N° of misspelled words : $W_p$
Initialize total N° of words : $Total$
Initialize the score tolerable : $S_t$
Initialize the score of post P : $S_p$
Initialize maximum score : $S_0$
for each tweet do
    $S_p \leftarrow \frac{W_p}{Total}$

    if $S_p \geq S_0$ then
        $Presentability \leftarrow 1$
    else
        if $S_p < S_0$ AND $S_p > S_t$ then
            $Presentability \leftarrow 1 - \frac{S_0 - S_p}{S_0 - S_t}$
        else
            if $S_p \leq S_t$ then
                $Presentability \leftarrow 0$
            end if
        end if
    end if
end for
return $Presentability$
```

To assess the effectiveness of our metrics, we intentionally included tweets from users with varying follower counts, as well as tweets with diverse levels of likes, retweets, and comments across both datasets, ensuring that our evaluation model was subjected to a spectrum of social influence scenarios and encompass the full spectrum of user interactions within the social media ecosystem.

Algorithm 3: CREDIBILITY ASSESSMENT ALGORITHM

```
Input: Tweet ID
Output: Credibility score
Initialize N° of followers tolerable : $F_t$
Initialize N° of followers of post P : $F_p$
Initialize maximum N° of followers : $F_0$
for each tweet do
    if $F_p \geq F_0$ then
        $Credibility \leftarrow 1$
    else
        if $F_p < F_0$ AND $F_p > F_t$ then
            $Credibility \leftarrow 1 - \frac{F_0 - F_p}{F_0 - F_t}$
        else
            if $F_p \leq F_t$ then
                $Credibility \leftarrow 0$

            end if
        end if
    end if
end for
return $Credibility$
```

Algorithm 4: POPULARITY ASSESSMENT ALGORITHM

```
Input: Tweet ID
Output: Popularity score
Initialize N° of retweet tolerable : $S_t$
Initialize N° of retweet of post P : $S_p$
Initialize maximum N° of retweet : $S_0$


for each tweet do
    if $S_p \geq S_0$ then
        $Popularity \leftarrow 1$
    else
        if $S_p < S_0$ AND $S_p > S_t$ then
            $Popularity \leftarrow 1 - \frac{S_0 - S_p}{S_0 - S_t}$
        else
            if $S_p \leq S_t$ then
                $Popularity \leftarrow 0$
            end if
        end if
    end if
end for
return $Popularity$
```

Additionally, it is worth noting that the two datasets were extracted at different points in time, as depicted in Fig. 12h. The timestamps reveal that the data for the cancer dataset was collected from “2022-11-10 17:09” to “2022-11-10 15:23,” whereas the earliest tweet in the coronavirus ataset was posted at “2022-11-10 16:49,” and the latest one at “2022-11-09 17:11.” This temporal contrast highlights an interesting observation: the timeliness of the cancer dataset is marginally superior to that of the coronavirus dataset. It implies that

Algorithm 5: RELEVANCY ASSESSMENT ALGORITHM

```
Input: Tweet ID
Output: Relevancy score
Initialize N° of positive words : $P_w$
Initialize N° of negative words : $N_w$
Initialize total N° of words : $Total$
Initialize the Sentiment tolerable : $S_t$
Initialize the Sentiment of post P : $S_p$
Initialize maximum Sentiment : $S_0$
for each tweet do
    $S_p \leftarrow \frac{P_w - N_w}{Total}$
    if $(S_p > 0$ and $S_p \geq S_0)$ OR $(S_p < 0$and$S_p \leq S_0)$ then
        $Relevancy \leftarrow 1$
    else
        if $S_p > 0$ and $S_p \leq S_0$ AND $S_p \geq S_t$ then
            $Relevancy \leftarrow 1 - \frac{S_0 - S_p}{S_0 - S_t}$
        else
            if $S_p < 0$ and $S_p \geq S_0$ AND $S_p \leq S_t$ then
                $Relevancy \leftarrow \frac{S_t + S_p}{S_t + S_0}$
            else
                if $(S_p > 0$ and $S_p \leq S_t)$ OR $(S_p < 0$ and $S_p \geq S_t)$ then
                    $Relevancy \leftarrow 0$
                end if
            end if
        end if
    end if
end for
return $Relevancy$
```

Algorithm 6: RELIABILITY ASSESSMENT ALGORITHM

```
Input: Tweet ID
Output: Reliability score
Initialize N° of likes of post P : $L_p$
Initialize N° of likes tolerable : $L_t$
Initialize maximum N° of likes : $L_0$
Initialize coefficients : $m_1$ $m_2$ $m_3$
// Calculate : Reputation
for each tweet do
    if $L_p \geq L_0$ then
        $Reputation \leftarrow 1$
    else
        if $L_p < L_0$ AND $L_p > L_t$ then
            $Reputation \leftarrow 1 - \frac{L_0 - L_p}{L_0 - L_t}$
        else
            if $L_p \leq L_t$ then
                $Reputation \leftarrow 0$
            end if
        end if
    end if
    // Calculate : Credibility
    $C \leftarrow$ Credibility(tweet)

    // Calculate : Popularity
    $P \leftarrow$ Popularity(tweet)

    // Calculate : Reliability
    $Reliability \leftarrow \frac{m_1 \times C + m_2 \times P + m_3 \times Reputation}{3}$

end for

return $Reliability$
```

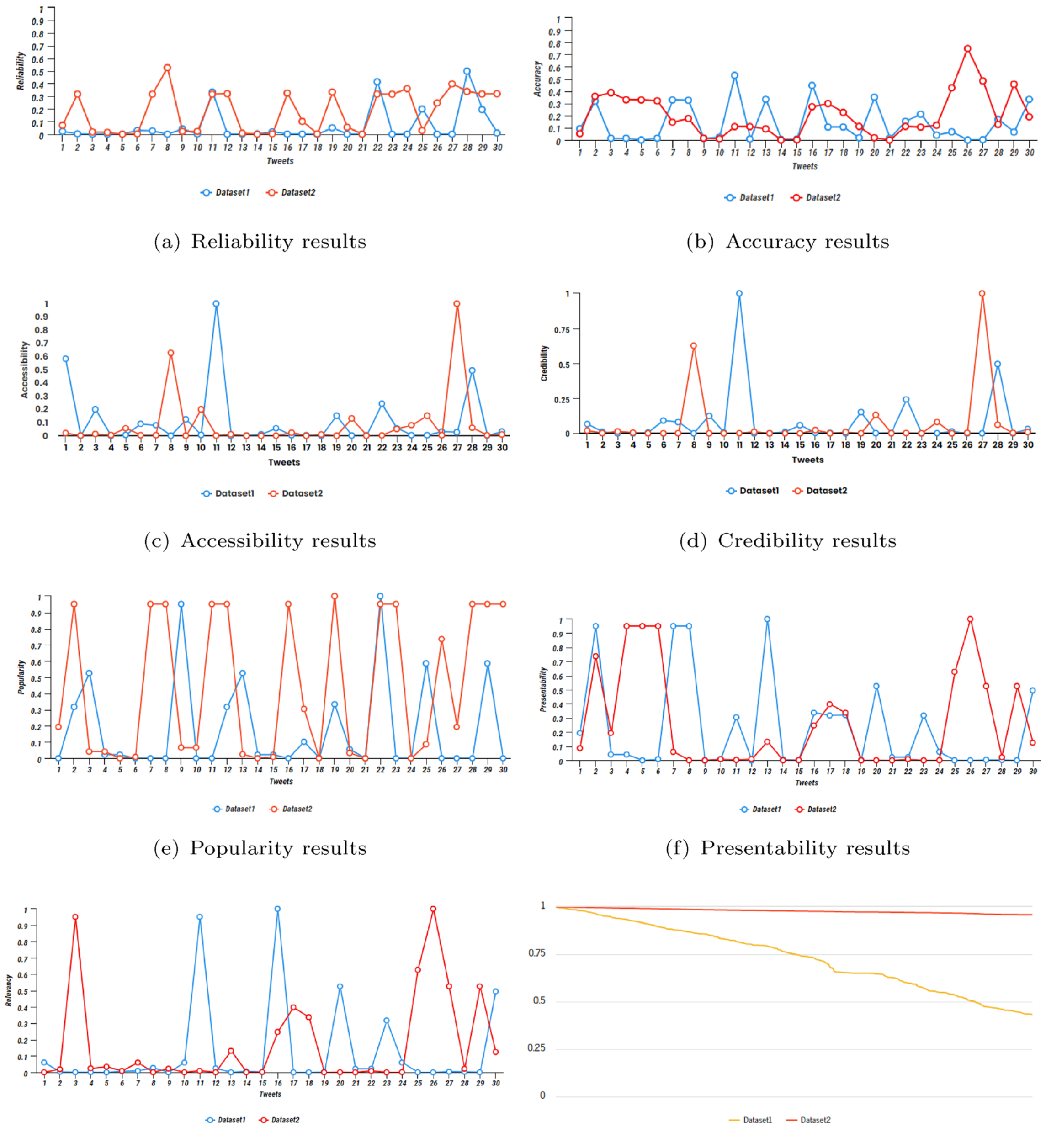
Algorithm 7: ACCESSIBILITY ASSESSMENT ALGORITHM

```
Input: Tweet ID
Output: Accessibility score
Initialize N° of comments of post P : $C_p$
Initialize N° of comments tolerable : $C_t$
Initialize maximum N° of comments : $C_0$
for each tweet do
    if $C_p \geq C_0$ then
        $Accessibility \leftarrow 1$
    else
        if $C_p < C_0$ AND $C_p > C_t$ then
            $Accessibility \leftarrow 1 - \frac{C_0 - C_p}{C_0 - C_t}$
        else
            if $C_p \leq C_t$ then
                $Accessibility \leftarrow 0$
            end if
        end if
    end if
end for
return $Accessibility$
```

Algorithm 8: ACCURACY ASSESSMENT ALGORITHM

```
Input: Tweet ID
Output: Accuracy score
Initialize coefficients : $m_1$ $m_2$ $m_3$

for each tweet do
    // Calculate : Presentability
    $P \leftarrow$ Presentability(tweet)

    // Calculate : Relevancy
    $R \leftarrow$ Relevancy(tweet)

    // Calculate : Reliability
    $Rel \leftarrow$ Reliability(tweet)

    // Calculate : Accuracy
    $Accuracy \leftarrow \frac{m_1 \times P + m_2 \times R + m_3 \times Rel}{3}$
end for
return $Accuracy$
```

tweets within the cancer dataset are more contemporaneous and aligned with the timeline of our data collection efforts. This finding underscores the robustness of our model in effectively evaluating tweets across various temporal contexts. It is a testament to the model's high performance and the quality of results it can produce, reaffirming its efficacy in assessing tweet quality across a spectrum of characteristics.



(a) Reliability results

(b) Accuracy results

(c) Accessibility results

(d) Credibility results

(e) Popularity results

(f) Presentability results

(g) Relevancy results

(h) Timeliness results

Fig. 12 Representation of **a** reliability **b** accuracy **c** accessibility **d** credibility **e** popularity **f** presentability **g** relevancy and **h** timeliness metrics

7 Conclusion and future work

In this paper, we presented the implementation of our new model FULMQA; FUzzy Logic-based Model for Quality Assessment; a quality model that is primarily designed to assess the quality of data derived from social media platforms. Our model includes eight data quality metrics namely: timeliness, credibility, reliability, presentation, popularity, relevancy, accessibility, and accuracy. To obtain a better understanding of how we assess the quality of data from social media, we run an extensive experiments to evaluate the performance of the model using a real implementation on two different datasets from 30 tweets. In the future work, we will propose a holistic model that includes other metrics.

Author Contributions Both authors equally contributed to this work.

Declarations

Competing interests The authors declare no competing interests.

References

Abbasi MA, Liu H (2013) Measuring user credibility in social media. In: International conference on social computing, behavioral-cultural modeling, and prediction. Springer, Berlin, Heidelberg, pp 441–448

Ajarroud O, Zellou A, Idri A (2018) A new filtering-based query processing: improving semantic caching efficiency in mediation systems. In: Proceedings of the 12th international conference on intelligent systems: theories and applications, SITA'18, October 2018, Article no. 12, ACM international conference proceeding series. Rabat, Morocco, pp 1–6

Al-Hajjar D, Jaafar N, Al-Jadaan M, Alnutaifi R (2015) Framework for social media big data quality analysis. New Trends Database Inf Syst II:301–314. https://doi.org/10.1007/978-3-319-10518-5-23

Alizamini FG, Pedram MM, Alishahi M, Badie K (2010) Data quality improvement using fuzzy association rules. In: 2010 International conference on electronics and information engineering, vol 1. IEEE, pp V1–468

Alrubaian M, Al-Qurishi M, Al-Rakhami M, Hassan MM, Alamri A (2017) Reputation-based credibility analysis of twitter social network users. Concurr Comput Pract Exp 29(7):e3873

Alrubaian M, Al-Qurishi M, Alamri A, Al-Rakhami M, Hassan MM, Fortino G (2018) Credibility in online social networks: a survey. IEEE Access 7:2828–2855

Ardagna D, Cappiello C, Samá W, Vitali M (2018) Context-aware data quality assessment for big data. Futur Gener. Comput. Syst. 89:548–562

Arolfo F, Rodriguez KC, Vaisman A (2020) Analyzing the quality of twitter data streams. Inf Syst Front 24:1–21. https://doi.org/10.1007/s10796-020-10072-x

Berlanga R, Lanza-Cruz I, Aramburu MJ (2019) Quality indicators for social business intelligence. In: 2019 6th international conference on social networks analysis, management and security (SNAMS). https://doi.org/10.1109/snams.2019.8931862

Berti-Équille L (1999) Qualité des données multi-sources et recommandation multi-critère. In: Actes du congrès francophone INFormatique des ORganisations et systèmes d'INFormation décisionnels (INFORSID'99), pp 185–204

Bird S (2006) NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL 2006 interactive presentation sessions, pp 69–72

Caballero I, Verbo E, Serrano M, Calero C, Piattini M (2009) Tailoring data quality models using social network preferences. In: International conference on database systems for advanced applications, Springer, Berlin, Heidelberg, pp 152–166

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

Chai K, Potdar V, Dillon T (2009) Content quality assessment related frameworks for social media. Lecture notes in computer science, pp 791–805. https://doi.org/10.1007/978-3-642-02457-3-65

Crosby PB (1979) Quality is free. McGraw-Hill, New York, p 309

Deming WE (1982) Quality, productivity and competitive position. Massachusetts Institute of Technology Center for Advanced Engineering Study, Cambridge, MA, USA

Earley J (1970) An efficient context-free parsing algorithm. Commun ACM 13(2):94–102

Ehrlinger L, Wöß W (2018) A novel data quality metric for minimality. In: International workshop on data quality and trust in big data, Springer, Cham, pp 1–15

El Alaoui I, Gahi Y, Messoussi R (2019) Big data quality metrics for sentiment analysis approaches. In: Proceedings of the 2019 international conference on big data engineering, pp 36–43

Elmasri R, Navathe SB (2000) Fundamentals of database systems, 3rd edn. Addison-Wesley, Reading, MA

Even A, Shankaranarayanan G (2009) Dual assessment of data quality in customer databases. J Data Inf Qual (JDIQ) 1(3):1–29

Fagroud FZ, Ajallouda L, Lahmar EHB, Zellou A, El Filali S (2021) A brief survey on internet of things (IoT). In: 1st International conference on digital technologies and applications. Lecture notes in networks and systems, 211 LNNS, ICDTA, pp 335–344

Firmani D, Mecella M, Scannapieco M, Batini C (2015) On the meaningfulness of big data quality (invited paper). Data Sci Eng 1(1):6–20. https://doi.org/10.1007/s41019-015-0004-7

Gabr MI, Yehia MH, Doaa SE (2021) Data quality dimensions, metrics, and improvement techniques. Futur Comput Inform J 6(1):3

Gupta P, Pathak V, Goyal N, Singh J, Varshney V, Kumar S (2019) Content credibility check on twitter. Commun Comput Inf Sci 899:197–212

Hassenstein MJ, Vanella P (2022) Data quality-concepts and problems. Encyclopedia 2022(2):498–510

Hitzler P, Zaveri A, Rula A, Maurino A, Pietrobon R, Lehmann J, Auer S (2016) Quality assessment for linked data: a survey a systematic literature review and conceptual framework. Semant Web 1:1–5

Hoyle D (2006) ISO 9000 quality systems handbook. Routledge, Oxford http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=42180

Hutto C, Gilbert E (2014) Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI conference on web and social media, vol. 8, pp 216–225

Immonen A, Paakkonen P, Ovaska E (2015) Evaluating the quality of social media data in big data architecture. IEEE Access 3:2028–2043. https://doi.org/10.1109/access.2015.2490723

International Organization for Standardization ISO/IEC 25012:2008(E) (2008) Software engineering-software product quality requirementsand evaluation (SQuaRE)-data quality model. International Organization for Standardization, Geneva, Switzerland

International Organization for Standardization–ISO (1994) Quality management and quality assurance: vocabulary ISO 8402:1994

International Standards Organization (ISO) 8402 (1994) Quality management and quality assurance

Juran JM (2003) Juran on leadership for quality. Simon and Schuster, New York

Laranjeiro N, Soydemir SN, Bernardino J (2015) A survey on data quality: classifying poor data. In: 2015 IEEE 21st Pacific rim international symposium on dependable computing (PRDC), IEEE, pp 179–188

Larousse. Qualité, www.larousse.fr/dictionnaires/francais/qualit%C3%A9/65477

Larsen PM (1980) Industrial application of fuzzy logic control. Int J Man Mach Stud 12:3–10

Lee YW, Pipino LL, Funk JD, Wang RY (2006) Journey to data quality. MIT Press, Cambridge, MA

Müller H, Naumann F, Freytag JC (2003) Data quality in genome databases. Humboldt University of Berlin, Berlin

Nikiforova A (2020) Definition and evaluation of data quality: user-oriented data object-driven approach to data quality assessment. Balt J Modern Comput 8(3):391–432

Olson JE (2003) Data quality: the accuracy dimension. Elsevier, Amsterdam

Ossorio Arroyo A, Onorati T, Diaz P (2018) Quality assessment of social media: lessons learnt from the literature. In: 2018 22nd International conference information visualisation (IV). https://doi.org/10.1109/iv.2018.00055

Pääkkönen P, Jokitulppo J (2017) Quality management architecture for social media data. J Big Data 4(1):1–26. https://doi.org/10.1186/s40537-017-0066-7

Radulovic F, Mihindukulasooriya N, García-Castro R, Gómez-Pérez A (2018) A comprehensive quality model for linked data. Semant Web 9(1):3–24

Reda O, Zellou A (2023) Assessing the quality of social media data: a systematic literature review. Bull Electr Eng Inform 12(2):1115–1126

Reda O, Sassi I, Zellou A, Anter S (2020) Towards a data quality assessment in big data. In: Proceedings of the 13th international conference on intelligent systems: theories and applications. https://doi.org/10.1145/3419604.3419803.

Reda O, Zellou A (2022) SMDQM-social media data quality assessment model. In: 2022 2nd International conference on innovative research in applied science, engineering and technology (IRASET), IEEE, pp 1–7

Reuter C, Ludwig T, Ritzkatis M, Pipek V (2015, May) Social-QAS: tailorable quality assessment service for social media content. In: International symposium on end user development, Springer, Cham, pp 156–170

Ross TJ (2012) Fuzzy logic with engineering applications, 3rd edn. Wiley, New York, p 585

Salvatore C, Biffignandi S, Bianchi A (2020) Social media and twitter data quality for new social indicators. Soc Indic Res. https://doi.org/10.1007/s11205-020-02296-w

Scannapieco M (2006) Data quality: concepts, methodologies and techniques. Data-centric systems and applications. Springer, Cham

Sint R, Schaffert S, Stroka S, Ferstl R (2009) Combining unstructured, fully structured and semi-structured information in semantic wikis. In: 4th Workshop on semantic wikis-the semantic Wiki web 6 the European semantic web conference Hersonissos, Crete, Greece, June 2009, pp 73

Tayi GK, Ballou DP (1998) Examining data quality. Commun ACM 41(2):54–57

Verma PK, Sharma V, Agarwal S (2019) Credibility investigation for tweets and its users. In: Proceedings of the 3rd international conference on computing methodologies and communication, ICCMC 2019, pp 925–928

Wang RY, Strong DM (1996) Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 12:5–33. https://doi.org/10.1080/07421222.1996.1151809

Wang X, Ruan D, Kerre EE (2009) Mathematics of fuzziness-basic issues. Studies in fuzziness and soft computing, vol 245. Springer, Berlin/Heidelberg, p 220

Wayne SR (1983) Quality control circle and company wide quality control. Qual Prog 16(10):14–17

Woodall P, Parlikad AK (2010) A hybrid approach to assessing data quality. In ICIQ

Yang J, Yu M, Qin H, Lu M, Yang C (2019) A twitter data credibility framework-hurricane Harvey as a use case. ISPRS Int J Geo Inf 8(3):111

Yousfi A, El Yazidi MH, Zellou A (2018) Assessing the performance of a new semantic similarity measure designed for schema matching for mediation systems. In: Nguyen N, Pimenidis E, Khan Z, Trawinski B (eds) Computational collective intelligence: 10th international conference on computational collective intelligence. ICCCI'18, Bristol, UK, September 5-7, Proceeding, part I, vol 11055. Springer, pp. 64–74. Print_ISBN: 978-3-319-98442-1. Online_ISBN: 978-3-319-98443-8

Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.