Preliminary Framework for Verifying the Syntactic Quality of Digital Voice-of-Customer for Product Quality Monitoring

Book of Proceedings

of the 7th edition of the International Conference on Quality Engineering and Management

ICQEM 2026

TECHNICAL RECORD

Title

Proceedings book of the 7th International Conference on Quality Engineering and Management, 2026

Authors/Editors

Carvalho, André; Sampaio, Paulo; Domingues, Pedro; Salvadorinho, Juliana; Morgado, Teresa; Casadesús, Martí; Marimon, Frederic; Pires, António Ramos; Saraiva, Pedro

Publisher

International Conference on Quality Engineering and Management

Date

July 2026

Cover Design

ICQEM

ISBN

978-989-54911-3-1

ISSN

2184-3481

1st Edition

200 units

7th International Conference on Quality Engineering and Management

2-3 July 2026, Universidade NOVA de Lisboa, Lisbon, Portugal

Preliminary Framework for Verifying the Syntactic Quality of Digital Voice-of-Customer for Product Quality Monitoring

Barravecchia Federico[1], Mastrogiacomo Luca[2] and Franceschini Fiorenzo[3]

[1] Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino 10129, Italy; federico.barravecchia@polito.it
[2] Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino 10129, Italy; luca.mastrogiacomo@polito.it
[3] Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino 10129, Italy; fiorenzo.franceschini@polito.it

***Abstract*—Voice-of-Customer (VoC) data, such as online customer reviews and unsolicited customer feedback, has become a valuable complement to traditional quality information in manufacturing contexts and within the broader paradigm of Quality 4.0. In particular, digital VoC provides post-market evidence on product quality, supporting the monitoring of field performance and the early detection of customer-perceived defects. However, the reliability of digital VoC-based monitoring depends critically on the quality of the underlying data. When digital VoC datasets are noisy, incomplete, inconsistent, or manipulated, even advanced analytics may generate misleading signals, reflecting the well-known "garbage in, garbage out" effect. This paper addresses the problem of digital VoC data reliability for product quality monitoring by grounding the discussion in established data-quality principles, with particular reference to ISO 8000. The standard distinguishes three complementary perspectives on data quality: syntactic, semantic, and pragmatic, which together determine whether data can support reliable analysis and decision-making. Building on this conceptual structure, the paper develops a preliminary framework for assessing the syntactic quality of digital VoC datasets. Syntactic quality concerns the formal correctness of data, independently of their meaning or usefulness. The proposed framework operationalizes syntactic quality through a set of subdimensions and associated metrics, providing a foundation for assessing the reliability of VoC datasets before performing downstream analytics in manufacturing quality monitoring. While the broader problem of VoC reliability includes semantic and pragmatic considerations, this work focuses specifically on the syntactic dimension as a necessary first step toward a comprehensive reliability assessment framework for digital VoC.**

Keywords – Digital Voice of Customer; Data Quality; Quality 4.0; Customer Reviews; ISO 8000; Reliability; Product Quality Monitoring.

I. Introduction

Digital Voice-of-Customer (VoC) has established itself as a rich and timely source of customer feedback (Özdağoğlu et al., 2018; Barravecchia et al., 2023). Digital VoC is produced continuously and without prompting, providing a stream of information about real-world usage, perceived performance, and emerging dissatisfaction on products and services. In manufacturing industries, these data are particularly valuable because they provide post-market evidence on how products perform in real usage conditions, complementing internal quality indicators such as non-conformities, warranty claims, service reports, and in-process quality data. In Quality 4.0, commonly understood as the digitalization of quality management through connectivity, advanced analytics, and new data sources (Zonnenshain and Kenett, 2020; Broday, 2022), digital VoC can act as a complementary "external sensor" that enriches internal quality signals (Barravecchia et al., 2022b).

Quality management research has explored how digital VoC can support monitoring and tracking quality (Barravecchia et al., 2023; Barravecchia et al., 2022a). In particular, online reviews and user-generated feedback have been used as a complementary source of information to detect early signals of quality evolution and customer-perceived issues. Because these data reflect real-world product usage, they can provide indications of emerging problems that may not yet appear in traditional internal quality metrics(Barravecchia et al., 2023). In the manufacturing domain, such analyses are especially relevant when digital VoC is used to identify recurring defect-related complaints, detect early signs of product quality deterioration, and support the prioritization of engineering investigations. Several studies have proposed approaches for quality tracking based on the analysis of review texts, using techniques such as sentiment analysis (Kim, 2021), topic modelling (Yan et al., 2025), and large language model-based methods (Park et al., 2026) to identify recurring defect-related discussions or dissatisfaction patterns. Research have also combined star ratings and textual content to detect shifts in customer perception over time (Barravecchia et al., 2026), applying time-series analysis (See-To and Ngai, 2018) or control-chart approaches (Kakde et al., 2024) to identify anomalies and potential quality degradation. These contributions illustrate that digital VoC analytics can help identify quality-relevant patterns. At the same time, the practical value of VoC is already evident in many organizational contexts, where actionable insights are obtained from reviews, and in many cases digital VoC can reveal issues earlier than conventional post-market surveillance channels (Kim and Park, 2017). The goal of this paper is to show that digital VoC-based quality monitoring is stronger and more defensible when it incorporates an explicit layer of input-data reliability assessment.

In practice digital VoC approaches frequently emphasize model development and insight extraction more than the evaluation of dataset quality. In domains like online reviews, the need is particularly critical because there is limited quality control over who can contribute and what they can write; this creates conditions for low-quality content and deliberate manipulation.

In this context, the core premise is that monitoring results are only as reliable as the data used to generate them. This is a general principle in information systems and decision-making: if data is incomplete, inconsistent, or unfit for purpose, decisions degrade correspondingly (Shankaranarayanan and Cai, 2006). In digital VoC, the risk of "garbage in, garbage out" is amplified by unstructured language, off-topic content, selection effects, and deception. In product quality monitoring, this issue is particularly critical because unreliable digital VoC data may distort the identification of defect patterns, mask emerging field failures, or trigger unnecessary alarms, thereby affecting downstream quality decisions.

This paper addresses the general problem of VoC data reliability for quality monitoring by grounding the discussion in established data-quality principles and, in particular, in the ISO 8000 standard (International Organization for Standardization, 2011) that conceptualizes data quality through three complementary perspectives: syntactic, semantic, and pragmatic, each capturing a different aspect of how data support analysis and decision-making. Within this broader perspective, the present study focuses specifically on syntactic quality, which concerns the formal correctness of data. The paper develops a preliminary framework that operationalizes syntactic quality for digital VoC datasets through a set of subdimensions and associated metrics.

Against this background, the study addresses the following research questions:

- RQ1: How can the syntactic quality of digital VoC data for product quality monitoring be conceptually defined and structured?
- RQ2: How can syntactic be operationalized through measurable subdimensions, metrics, and a structured verification-and-improvement protocol for digital VoC datasets?

The remainder of the paper is structured as follows. Section II discusses why Digital VoC reliability constitutes a distinct challenge in quality monitoring contexts. Section III reviews established data-quality principles. Section IV introduces ISO 8000 as the conceptual foundation of the proposed framework. Section V operationalizes the syntactic quality dimension for Digital VoC datasets. Section VI introduces a structured approach for improving syntactic quality through data cleaning activities. Section VII presents an application case to illustrate the practical implementation of the proposed framework. Finally, Section VIII concludes the paper and outlines directions for future research.

II. Why VoC Reliability Is a Distinct Problem in Quality Monitoring

Digital VoC data differs from many traditional quality datasets because it is user-generated, largely unstructured, and not governed by the organization that uses it. This creates different structural reliability challenges.

First, review texts can be extremely brief, off-topic, or highly subjective, and real-world review corpora often contain a substantial amount of spam content, or unhelpful information (Mewada and Dewang, 2023). A second challenge is a systematic bias in who writes reviews and when. Research has highlighted that online ratings often follow a J-shaped distribution, reflecting self-selection and incentives that are not random samples of the customer population (Ullah et al., 2016). In monitoring terms, this implies that the dataset may exaggerate extreme experiences, complicating the interpretation of trends as "true" quality shifts (Subroto and Christianis, 2021). A third challenge is manipulation. Opinion spam and the trustworthiness of online opinions have been identified as critical issues in review analysis (Hu et al., 2012). Early large-scale studies of reviews argued that trustworthiness was a neglected problem in review mining and showed how untruthful opinions, brand-only reviews, and non-reviews can distort the dataset (Esposito et al., 2021). Finally, even when reviews are genuine, the practical usefulness of text is influenced by clarity and interpretability. Writing quality, readability, and error-related signals have been treated as important factors that shape perceived helpfulness and, by extension, the practical value of reviews as information (Santhiran et al., 2024).

These issues do not reduce the usefulness of digital VoC. Instead, they highlight the importance of carefully assessing the reliability of the data. If digital VoC is used as an input to quality monitoring, especially for decisions that might influence engineering investigations, design changes, supplier actions, or customer communication, then the dataset should be treated like an "information product" whose quality must be defined and validated for use.

In manufacturing contexts, these limitations are not merely analytical inconveniences. When digital VoC is incorporated into product quality monitoring systems, syntactic and structural defects in the data can affect defect detection, distort product-level comparisons, and weaken the evidential basis for decisions concerning root-cause analysis, supplier assessment, redesign priorities, or corrective actions. For this reason, Digital VoC data should be treated as a quality-relevant information asset whose reliability must be verified before being used in product quality evaluation.

III. Recognized Data-Quality Principles

A key reason why digital VoC reliability is often hard to discuss is the lack of shared vocabulary. Data-quality research has long established that quality is multidimensional. A widely cited perspective is that data quality is not only accuracy: it includes intrinsic dimensions (e.g., accuracy, believability), contextual dimensions (e.g., relevance, timeliness), representational dimensions (e.g., interpretability, consistency of representation), and accessibility-related dimensions (Wand and Wang, 1996). The "Beyond Accuracy" framework articulates these categories and shows that data consumers evaluate quality across multiple criteria, not a single one (Wang and Strong, 1996).

Several surveys emphasize the breadth of dimensions and the need to interpret them relative to context, particularly for large-scale and heterogeneous datasets (Fox et al., 1994; Ridzuan and Zainon, 2024; Wang et al., 2024).

Within this broader landscape, ISO 8000 (International Organization for Standardization, 2011) provides a relevant foundation for a Quality 4.0 setting because it makes data quality conceptually compatible with quality management thinking. ISO vocabulary defines data quality as the degree to which inherent characteristics of data fulfil requirements (International Organization for Standardization, 2015).

IV. ISO 8000 AS CONCEPTUAL FOUNDATION FOR MEASURING DIGITAL VOC DATA QUALITY

According to ISO 8000-8 (International Organization for Standardization, 2015), information and data quality can be understood through three complementary perspectives, each addressing a different aspect of how data support decision-making and analysis (see Figure 1):

- Syntactic quality refers to the formal correctness of data: it concerns whether data conform to predefined schemas, formats, and integrity rules, ensuring that records are structurally valid and processable by information systems, independently of their meaning or usefulness.
- Semantic quality concerns whether data correspond correctly, unambiguously, and meaningfully to the real-world domain they are intended to represent. In ISO 8000-8, semantic quality is expressed in terms of a correct and complete mapping between identifiable data units and the entities or concepts of the domain of interest.
- Pragmatic quality concerns whether data are suitable for a specific purpose or application. It focuses on whether the data, even if formally correct and semantically meaningful, are actually useful for supporting the decisions or activities they are intended to inform.

This triad is particularly useful for digital VoC because it clarifies that "being well-formed" is not equivalent to "being trustworthy," and "being correct" is not equivalent to "being fit for monitoring decisions". This ISO structure suggests a structured approach: digital VoC quality should be evaluated at all three levels, because failures at any level can undermine monitoring conclusions.

ISO 8000 also distinguishes verification and validation (International Organization for Standardization, 2015): some aspects of data quality are checked against specified requirements (verification), while others are judged against intended use (validation). This distinction becomes central when digital VoC data is syntactically well-formed but still unreliable for monitoring.

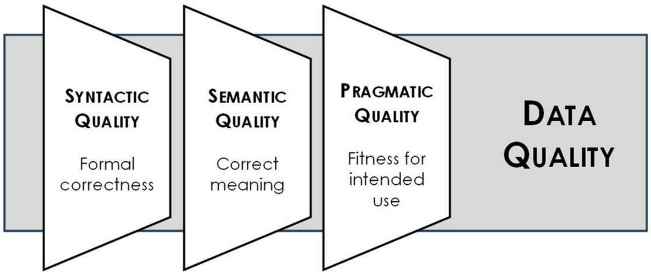

Figure 1. Conceptual representation of data quality according to ISO 8000-8 (International Organization for Standardization, 2015)

V. OPERATIONALIZING SYNTACTIC QUALITY OF DIGITAL VOC DATA

As seen in the previous section, ISO 8000-8 conceptualizes data quality through three complementary dimensions (International Organization for Standardization, 2015). While all three dimensions are relevant for assessing the reliability of Digital VoC data, the focus of this article is exclusively on the syntactic dimension. According to ISO 8000-8, syntactic quality refers to the degree to which data conform to formally specified metadata (International Organization for Standardization, 2015). In this study, syntactic quality is operationalized by translating the ISO 8000 notion of conformance to formally specified metadata into classes of verifiable constraints for Digital VoC datasets. Specifically, the six subdimensions were derived by identifying the main types of metadata requirements that can be formally specified and checked before analysis: data-type requirements, mandatory-field requirements, domain and format requirements, identifier uniqueness requirements, inter-attribute integrity requirements, and text representation requirements. These requirement classes were then mapped respectively onto six syntactic subdimensions: structural validity, completeness, format validity, entity integrity, cross-field consistency, and text processability. In particular, they were derived by grouping the main categories of formally verifiable defects discussed in the literature, such as type mismatches, missing mandatory values, domain and format violations, duplicate records, inter-field inconsistencies, and text representation problems (Wang and Strong, 1996). Consistent with this perspective, these dimensions were defined to capture typical situations of syntactic non-conformance, such as incorrect data types, missing mandatory values, domain or format violations, duplicate records, inconsistencies between related attributes, and problems in text representation. Similar categories of data defects are widely discussed in data cleaning and data quality assessment research (Rahm and Do, 2000). Each subdimension therefore represents a specific category of syntactic defect.

Table 1 provides an overview of the syntactic subdimensions and associated metrics proposed in this study. While these subdimensions are informed by established data-quality literature and the ISO 8000 conceptualization of syntactic quality, their specific definition and organization represent an original contribution of this work.

Table 1. Syntactic quality subdimensions and releted metrics for digital VoC data. All indicators are normalized in the interval [0,1], where 1 represents full compliance.

Syntactic Subdimension	ISO 8000 Metadata Requirement	What It Verifies	Metric
STRUCTURAL VALIDITY	Data type	Conformance of attribute values to declared data types	*Type Conformance Rate (TCR)*
COMPLETENESS	Mandatory attribute	Presence of mandatory attributes defined in metadata	*Mandatory Attribute Completeness Ratio (MACR)*
FORMAT VALIDITY	Domain / representation	Compliance with declared domains and representation formats	*Domain Compliance Rate (DCR)*
ENTITY INTEGRITY	Uniqueness	Uniqueness of records according to declared identifiers	*Uniqueness Rate (UR)*
CROSS-FIELD CONSISTENCY	Integrity	Logical consistency between related attributes	*Compliance Score (CS)*
TEXT PROCESSABILITY	Representation	Compliance of textual data with encoding and normalization standards	*Text Representation Compliance Rate (TRCR)*

A. Structural Validity

Structural validity concerns adherence to declared data types. ISO 8000-8 explicitly anchors syntactic quality in conformance to requirements stated in metadata. Data type inconsistencies are recognized as a primary source of structural defects, particularly in multi-source environments (Rahm & Do, 2000).

In Digital VoC contexts, structural validity is critical because review datasets are typically aggregated from heterogeneous platforms. For example, one platform may encode ratings as integers (1–5), while other stores ratings as textual strings (e.g., “5 stars”). If data type harmonization is incomplete, data analytics may produce failures. Such defects represent violations of formally declared structure and therefore fall within syntactic quality.

To illustrate violations of structural validity, Table 2 reports examples of records that do not conform to the declared data types. These records contain type mismatches across multiple attributes. The Rating field includes non-integer values, the Timestamp field contains a non-standard textual expression, and Helpful_Votes is expressed as a string. Such violations prevent correct parsing and compromise the structural integrity of the dataset.

Structural validity can be operationalized through a Type Conformance Rate (TCR), defined as the proportion of records that conform to the declared data types. Let N be the total number of records and N_s the number of records satisfying all constraints. The metric can be defined as:

$$TCR = \frac{N_s}{N} \qquad \mathrm{TCR} \in [0,1] \tag{1}$$

A record is considered-compliant if all attributes respect the declared data types.

Table 2. Example of structural validity violations. The table reports review identifier (Review_ID), unstructured textual content (Review_Text), product identifier (Product_ID), customer rating (Rating), timestamp (Timestamp), and helpfulness indicators (Helpful_Votes, Total_Votes).

Review_ID	Review_Text	Product_ID	Rating	Timestamp	Helpful_Votes	Total_Votes
R1	I’ve been using this device for a couple of weeks and overall it works quite well, especially considering the price.	P100	"five"	2026-01-10	3	5
R2	The product arrived on time and works as expected, although the battery life could be slightly better.	P101	4	"yesterday"	2	3
R3	The performance is decent, but I experienced occasional lag when running multiple apps at the same time.	P102	3.5	2026-02-01	"ten"	12

B. Completeness

Completeness refers to the presence of mandatory attributes as declared in metadata. In data-quality assessment literature, completeness is a core measurable dimension and is commonly operationalized through missing-value ratios (Pipino et al., 2002). ISO/IEC 25012 similarly identifies completeness as an inherent data-quality characteristic (Guerra-García et al., 2023).

Within this framework, completeness is syntactic when it is defined relative to formally declared requirements. If a digital VoC dataset specifies that each record must include a timestamp, rating, and product identifier, then missing values constitute non-conformance.

Table 3 illustrates typical completeness violations, where mandatory attributes defined in the metadata are missing. In these cases, mandatory fields such as Timestamp, Product_ID, and Rating are missing. Since these attributes are required by the metadata specification, their absence constitutes syntactic non-conformance and reduces the usability of the dataset for monitoring purposes.

Completeness can be measured through a Mandatory Attribute Completeness Ratio (MACR). Let M be the set of attributes declared mandatory in the metadata specification. For each attribute $m \in M$, completeness can be computed as:

$$C_m = \frac{\text{Number of non-null values for } m}{N} \tag{2}$$

An overall completeness index can then be defined as:

$$MACR = \frac{1}{\text{M}} \sum_{m \in M} C_m \qquad MACR \in [0,1] \tag{3}$$

Table 3. Example of completeness violations. The table reports review identifier (Review_ID), unstructured textual content (Review_Text), product identifier (Product_ID), customer rating (Rating), timestamp (Timestamp), and helpfulness indicators (Helpful_Votes, Total_Votes).

Review_ID	Review_Text	Product_ID	Rating	Timestamp	Helpful_Votes	Total_Votes
R4	Excellent product, very easy to use and the build quality feels solid. I would definitely recommend it.	P200	5	NULL	2	3
R5	Works well overall, but I had some issues with the initial setup which took longer than expected	NULL	4	2026-01-12	1	2
R6	The product is okay for basic use, but it does not meet expectations for more advanced tasks.	P202	NULL	2026-01-15	0	1

C. Format Validity

Format validity evaluates whether individual attribute values respect declared domains. Format validity concerns domain constraints and standardized representations. Data-cleaning research identifies domain violations as recurring classes of dirty data (Rahm & Do, 2000).

In Digital VoC datasets, format violations often arise in timestamp and categorical fields. A crawler may collect dates in ambiguous forms such as “03/04/2026” without day/month specification. If interpreted inconsistently in processing, such ambiguity can distort monitoring. Similarly, ratings outside the permitted domain (e.g., “6” in a five-point scale) represent domain violations. These defects are verifiable against declared format and domain rules and thus belong to syntactic quality. Table 4 provides examples of violations of domain and format constraints in digital VoC data.

These records violate predefined domains and representation formats. The Rating value exceeds the allowed range, and Timestamp values are expressed using inconsistent formats instead of the required standard. Such inconsistencies can lead to errors in aggregation and temporal analysis.

Format validity can be assessed using a Domain Compliance Rate (DCR), representing the proportion of attribute values that fall within the permitted domain and match declared formats. For a given attribute m with defined domain constraints:

$$DCR_m = \frac{\text{Number of values conforming to domain/format}}{\text{Total values of } m} \qquad DCR_a \in [0,1] \tag{4}$$

An aggregate format validity score can be computed across all constrained attributes:

$$DCR = \frac{1}{K} \sum_{k=1}^{K} DCR_{m_k} \qquad \text{DCR} \in [0,1] \tag{5}$$

where K is the number of attributes with domain constraints.

Table 4. Example of format validity violations. The table reports review identifier (Review_ID), unstructured textual content (Review_Text), product identifier (Product_ID), customer rating (Rating), timestamp (Timestamp), and helpfulness indicators (Helpful_Votes, Total_Votes).

Review_ID	Review_Text	Product_ID	Rating	Timestamp	Helpful_Votes	Total_Votes
R7	The product exceeded my expectations in almost every aspect, especially performance and ease of use.	P300	6	2026-01-10	2	3
R8	Good overall experience, but the packaging was slightly damaged when it arrived.	P301	4	10/01/2026	1	2
R9	Not very satisfied with this purchase, as the device tends to overheat after extended use.	P302	2	Jan-10-2026	0	1

D. Entity Integrity

Entity integrity addresses uniqueness constraints and the absence of duplicate records. ISO 8000-8 explicitly recognizes that violations of entity integrity rules affect syntactic quality when uniqueness is formally specified (International Organization for Standardization, 2015). In data-quality research, duplicate detection and record linkage are established problem domains. The work of Fellegi and Sunter (1969) formalized record linkage as the identification of records representing the same real-world entity, while Elmagarmid et al. (2007) survey duplicate detection as a central data-quality challenge.

In Digital VoC environments, duplicates frequently arise from repeated crawling or syndicated reviews. If identical review texts are stored multiple times, sentiment frequencies and topic distributions may be artificially inflated. Such distortions can trigger false monitoring alarms. Since uniqueness can be formally specified (e.g., one record per review identifier), duplicate presence constitutes a syntactic integrity violation.

To illustrate violations of entity integrity, Table 5 shows duplicate records that violate uniqueness constraints. The two records share the same Review_ID, violating the uniqueness constraint defined in the metadata. Such duplicates may arise from repeated data collection and can distort statistical analyses by artificially inflating the presence of specific reviews.

Entity integrity can be quantified through a Uniqueness Rate (UR). Let N_d be the number of detected duplicate records. Then:

$$UR = 1 - \frac{N_d}{N} \qquad \mathrm{DR} \in [0,1] \tag{6}$$

Table 5. Example of entity integrity violations. The table reports review identifier (Review_ID), unstructured textual content (Review_Text), product identifier (Product_ID), customer rating (Rating), timestamp (Timestamp), and helpfulness indicators (Helpful_Votes, Total_Votes).

Review_ID	Review_Text	Product_ID	Rating	Timestamp	Helpful_Votes	Total_Votes
R10	This is one of the best products I've purchased recently, very reliable and easy to configure.	P400	5	2026-01-10	2	3
R10	This is one of the best products I've purchased recently, very reliable and easy to configure.	P400	5	2026-01-10	2	3

E. Cross-field Consistency

Cross-field consistency concerns referential and inter-attribute integrity constraints. Consistency is recognized as a measurable data-quality dimension in established assessment frameworks (Pipino et al., 2002), and integrity constraint enforcement is a foundational principle in database systems and data cleaning (Rahm and Do, 2000). ISO/IEC 25012 also identifies consistency as an inherent characteristic of data quality (Guerra-García et al., 2023).

In VoC datasets, cross-field inconsistencies occur when attribute combinations violate declared logical rules. For instance, a record may report more helpful votes than total votes, or it may indicate a specific platform while containing attributes exclusive to another platform. Table 6 reports examples of violations of cross-field integrity rules defined between related attributes. These records violate a temporal consistency constraint between related attributes. The Review_Date precedes the Purchase_Date, which is logically inconsistent under the assumption that a review can only be written after a purchase.

Cross-field consistency can be measured using an Integrity Violation Rate (IVR). Let R be the set of declared cross-attribute or referential integrity rules (e.g., helpful_votes ≤ total_votes). For each rule r, define:

$$IVR_r = \frac{\text{Number of records violating } r}{N} \qquad IVR_r \in [0,1] \tag{7}$$

An overall Consistency Score (CS) can be defined as:

$$CS = 1 - \frac{1}{\mathrm{R}} \sum_{r \in R} IVR_r \qquad CS \in [0,1] \tag{8}$$

Table 6. Example of cross-field consistency violations. The table reports review identifier (Review_ID), unstructured textual content (Review_Text), product identifier (Product_ID), review date (Review_Date), purchase date (Purchase_Date), and customer rating (Rating).

Review_ID	Review_Text	Product_ID	Review_Date	Purchase_Date	Rating
R11	I am very satisfied with the product, but it is strange that the review date appears before the purchase date.	P500	2026-01-10	2026-01-15	5
R12	The product works fine overall, but the recorded dates seem inconsistent with the actual purchase.	P501	2026-01-05	2026-01-20	4

F. Text Processability

Digital VoC differs from many traditional datasets due to the presence of unstructured textual content. Nevertheless, text fields remain subject to formal representation constraints. ISO 8000-8 frames syntactic quality in terms of conformance to formal expression rules (International Organization for Standardization, 2015). At the representation level, standards such as RFC 3629 (UTF-8) and the Unicode Standard define encoding and normalization rules that ensure consistent machine-readable representation.

In practice, encoding inconsistencies can distort analytics. A review containing the word "caffè" may appear in different Unicode forms or may be mis-decoded (e.g., "caffÃ¨"). Without normalization, tokenization and dictionary matching treat these variants as distinct terms, fragmenting frequency counts and affecting performance of text analytics.

Table 7 illustrates violations related to text encoding and normalization requirements. These records exhibit encoding inconsistencies, escaped Unicode sequences, and non-standard symbolic content. Such issues hinder text normalization and tokenization processes, affecting the reliability of downstream text analytics.

Text processability can be assessed through a Text Representation Compliance Rate (TRCR). This metric evaluates the proportion of textual records conforming to declared encoding and normalization standards. Let N_t be the number of records containing text, and N_p the number of text records successfully validated, then:

$$TRCR = \frac{N_p}{N_t} \qquad TRCR \in [0,1] \tag{9}$$

Table 7.Example of text processability violations. The table reports review identifier (Review_ID), unstructured textual content (Review_Text), product identifier (Product_ID), customer rating (Rating), timestamp (Timestamp), and helpfulness indicators (Helpful_Votes, Total_Votes).

Review_ID	Review_Text	Product_ID	Rating	Timestamp	Helpful_Votes	Total_Votes
R14	The batterÃ-a life is excellent and the device performs consistently even under heavy usage.	P600	5	2026-01-10	2	3
R15	The product is \u00E8asy to use and the interface is intuitive, although some features could be improved.	P601	4	2026-01-11	1	2
R16	👍 👍 👍 Works fine for basic tasks, but lacks some advanced functionalities expected at this price point.	P602	3	2026-01-12	0	1

VI. Data Cleaning for Improving Syntactic Quality in Digital VoC

In practical Digital VoC analytics, data cleaning is almost always performed. Records with obvious defects are removed, duplicates are filtered, missing values are handled, and textual preprocessing steps such as lowercasing or token normalization are applied. However, these activities are typically conducted in an ad hoc manner, rather than within a formally declared verification framework. Cleaning decisions are often motivated by practical convenience rather than by explicit conformance to declared metadata and integrity constraints.

This lack of structure creates two problems. First, cleaning procedures become difficult to reproduce and audit. Second, there is no explicit link between cleaning actions and formally defined data-quality requirements. As a result, sophisticated analytics can be applied to datasets whose syntactic reliability has never been systematically verified.

Within the ISO 8000 perspective adopted in this paper, syntactic quality is not an emergent property of preprocessing but the outcome of verification against formally specified rules. Data cleaning, therefore, should not be viewed as a heuristic preprocessing step, but as a structured verification-and-correction process aimed at restoring conformance to declared metadata and integrity constraints.

To operationalize syntactic quality improvement in Digital VoC datasets, this study proposes a structured four-stage protocol (see Figure 2).

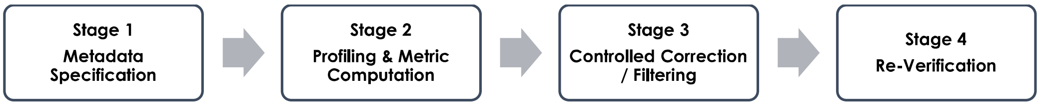

Figure 2. Proposed syntactic quality improvement protocol

A. Stage 1 – Metadata Specification

The protocol begins with the formal specification of metadata. Before any inspection or correction can be performed, the dataset must be associated with a declared structural definition. This includes the schema (attributes and data types), identification of mandatory fields, specification of permitted domains and format standards, explicit uniqueness constraints, cross-field integrity rules, and declared text representation requirements. Without such specification, it is not possible to distinguish between compliant and non-compliant records, and cleaning decisions inevitably become arbitrary.

To make Stage 1 operational and reproducible, Table 2 specifies the metadata components that must be formally declared.

Table 8. Stage 1: Metadata Specification Requirements for Syntactic Verification.

Metadata Component	What Must Be Defined
Data type Definition	List of attributes and associated data types (e.g., integer, string, datetime, boolean)
Mandatory Attributes	Identification of fields required for syntactic completeness (e.g., timestamp, rating, product ID)
Permitted Domains	Allowed value ranges or enumerations for constrained attributes (e.g., rating ∈ {1..5})
Format Standards	Required representation formats (e.g., ISO 8601 for timestamps, standardized language codes)
Uniqueness Constraints	Definition of primary key(s) or uniqueness rule (e.g., one record per reviewer ID)
Cross-field Integrity Rules	Logical consistency rules between attributes (e.g., helpful_votes ≤ total_votes)
Text Representation Standards	Encoding standard and normalization form for textual content (e.g., UTF-8, NFC normalization)

B. Stage 2 – Profiling and Violation Detection

Once the specification is established, the dataset is subjected to systematic profiling in order to detect violations. At this stage, the syntactic metrics introduced in Section V are computed. This step transforms qualitative concerns about "dirty data" into quantifiable metrics of structural non-conformance. Rather than directly modifying the dataset, the objective is first to measure the magnitude and distribution of defects across different constraint families. Profiling thus provides a diagnostic baseline that guides subsequent corrective actions.

C. Stage 3 – Controlled Correction or Filtering

Corrective interventions are subsequently carried out in a controlled manner. At this stage, each detected violation is addressed through actions that are explicitly linked to a declared syntactic constraint. Cleaning is therefore not guided by heuristic judgment or convenience, but by formal verification requirements established during metadata specification. The objective is to restore conformance to declared structural rules.

Corrective actions may include:

- Type harmonization, when structural validity is violated. This may involve aligning heterogeneous attribute names across sources, enforcing declared data types (e.g., converting numeric strings into integers when unambiguous), and removing records that cannot be parsed according to the declared schema.
- Removal or formally justified imputation of incomplete records, when mandatory attribute constraints are not satisfied. In contexts where a missing mandatory field cannot be deterministically reconstructed, records should be excluded rather than arbitrarily completed.
- Normalization of domain values and format standardization, in cases of format validity violations. This includes enforcing declared value ranges (e.g., rating bounds), converting dates to a standardized representation, and correcting enumerations when a one-to-one mapping to the declared domain is clearly defined.
- Duplicate detection and elimination according to declared uniqueness constraints, when entity integrity is compromised. Removal decisions should be based on explicit matching criteria (e.g., identical reviewer identifier).
- Enforcement of cross-field integrity constraints, when inter-attribute logical inconsistencies are detected. Records violating declared rules (e.g., helpful_votes exceeding total_votes) should either be corrected or excluded.
- Encoding normalization and text representation standardization, when textual processability requirements are not satisfied. This includes enforcing declared character encoding (e.g., UTF-8), applying Unicode normalization forms, and removing non-parseable or corrupted characters that violate representation standards.

D. Stage 4 – Re-Verification

Following correction, all syntactic metrics must be recomputed. By comparing pre- and post-cleaning metrics, it becomes possible to quantify improvement and evaluate the effectiveness of corrective actions. This step closes the verification loop and ensures that syntactic quality is demonstrably enhanced rather than assumed.

VII. Application Case

To illustrate the practical relevance of the proposed framework, a Digital VoC dataset concerning a smartphone in the consumer electronics sector was analyzed. The dataset was constructed from publicly available online customer reviews posted on three major e-commerce and product-review platforms. Reviews were collected through a structured web-scraping procedure over a 12-month observation window, using the smartphone model identifier and related product-page URLs as selection criteria. The final raw dataset included 52,438 reviews. The dataset was selected because it reflects a typical post-market monitoring scenario in which customer feedback is integrated from heterogeneous external sources rather than generated within a controlled internal quality system. Each record contains the following attributes: Review_ID, Reviewer_ID, Platform, Product_ID, Rating, Review_Text, Timestamp, Helpful_Votes, and Total_Votes. The objective of this application is not only to assess syntactic quality, but to demonstrate how syntactic defects can affect the reliability of product quality monitoring.

A. Stage 1 – Metadata Specification

A formal metadata specification was defined prior to analysis, including: (i) rating domain (integer ∈ {1–5}), (ii) timestamp format (ISO 8601), (iii) mandatory attributes (Review_ID, Product_ID, Rating, Timestamp), (iv) uniqueness constraint (one record per Review_ID), (v) cross-field consistency rules, and (vi) UTF-8 encoding with NFC normalization for textual data. This specification provides the formal reference for distinguishing compliant from non-compliant records.

B. Stage 2 – Profiling and Impact of Syntactic Defects

The dataset was profiled against the declared constraints. Several categories of syntactic defects were identified. *Structural validity* issues included ratings expressed as strings (e.g., "five") or non-integer values (e.g., 3.5). *Format violations* were observed in timestamps represented in inconsistent forms (e.g., "08/14/2025" instead of ISO format). *Completeness* issues mainly involved missing timestamps (5.6% of records). *Entity integrity* violations were due to duplicate Review_IDs (4.1%), primarily caused by repeated crawling. *Text processability* problems included encoding errors (e.g., "batterÃa", "\u00E8") and non-normalized Unicode strings. Overall, only 86.7% of records satisfied all syntactic constraints.

These defects were not only structural but had a direct impact on monitoring outputs. This effect becomes particularly relevant in manufacturing contexts, where digital VoC data is used to identify product-related defects and support engineering decisions. For example, several reviews referring to battery overheating issues were duplicated due to repeated data extraction, artificially amplifying the perceived frequency of this defect. Similarly, inconsistent timestamps distorted the temporal distribution of complaints related to device malfunctions, creating misleading signals of sudden quality degradation. In addition, format and encoding inconsistencies fragmented defect-related information. For instance, reviews mentioning "battery issue", "battery problem", and mis-encoded variants (e.g., "batterÃ-a issue") were treated as distinct terms prior to normalization, reducing the accuracy of defect aggregation and topic identification. These issues illustrate how syntactic defects can directly affect the identification, quantification, and prioritization of product quality problems in manufacturing environments.

C. *Stage 3 – Controlled Correction*

Corrective actions were applied in relation to the violated constraints. Type casting harmonization were applied to 1,214 records to restore structural validity. A total of 2,937 records with missing mandatory attributes were removed. Domain and format violations were addressed by excluding 947 out-of-range ratings and standardizing timestamp representations. Entity integrity was restored by removing 1,508 duplicate records. Cross-field inconsistencies (936 records) were either corrected (421 cases) or removed. Finally, 2,468 records required text normalization to ensure proper encoding and representation. After correction, the dataset size decreased from 52,438 to 47,526 records (−9.4%). All interventions were directly derived from the metadata specification and applied in a reproducible manner.

D. *Stage 4 – Re-Verification and Impact on Monitoring*

Following correction, syntactic quality metrics were recomputed. The results are summarized in Table 9.

Table 9. Syntactic Quality Metrics Before and After Cleaning

Metric	Pre-Cleaning	Post-Cleaning	Change
Type Conformance Rate (TCR)	0.962	0.998	+3.6%
Mandatory Attribute Completeness Ratio (MACR)	0.944	1.000	+5.6%
Domain Compliance Rate (DCR)	0.960	1.000	+4.0%
Uniqueness Rate (UR)	0.971	1.000	+2.9%
Compliance Score (CS)	0.982	0.997	+1.5%
Text Representation Compliance Rate (TRCR)	0.953	0.999	+4.6%

All syntactic dimensions improved after correction. The proportion of fully compliant records increased from 86.7% to 99.4%. More importantly, the removal of syntactic defects led to more stable and reliable monitoring signals. In particular, the reduction of duplicate records decreased artificially inflated defect frequencies, while timestamp normalization reduced noise in time-based analyses. This resulted in more consistent product-level comparisons and a more reliable identification of potential quality issues. This application shows that syntactic verification is not merely a preprocessing step, but a necessary condition for reliable Digital VoC-based product quality monitoring. Without such verification, monitoring systems may generate misleading signals, potentially leading to incorrect engineering decisions, such as unnecessary investigations or missed detection of emerging defects.

VIII. Conclusions

This paper addressed the problem of digital VoC data reliability in product quality monitoring by adopting a data-quality perspective grounded in ISO 8000. While prior research has extensively explored methods for extracting insights from digital VoC, limited attention has been devoted to the explicit verification of the quality of the underlying data. The main contribution of this study is the development of a structured and operational framework for assessing the syntactic quality of digital VoC datasets. In particular, the paper provides three original elements. First, it translates the ISO 8000 conceptualization of data quality into the specific context of digital VoC, clarifying the role of syntactic quality as a prerequisite for reliable analytics. Second, it proposes a set of six syntactic subdimensions—structural validity, completeness, format validity, entity integrity, cross-field consistency, and text processability—that organize the main categories of formally verifiable defects in VoC data. Third, it operationalizes these subdimensions through measurable metrics and integrates them into a structured four-stage protocol for syntactic verification and improvement. These subdimensions are not intended as an exhaustive taxonomy of data quality, but as an operational decomposition of ISO 8000 syntactic conformance into verifiable metadata requirement classes for Digital VoC datasets.

The application case demonstrated the practical relevance of the proposed approach, showing that non-negligible levels of syntactic defects may be present even in large-scale VoC datasets and that systematic verification can substantially improve data compliance prior to analysis. In this sense, the framework contributes to strengthening the reliability and auditability of digital VoC-based quality monitoring.

The study has some limitations. First, it focuses exclusively on syntactic quality and does not address semantic aspects, such as the correctness or relevance of review content, nor pragmatic aspects related to fitness for specific decision-making contexts. Second, although the proposed metrics are formally defined, the identification of acceptable thresholds and governance rules remains context-dependent and requires further investigation.

Future research may extend this work by integrating semantic and pragmatic dimensions into a comprehensive reliability assessment framework for digital VoC. Additional studies could explore cross-domain applications, benchmark ranges for syntactic quality metrics, and the integration of syntactic verification into automated VoC analytics pipelines for Quality 4.0 systems.

References

Barravecchia, F., Franceschini, F., & Mastrogiacomo, L. (2018). A service network perspective to evaluate service matching in early design. *Journal of Service Theory and Practice*, *28*(3), 356-383.

Barravecchia, F., Mastrogiacomo, L., & Franceschini, F. (2022a). KA-VoC map: classifying product key-attributes from digital voice-of-customer. *Quality Engineering*, *34*(3), 344-358.

Barravecchia, F., Mastrogiacomo, L., & Franceschini, F. (2022b). Digital voice-of-customer processing by topic modelling algorithms: insights to validate empirical results. *International Journal of Quality & Reliability Management*, *39*(6), 1453-1470.

Barravecchia, F., Mastrogiacomo, L. and Franceschini, F. (2023) 'Product quality tracking based on digital Voice-of-Customers', Total Quality Management & Business Excellence, 34 (11–12), 1386–1409.

Barravecchia, F., Mastrogiacomo, L. and Franceschini, F. (2026) 'Digital VoC analysis for product/service quality tracking in the era of Quality 4.0', International Journal of Quality & Reliability Management, 43 (2), 499–515.

Broday, E. E. (2022) 'The evolution of quality: from inspection to quality 4.0', International Journal of Quality and Service Sciences, 14 (3), 368–382.

Elmagarmid, A. K., Ipeirotis, P. G. and Verykios, V. S. (2007) 'Duplicate record detection: A survey', IEEE Transactions on knowledge and data engineering, 19 (1), 1–16.

Esposito, C., Moscato, V. and Sperlí, G. (2021) 'Trustworthiness assessment of users in social reviewing systems', IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52 (1), 151–165.

Fellegi, I. P. and Sunter, A. B. (1969) 'A theory for record linkage', Journal of the American statistical association, 64 (328), 1183–1210.

Fox, C., Levitin, A. and Redman, T. (1994) 'The notion of data and its quality dimensions', Information processing & management, 30 (1), 9–19.

Franceschini, F., Galetto, M., Maisano, D., & Neely, A. D. (2019). Designing performance measurement systems, Theory and Practice of Key Performance Indicators. *Springer, Cham (Switzerland)*

Guerra-García, C., Nikiforova, A., Jiménez, S., Perez-Gonzalez, H. G., Ramírez-Torres, M. and Ontañon-García, L. (2023) 'ISO/IEC 25012-based methodology for managing data quality requirements in the development of information systems: Towards Data Quality by Design', Data & Knowledge Engineering, 145 , 102152.

Hu, N., Bose, I., Koh, N. S. and Liu, L. (2012) 'Manipulation of online reviews: An analysis of ratings, readability, and sentiments', Decision support systems, 52 (3), 674–684.

International Organization for Standardization (2011) 'ISO 8000-1: Data quality — Part 1: Overview'.

International Organization for Standardization (2015) 'ISO 8000-8: Data quality — Part 8: Information and data quality: Concepts and measuring'.

Kakde, S. T., Roychowdhury, S., Bhosale, A. T. and Maiti, J. (2024) 'CPAN chart: A novel customer perception analysis system using natural language processing and attribute control charting', IEEE Transactions on Engineering Management, 71 , 7609–7622.

Kim, R. Y. (2021) 'Using online reviews for customer sentiment analysis', IEEE Engineering Management Review, 49 (4), 162–168.

Kim, W. G. and Park, S. A. (2017) 'Social media review rating versus traditional customer satisfaction: Which one has more incremental predictive power in explaining hotel performance?', International Journal of Contemporary Hospitality Management, 29 (2), 784–802.

Mastrogiacomo, L., Barravecchia, F., Franceschini, F., & Marimon, F. (2021). Mining quality determinants of product-service systems from user-generated contents. *Quality Engineering*, *33*(3), 425-442.

Mewada, A. and Dewang, R. K. (2023) 'A comprehensive survey of various methods in opinion spam detection', Multimedia Tools and Applications, 82 (9), 13199–13239.

Özdağoğlu, G., Kapucugil-İkiz, A. and Çelik, A. F. (2018) 'Topic modelling-based decision framework for analysing digital voice of the customer', Total Quality Management & Business Excellence, 29 (13–14), 1545–1562.

Park, S., Joung, J. and Kim, H. (2026) 'Large Language Model-Based Online Review Classification for Subfeature-Level Customer Opinion Analysis', Journal of Mechanical Design, 148 (4), 041701.

Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002) 'Data quality assessment', Communications of the ACM, 45 (4), 211–218.

Rahm, E. and Do, H. H. (2000) 'Data cleaning: Problems and current approaches', IEEE Computer Society Data Engineering Bulletin, 23 (4), 3–13.

Ridzuan, F. and Zainon, W. M. N. W. (2024) 'A review on data quality dimensions for big data', Procedia computer science, 234 , 341–348.

Santhiran, R., Varathan, K. D. and Chiam, Y. K. (2024) 'Feature extraction from customer reviews using enhanced rules', Peerj Computer Science, 10 , e1821.

See-To, E. W. K. and Ngai, E. W. T. (2018) 'Customer reviews for demand distribution and sales nowcasting: a big data approach', Annals of Operations Research, 270 (1), 415–431.

Shankaranarayanan, G. and Cai, Y. (2006) 'Supporting data quality management in decision-making', Decision Support Systems, 42 (1), 302–317.

Subroto, A. and Christianis, M. (2021) 'Rating prediction of peer-to-peer accommodation through attributes and topics from customer review', Journal of Big Data, 8 (1), 9.

Ullah, R., Amblee, N., Kim, W. and Lee, H. (2016) 'From valence to emotions: Exploring the distribution of emotions in online product reviews', Decision Support Systems, 81 , 41–53.

Wand, Y. and Wang, R. Y. (1996) 'Anchoring data quality dimensions in ontological foundations', Communications of the ACM, 39 (11), 86–95.

Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S. and Aggarwal, S. (2024) 'Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality', Journal of the Knowledge Economy, 15 (1), 1159–1178.

Wang, R. Y. and Strong, D. M. (1996) 'Beyond accuracy: What data quality means to data consumers', Journal of management information systems, 12 (4), 5–33.

Yan, L., Tao, B., Han, Z. and Ouyang, L. (2025) 'Online reviews analysis in product defects and customer requirements via two-stage model', Total Quality Management & Business Excellence, 36 (7–8), 788–810.

Zonnenshain, A. and Kenett, R. S. (2020) 'Quality 4.0—the challenging future of quality engineering', Quality Engineering, 32 (4), 614–626.