A Categorical Error Sensitivity Index (ISEC): A Preventive Ordinal Decision-Support Index for Irrecoverable Errors in Data Entry Systems

Ricardo Raúl Palma
Universidad Nacional de Cuyo
Instituto de Ingeniera
Mendoza, Argentina
rpalma@uncu.edu.ar

Mauro Anibal Benetti
Universidad Tecnológica Nacional
Facultad Regional San Rafael
Mendoza, Argentina
mbenetti@frsr.utn.edu.ar

Fabricio Orlando Sanchez Varretti
Universidad Tecnológica Nacional
Facultad Regional San Rafael
Mendoza, Argentina
fsanchez@frsr.utn.edu.ar

Abstract. Data entry systems remain structurally vulnerable to categorical misclassifications, particularly in small and medium enterprises. When nominal categories exhibit semantic or morphological proximity, human–machine interaction may produce errors that are irrecoverable ex post. In the absence of automated input controls, manual data entry frequently generates irrecoverable categorical distortions that propagate into Key Performance Indicators (KPIs), thereby misleading the managerial decision-making.

State of the art normalization tools typically evaluate semantic and morphological dimensions in isolation and rely heavily on standard dictionaries, rendering them ineffective for SME master data rich in custom SKUs, abbreviations, and domain-specific technical jargon.

This paper introduces the Categorical Error Sensitivity Index (ISEC), an ordinal composite index designed to rank category pairs according to their structural susceptibility to confusion. ISEC integrates semantic distance (via word embeddings), custom-weighted morphological transformation costs (through an adapted Damerau–Levenshtein algorithm), and empirical frequency into a unified, mathematically robust preventive framework. By leveraging vector database architecture, ISEC also reduces computational complexity, achieving approximately a 195× performance improvement over brute-force methods. Validated across three heterogeneous datasets: governmental judicial records, retail inventory, and a synthetic ISO-coded metalworking catalog. The ISEC provides a scalable and proactive data governance instrument that enables SMEs to detect latent structural risk embedded within their categorical data assets.

Keywords: Data Quality, ISEC, User Generated Text, Semantic Similarity, Damerau-Levenshtein

Un índice de sensibilidad para errores categóricos (ISEC): Un índice ordinal preventivo para errores irrecuperables en los sistemas de entrada de datos

Resumen. Los sistemas de entrada de datos siguen siendo estructuralmente vulnerables a clasificaciones erróneas categóricas, especialmente en las pequeñas y medianas empresas (pymes). Cuando las categorías nominales exhiben proximidad semántica o morfológica, la interacción hombre-máquina puede producir errores que son irrecuperables ex post. En ausencia de controles de entrada, la entrada manual genera frecuentemente distorsiones categóricas irrecuperables que se propagan a los indicadores clave de rendimiento (KPI) desinformando así la toma de decisiones gerenciales.

Las herramientas de normalización de última generación suelen evaluar las dimensiones semánticas y morfológicas de forma aislada y dependen en gran medida de diccionarios estándar, lo que las vuelve ineficaces para los datos maestros de las pymes ricos en SKU personalizados, abreviaturas y jerga técnica específica del dominio.

Este artículo presenta el Índice de Sensibilidad al Error Categórico (ISEC), un índice compuesto ordinal diseñada para ordenar pares de categorías según su susceptibilidad estructural a la confusión. El ISEC integra la distancia semántica (a través de incrustaciones de palabras), costos de transformaciones morfológica ponderadas personalizadas (utilizando Damerau-Levenshtein adaptado) y la frecuencia empírica en un marco preventivo unificado y matemáticamente sólido. Al aprovechar las arquitecturas de bases de datos vectoriales, ISEC también reduce la complejidad computacional, logrando aproximadamente una mejora de rendimiento de 195 veces con respecto a métodos de fuerza bruta. Validado en tres conjuntos de datos heterogéneos: registros judiciales gubernamentales, inventario minorista y códigos de la ISO-1832, el ISEC proporciona un instrumento de gobernanza de datos escalable y proactivo para la pyme.

Palabras clave: Calidad de datos, ISEC, texto generado por usuarios, similitud semántica, Damerau-Levenshtein.

1 Introduction

Small and medium enterprises (SMEs) operate under structural constraints in technological resources, data governance capacity, and formalized information management processes (Alsousi & Asadullah Shah, 2022; Goh Zhi Ji & Jugindar Singh Kartar Singh, 2023; Günther et al., 2019). In such environments, the human–machine interface remains as a weak link in the data ingestion process (Aliero et al., 2023). Unlike large corporations with automated validation pipelines and structured master data governance, SMEs frequently rely on manual categorical inputs embedded in information systems, spreadsheets, and operational databases (Feit et al., 2016; Goh Zhi Ji & Jugindar Singh Kartar Singh, 2023; Günther et al., 2019).

This structural reliance on manual data capture introduces a critical yet under-theorized vulnerability: categorical fragility (Unterkalmsteiner & Abdeen, 2023). Nominal categories, product codes, supplier names, incident types, labels, are treated as stable semantic units. However, when entered manually, these categories become susceptible to typographical variations, morphological distortions, semantic ambiguity, and inconsistent abbreviations (Feit et al., 2016; Navarro, 2001; van der Goot & van Noord, 2017).

This paper argues that categorical sensitivity to error is not a downstream data quality issue (Wand & Wang, 1996), but a measurable property of system design. We introduce the **Categorical Errors Sensitivity Index (ISEC)**, an ordinal metric designed to rank nominal taxonomies according to their latent vulnerability to confusion prior to error occurrence (Lorena et al., 2020).

ISEC shifts data quality management from reactive correction toward preventive risk modeling (Cichy & Rass, 2019; Miller et al., 2025), particularly at the level of Decision Support Systems (DSS), where aggregated categorical distortions may produce structured misinformation (Ghasemaghaei et al., 2018; Mendes Sampaio et al., 2015; Shrestha et al., 2021; Velden et al., 2024).

2 Problem Statement and Situational Context

2.1 Structured Misinformation and Algorithmic Indeterminacy

Decision Support Systems (DSS) rely heavily on aggregated nominal data to compute Key Performance Indicators (KPIs) (Mendes Sampaio et al., 2015). If the underlying taxonomy is fragile, aggregation may produce structured misinformation: internally consistent metrics derived from semantically unstable inputs (Simard et al., 2019; Wang & Strong, 1996).

Categorical fragility manifests critically in an operational regime where the effective structural separation between two valid categories is naturally short (Lorena et al.,

2020; Unterkalmsteiner & Abdeen, 2023). Crucially, an error becomes **irrecoverable** not necessarily because the structural distance between the intended label and another label drops to zero. Rather, an error is irrecoverable when a small stochastic perturbation (e.g. typographical error) places the erroneous entry equidistant between two valid categories, or strictly closer to an incorrect one (Navarro, 2001).

Once this state of structural ambiguity is reached, systems operating solely on internal similarity measures lack sufficient discriminatory power to deterministically assign the correct category (Malkov & Yashunin, 2020; Navarro, 2001).

2.2 Limitations of Current Metrics: The Disconnection of Dimensions

The literature reveals a critical gap in the treatment of non-standard words (e.g., SKUs, industry acronyms, and alphanumeric ISO codes) (Aliero et al., 2023; Miller et al., 2025). State of the art (SOTA) text normalization systems, demand massive computational power or extensive training corpora (Aliero et al., 2023; Johnson et al., 2024; Shrestha et al., 2021), resources unavailable to most SMEs and specially the Argentinean SME (Alsousi & Asadullah Shah, 2022; Günther et al., 2019).

Furthermore, traditional metrics suffer from the "disconnection of dimensions" (Cichy & Rass, 2019; Wang & Strong, 1996). They evaluate morphological distance (how a word is written) and semantic distance (what a word means) independently (Johnson et al., 2024; Mendes Sampaio et al., 2015; Po, 2020; Unterkalmsteiner & Abdeen, 2023). Standard spell-checkers evaluate structural edits but ignore the semantic plausibility of the correction (Po, 2020; Zhao & Sahni, 2019), while semantic models fail to account for the physical ergonomics of typing errors. The ISEC bridges this gap by unifying these dimensions into a single actionable index.

2.3 Irrecoverable Errors

An error is considered irrecoverable when:

- It produces a syntactically valid category.
- It does not violate database constraints.
- It remains undetected during aggregation (Simard et al., 2019).
- The original intended label cannot be reconstructed ex post (Aliero et al., 2023; Berger et al., 2021; van der Goot & van Noord, 2017).

Such errors differ fundamentally from missing data or format violations (Aliero et al., 2023; Cichy & Rass, 2019; Feit et al., 2016; Miller et al., 2025).

2.4 Human–Machine Interface Weakness

SMEs depend heavily on manual UI forms without adaptive validation or semantic distance monitoring (e.g. excel sheets) (Goh Zhi Ji & Jugindar Singh Kartar Singh,

2023; Günther et al., 2019; Shrestha et al., 2021). Standard spell-checkers are ineffective because SMEs master data often include:

- Technical nomenclatures
- Industry-specific acronyms
- Non-standard abbreviations
- Alpha-numeric codes
- Evolving names and abbreviations (Aliero et al., 2023; Miller et al., 2025; van der Goot & van Noord, 2017)

Thus, the interface itself becomes a structural generator of latent informational entropy (Lorena et al., 2020; Wand & Wang, 1996).

3 Theoretical Positioning

Rather than defining quality as conformance to specification ex post (Wand & Wang, 1996; Wang & Strong, 1996), we conceptualize quality as **ex ante structural robustness**. A taxonomy with high categorical proximity is intrinsically fragile regardless of whether errors have already been observed (Lorena et al., 2020; Unterkalmsteiner & Abdeen, 2023).

3.1 Distance Compression and Probabilistic Collapse

It is important to clarify that categorical sensitivity does not arise merely because two categories exhibit short structural distance. Proximity alone does not constitute an error. Rather, sensitivity emerges when an already small inter-category distance is exposed to stochastic perturbations induced by capture process (Feit et al., 2016; Wand & Wang, 1996; Zhao & Sahni, 2019).

Let $d(i,j)$ denote the composite structural distance between two categories (Navarro, 2001; Velden et al., 2024). Consider a random perturbation operator ε representing a typographical noise, abbreviation variability, or human transcription error (Aliero et al., 2023; van der Goot & van Noord, 2017). The effective post-perturbation distance becomes:

$$d'(i,j) = d(i,j) - \varepsilon$$

where ε is a non-negative stochastic variable bounded by, the maximum transformation cost induced by human input variability (Feit et al., 2016; Lorena et al., 2020; Navarro, 2001).

When $d(i,j)$ is sufficiently small, there exists a non-zero probability that:

$$d(i,j) \leq \delta$$

Under this condition, the effective discriminative margin between i and j falls below the admissible threshold δ (Lorena et al., 2020). From the perspective of a normalization or a correction algorithm, both categories may become plausible candidates within the tolerance bounds of the correction mechanism (Aliero et al., 2023; Zhao & Sahni, 2019). The classification decision ceases to be strictly deterministic and becomes probabilistically underdetermined (Simard et al., 2019).

Thus, categorical fragility is not simply a function of proximity, but of **proximity under perturbation** when some perturbations are more likely than others (Feit et al., 2016; Lorena et al., 2020). The smaller the initial structural separation, the lower the perturbation threshold required to generate classification ambiguity. This phenomenon may be termed **distance compression under stochastic noise**.

4 Formal Definitions

4.1 ISEC Definition

The Index of Sensitivity to Categorical Errors for a pair of categories (i, j) is defined as:

$$ISEC_{i,j} = \frac{1 + FMN_{(i,j)}}{DSN_{(i,j)}{}^{\alpha} \cdot CMP^{1-\alpha}{}_{(i,j)}}$$

Where: $FMN_{(i,j)}$ is the normalized mean frequency of the pair, $DSN_{(i,j)}$ is the normalized semantic distance, $CMP_{(i,j)}$ is the penalized mean transformation cost, computed using a weighted Damerau-Levenshtein distance in which each edit operation is assigned a cost according to a custom pair-wise character transformation matrix. The parameter $\alpha \in [0,1]$ is a tuning parameter that determines the semantic vs. morphological weight. When $\alpha \to 1$ prioritizes meaning (better for long categories and descriptions), while $\alpha \to 0$ prioritizes structure (better for SKUs and short alphanumeric codes). ISEC is an unbounded positive real-valued index used exclusively for ordinal ranking of categorical sensitivity, and it does not represent a probability measure.

4.2 Domain of ISEC

Let $C = \{i, j, \dots, n\}$ be a finite set of nominal categories with $n \geq 2$. The Index of Sensitivity to Categorical Errors is defined over the domain:

$$D = \{(i, j) \in C \times C \wedge i \neq j\}$$

that is, over ordered pairs of **distinct categories**. For every $(i, j) \in D$, assume:

$$DSN > 0 \ \wedge CMP > 0$$

These strict positivity conditions exclude trivial identity cases and prevent denominator degeneracy.

4.3 FMN (Normalized mean frequency of the pair):

$$FMN_{(i,j)} = log_{10}\left(FM_{(i,j)}\right)$$

Where:

$$FM_{(i,j)} = \frac{\left(F_i + F_j\right)}{2}$$

4.4 DSN (Normalized Semantic Distance):

It represents the semantic dissimilarity between categories, obtained as the complement of the cosine similarity between their *vector embeddings* (Xie et al., 2023)

$$DSN_{(a,b)} = \frac{\left(1 - CosenoSimilaridad_{(a,b)}\right)}{2}$$

Where:

$$CosenoSimilaridad_{(a,b)} = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \times \|\vec{b}\|} \in [-1,1]$$

4.5 Penalized Mean Cost (CMP):

$$CMP_{(i,j)} = CM_{(i,j)} + k \times CP_{(i,j)}$$

where $k \geq 0$ regulates the relative influence of cumulative linear structural divergence.

Mean Transformation Cost (CM).
Let the transformation of i into j require a sequence of N weighted edit operations under a customized Damerau-Levenshtein scheme. Let c_k denote the cost of the k-th operation as specified by the shared pair-wise character cost matrix. The mean transformation cost is defined as:

$$CM_{(i,j)} = \frac{1}{N}\sum_{K=1}^{N} c_K$$

Penalty Cost (CP).
We design *CP* to accumulate only insertions, deletions, and substitutions. Transpositions are excluded to preserve linear growth in string-length divergence, a modification of the classical Damerau-Levenshtein framework , which treats all four operations symmetrically (Zhao & Sahni, 2019).

$$CP_{(i,j)} = \sum_{L \in \mathcal{L}} w_L \times n_L(i,j)$$

Where: $\mathcal{L} = \{insertion, deletion, substitution\}$

Conceptually, CM captures the average morphological effort required to transform one category into another, while CP captures the total accumulated linear structural deviation. The penalization term $k \times CP$ amplifies the cases in which divergence is driven by numerous operations.

5 Implications for Data Quality and Information Systems

ISEC operates at the pre-aggregation layer (Cichy & Rass, 2019; Simard et al., 2019). Its strategic relevance lies in:

- Identifying fragile taxonomies before or after its deployment (Lorena et al., 2020; Unterkalmsteiner & Abdeen, 2023)
- Supporting UI/UX redesign decisions(Unterkalmsteiner & Abdeen, 2023; Wang & Strong, 1996)
- Guiding category consolidation if necessary (Lorena et al., 2020; Unterkalmsteiner & Abdeen, 2023)
- Preventing propagation of irrecoverable distortions.
- Provide additional metadata about the underlying information used for any KPI construction in particular (Ghasemaghaei et al., 2018; Mendes Sampaio et al., 2015; Shrestha et al., 2021)

Unlike post-hoc audits, ISEC models **latent structural risk** (Lorena et al., 2020; Miller et al., 2025; Simard et al., 2019). This is particularly relevant for environments, where constraints prevent continuous data cleansing cycles.

6 Computational Optimization done for ISEC

A brute-force calculation of both distances (semantic and morphological) for N categories results in $O(N^2)$ complexity (Navarro, 2001; Zhao & Sahni, 2019), making it nonviable for catalogs exceeding a few thousand items (Berger et al., 2021).

To ensure technological sovereignty and low implementation costs (Günther et al., 2019; Navarro, 2001; van der Goot & van Noord, 2017), ISEC is architected for open-source stacks using a **Hybrid Search Strategy** via Vector Databases (e.g., ChromaDB with algorithms (Malkov & Yashunin, 2020; Xie et al., 2023)) :

- **Breadth-First (Semantic)**: The system queries the VectorDB to retrieve only the Top-K semantic neighbors for a given category in approximately $O(\log N)$ time (Johnson et al., 2024; Malkov & Yashunin, 2020; Reimers & Gurevych, 2019).

- **Greedy (Morphological):** The *CMP* is calculated against these Top-K candidates in *O(K)* time (Malkov & Yashunin, 2020). When Top-K is equal to the number of all categories, the hybrid search is reduced to a simple greedy search.

This reduces overall complexity to $O(\log N + k)$ (Malkov & Yashunin, 2020). In empirical testing with 1,000 categories, processing time dropped from 8.4 hours to 2.6 minutes, representing an efficiency of approximately 195× over brute-force methods.

7 Empirical Validation Across Heterogeneous Domains

To evaluate the practical applicability of ISEC under real-world constraints, the metric was validated across three heterogeneous datasets, originally presented in doctoral research. All experiments were executed on a virtual machine equipped with 4 vCPUs (2 CPU cores at 3.8 GHz) and 8 GB of RAM, without GPU acceleration. The ISEC implementation used approximately 2.7 GB of RAM, of which 1 GB was allocated to the embedding model and the remained memory to the Python execution environment.

7.1 Case 1: Judicial Administrative Records (CAJ – Argentina)

The first validation scenario utilized open-access data from the *Centro de Acceso a la Justicia (CAJ)*, an Argentine Governmental legal assistance network. The dataset consisted of 1,069,280 records across 32 variables, collected between 2016 and 2019[1].This Dataset was selected because is a good example of a taxonomy well known (provinces in Argentina), and this dataset had an organic growth of the variations with the correspondent frequencies over time.

The ISEC identified extreme vulnerabilities in short, high-frequency abbreviations. For instance, "*caba*", the abbreviation of "*Ciudad Autónoma de Buenos Aires*" and "*cba*" (Córdoba) differ by a single character. The pair "*cba*" also heavily conflicted with other abbreviations like "*pba*", "*gba*", and "*ba*". The combined frequency of these highly sensitive abbreviations exceeded 56,000 records (accounting for 6-7% of total entries). It was flagged that a single typo in these abbreviations causes irrecoverable misclassification or loss of the datapoint given the irreversibility of the change. Identifying this vulnerability allowed for proactive recommendations such as targeted autocomplete suggestions only for those categories or requesting a confirmation from the user about the intention when the correspondent abbreviations is used.

[1] https://datos.gob.ar/it/dataset/justicia-consultas-efectuadas-centros-acceso-justicia--caj-

7.2 Case 2: Retail Supermarket Inventory

The second case examined a retail SME inventory database comprising 25,638 manually entered product descriptions distributed across 14 ontological groups (e.g., *"Aseo de hogar"*, *"Cuidado Personal"*, *"Charcutería"*). Unlike the CAJ dataset, this environment exhibited long descriptive strings (sentences) with high lexical variability and semantic density , the dataset can be found in link[2]. The analysis focused on an inter-group comparison, evaluating items from one group against items in all other groups to detect latent overlaps, yielding a massive search space of 297,404,181 unique pairs. ISEC successfully revealed structural duplicates and semantic collisions spanning different classification groups. Critical vulnerabilities identified included:

- "*Bizcocho achiras del Huila*" assigned to the *Panadería y Pastelería* group vs. "*Bizcocho achiras del huila*" assigned to *Dulces y postres*.
- "*Quitamanchas Fab ropa blanca líquido*" in *Aseo de hogar* vs. "*Quitamanchas Fab ropa color blanca líquido*" in *Cuidado de ropa y calzado*.

These findings empirically demonstrate that semantic embeddings alone are insufficient for quality assessment when naming conventions share critical structural tokens across domain boundaries (Po, 2020). Integrating penalized morphological costs allowed ISEC to surface the most affected categories by lower cost transformations.

7.3 Case 3: Synthetic ISO-Based Dataset

The third validation scenario employed a controlled synthetic dataset structured according to ISO 1832, the international standard defining nomenclature for indexable inserts in metal cutting tools. While the standard allows for nearly 100 million theoretically unique combinations, SMEs typically operate with a much smaller subset. A list of 1,000 valid codes was generated to test the index.

ISEC identified "silent" vulnerabilities where minor typographical noise generated entirely different but completely valid codes. For example, the code "AAGX110216" was highly sensitive to transforming into "AGAX110216" due to one transposition $A \leftrightarrow G$.

7.4 Cross-Case Observations

Across the three domains, several structural regularities emerged:

- Abbreviation systems significantly reduce effective inter-category distance (Aliero et al., 2023; van der Goot & van Noord, 2017).
- Semantic embeddings or morphologic distances alone are insufficient for discriminability(Johnson et al., 2024; Velden et al., 2024)

[2] https://www.kaggle.com/datasets/camiloemartinez/productos-consumo-masivo

- When empirical frequency data is unavailable, as in pre-deployment catalogs, the index operates exclusively on structural proximity, which is precisely the intended behavior.

Most critically, in all three cases, ISEC identified high-risk categorical zones before normalization procedures were applied, confirming its preventive capacity (Cichy & Rass, 2019; Günther et al., 2019; Wand & Wang, 1996). These results support the central claim of this paper: categorical fragility is a measurable structural property that precedes observable data quality failure (Lorena et al., 2020; Unterkalmsteiner & Abdeen, 2023).

8 Methodological Advantages

The ISEC framework introduces several methodological contributions that distinguish it from traditional data quality assessment and string similarity approaches.

- **Ex-Ante Structural Risk Modeling**: Conventional data quality methodologies operate reactively, identifying errors after they have been instantiated in the data (Cichy & Rass, 2019; Wand & Wang, 1996; Wang & Strong, 1996). In contrast, ISEC models the **latent structural vulnerability** prior to error occurrence (Simard et al., 2019; Unterkalmsteiner & Abdeen, 2023).
- **Integration of Morphological, Semantic, and Frequency Dimensions**: Traditional edit-distance metrics focus exclusively on syntactic transformations (Aliero et al., 2023; Navarro, 2001; Zhao & Sahni, 2019). Semantic similarity measures, in isolation also fail to capture typographical perturbation dynamics of real scenarios (Feit et al., 2016; Johnson et al., 2024; Reimers & Gurevych, 2019).

9 Discussion

This work reframes categorical data quality as a problem of **structural robustness under perturbation**. We conceptualize fragility as an emergent property of the taxonomy (Unterkalmsteiner & Abdeen, 2023; Velden et al., 2024). Furthermore, we understand the taxonomy fragility as the interaction of 3 elements: (1) Taxonomy geometry, (2) Capture process (source of the perturbation), (3) the algorithm discriminatory power. The ISEC calculates sensitivity using the first two elements, whereas the third, the normalization algorithm, acts as an external moderating factor determined by the state of the art. In the limit, assuming the existence of a perfectly discriminative normalization function, structural perturbations would not produce classification indeterminacy. Only under such conditions, the index collapses in practical relevance, as fragility would no longer produce normalization uncertainties.

9.1 Boundary Conditions

ISEC is designed for nominal categorical systems characterized by manual or semi-manual input processes (Feit et al., 2016). It is less applicable to:

- Ontologically constrained taxonomies with enforced unique identifiers (Unterkalmsteiner & Abdeen, 2023),
- High-dimensional continuous feature spaces.

9.2 Beyond SMEs

Although motivated by SME constraints (Alsousi & Asadullah Shah, 2022; Goh Zhi Ji & Jugindar Singh Kartar Singh, 2023; Günther et al., 2019), the framework generalizes to any environment where exist; categories normalization and aggregation, there is a capture or ingestion process that introduces perturbation such as speech to text (STT) or optical character recognition (OCR) (Aliero et al., 2023; Lorena et al., 2020).

9.3 From Reactive Data Quality to Information Robustness

Classical data quality frameworks conceptualize quality as conformance to specification after data instantiation (Cichy & Rass, 2019; Wang & Strong, 1996). Dimensions such as accuracy, completeness, and consistency evaluate records relative to predefined constraints (Miller et al., 2025; Wand & Wang, 1996).

The ISEC suggests a complementary perspective: quality as 'structural robustness' prior to error occurrence (Simard et al., 2019; Unterkalmsteiner & Abdeen, 2023). We propose **structural robustness** as the degree to which a categorical system preserves semantic separability under bounded stochastic perturbation. Under this perspective:

- Data accuracy evaluates correctness **ex post** (Wang & Strong, 1996).
- Information robustness evaluates separability **ex ante** (Unterkalmsteiner & Abdeen, 2023).

10 Conclusion

Categorical errors in SMEs are not merely operational nuisances; they are potential generators of structured misinformation within Decision Support Systems (Mendes Sampaio et al., 2015; Shrestha et al., 2021). Some of these errors are irrecoverable once embedded. The ISEC provides an ordinal, preventive, and scalable framework to measure categorical sensitivity prior to error manifestation (Lorena et al., 2020; Unterkalmsteiner & Abdeen, 2023). By modeling inter-category proximity and modeling transformation costs (Navarro, 2001; Zhao & Sahni, 2019), it operationalizes a previously unmeasured dimension of information quality: **Structural robustness**.

11 Acknowledgments

During the preparation of this work, the author(s) used Microsoft Copilot to improve the readability of the manuscript. After using this tool, the author(s) reviewed and edited the content as needed and assume full responsibility for the content.

12 References

Aliero, A. A., Adebayo, B. S., Aliyu, H. O., Tafida, A. G., Kangiwa, B. U., & Dankolo, N. M. (2023). Systematic Review on Text Normalization Techniques and its Approach to Non-Standard Words. *International Journal of Computer Applications*, *185*(33), 44–55. https://doi.org/10.5120/ijca2023923106

Alsousi, A., & Asadullah Shah. (2022). DATA GOVERNANCE FOR SME: SYSTEMATIC LITERATURE REVIEW. *Journal of Information Systems and Digital Technologies*, *4*(2), 2022. https://doi.org/10.31436/jisdt.v4i2.237

Berger, B., Waterman, M. S., & Yu, Y. W. (2021). Levenshtein Distance, Sequence Comparison and Biological Database Search. *IEEE Transactions on Information Theory*, *67*(6), 3287–3294. https://doi.org/10.1109/TIT.2020.2996543

Cichy, C., & Rass, S. (2019). An Overview of Data Quality Frameworks. *IEEE Access*, *7*, 24634–24648. https://doi.org/10.1109/ACCESS.2019.2899751

Feit, A. M., Weir, D., & Oulasvirta, A. (2016). How we type: Movement strategies and performance in everyday typing. *Conference on Human Factors in Computing Systems - Proceedings*, 4262–4273. https://doi.org/10.1145/2858036.2858233/SUPPL_FILE/P4262-FEIT.MP4

Ghasemaghaei, M., Ebrahimi, S., & Hassanein, K. (2018). Data analytics competency for improving firm decision making performance. *The Journal of Strategic Information Systems*, *27*(1), 101–113. https://doi.org/10.1016/j.jsis.2017.10.001

Goh Zhi Ji, & Jugindar Singh Kartar Singh. (2023). Digitalization and its Impact on Small and Medium-sized Enterprises (SMEs): An Exploratory Study of Challenges and Proposed Solutions. *International Journal of Business and Technology Management*, *5*. https://doi.org/10.55057/ijbtm.2023.5.4.22

Günther, L. C., Colangelo, E., Wiendahl, H.-H., & Bauer, C. (2019). Data quality assessment for improved decision-making: a methodology for small and medium-sized enterprises. *Procedia Manufacturing*, *29*, 583–591. https://doi.org/10.1016/j.promfg.2019.02.114

Johnson, S. J., Murty, M. R., & Navakanth, I. (2024). A detailed review on word embedding techniques with emphasis on word2vec. *Multimedia Tools and Applications*, *83*(13), 37979–38007. https://doi.org/10.1007/S11042-023-17007-Z/METRICS

Lorena, A. C., Garcia, L. P. F., Lehmann, J., Souto, M. C. P., & Ho, T. K. (2020). How Complex Is Your Classification Problem? *ACM Computing Surveys*, *52*(5), 1–34. https://doi.org/10.1145/3347711

Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 824–836. https://doi.org/10.1109/TPAMI.2018.2889473

Mendes Sampaio, S. de F., Dong, C., & Sampaio, P. (2015). DQ2S – A framework for data quality-aware information management. *Expert Systems with Applications*, *42*(21), 8304–8326. https://doi.org/10.1016/j.eswa.2015.06.050

Miller, R., Chan, S. H. M., Whelan, H., & Gregório, J. (2025). A Comparison of Data Quality Frameworks: A Review. *Big Data and Cognitive Computing*, *9*(4), 93. https://doi.org/10.3390/bdcc9040093

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, *33*(1), 31–88. https://doi.org/10.1145/375360.375365

Po, D. K. (2020). Similarity Based Information Retrieval Using Levenshtein Distance Algorithm. *International Journal of Advances in Scientific Research and Engineering*, *06*(04), 06–10. https://doi.org/10.31695/IJASRE.2020.33780

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 3982–3992. https://doi.org/10.18653/V1/D19-1410

Shrestha, Y. R., Krishna, V., & von Krogh, G. (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Journal of Business Research*, *123*, 588–603. https://doi.org/10.1016/j.jbusres.2020.09.068

Simard, V., Rönnqvist, M., Lebel, L., & Lehoux, N. (2019). A General Framework for Data Uncertainty and Quality Classification. *IFAC-PapersOnLine*, *52*(13), 277–282. https://doi.org/10.1016/j.ifacol.2019.11.181

Unterkalmsteiner, M., & Abdeen, W. (2023). A compendium and evaluation of taxonomy quality attributes. *Expert Systems*, *40*(1). https://doi.org/10.1111/exsy.13098

van der Goot, R., & van Noord, G. (2017). *MoNoise: Modeling Noise Using a Modular Normalization System*. https://doi.org/https://doi.org/10.48550/arXiv.1710.03476

Velden, M. van de, D'Enza, A. I., Markos, A., & Cavicchia, C. (2024). A general framework for implementing distances for categorical variables. *Pattern Recognition*, *153*, 110547. https://doi.org/10.1016/j.patcog.2024.110547

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95. https://doi.org/10.1145/240455.240479

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Xie, X., Liu, H., Hou, W., & Huang, H. (2023). A Brief Survey of Vector Databases. *2023 9th International Conference on Big Data and Information Analytics (BigDIA)*, 364–371. https://doi.org/10.1109/BigDIA60676.2023.10429609

Zhao, C., & Sahni, S. (2019). String correction using the Damerau-Levenshtein distance. *BMC Bioinformatics*, *20*(11), 1–28. https://doi.org/10.1186/S12859-019-2819-0/FIGURES/24