



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Metrics for Measuring the Quality of Unstructured Data

TESI DI LAUREA MAGISTRALE IN
COMPUTER SCIENCE AND ENGINEERING -
INGEGNERIA INFORMATICA

Author: **Javier Alejandro Ortiz Silva**

Student ID: 10913979

Advisor: Prof. Cinzia Cappiello

Academic Year: 2024-2025

Abstract

Subjectivity can contribute to a holistic understanding and assessment of fitness for heterogeneous data. Questionnaire-based methods prove valuable for measuring quality, diagnosing problems, and raising quality awareness. Models trained on human-annotated datasets can learn quality judgment patterns. Although practice is domain-dependent, past works of the data quality discipline provide foundations at the high and intermediate levels of abstraction. These include consumer-oriented views, dimension descriptions, generic metric formulations, and conceptual modeling to separate intentionality from data representations. This work explores the definition and assessment of a quality model for text data. Our literature review indicates that the use of domain-validated representations and methods, and the perspective of model predictions as quality estimations are fundamental for understanding and implementing automatic-metrics. Approaches include domain-driven rule-based, stochastic, and parametric models, clustering to study data groups in terms of quality properties, and high-performance neural networks able to extract task-relevant features. Lightweight pretrained models and comprehensive libraries reduce cost and expertise barriers. To illustrate practical aspects, we train DistilBERT-based classifiers as automatic-metrics for the Comprehensible, Repetition-free, and Grammatical dimensions of the SEAHORSE dataset. Our models exhibit above 88% ROC AUC at a parameter ratio of 0.12 relative to the mT5-SEAHORSE reference. While local metrics can be validated with test-splits of identically distributed data, generalization claims require out-of-distribution evaluations. High-quality data and high-performance models form a virtuous circle: quality data enables better models, which can serve as automatic metrics to quantify quality properties.

Keywords: data quality assessment, questionnaire-based quality, unstructured data, subjective quality, context-aware data quality, quality metrics

Abstract in lingua italiana

La soggettività può contribuire a una comprensione e valutazione olistica dell'idoneità di dati eterogenei. I metodi basati su questionari si dimostrano preziosi per misurare la qualità, diagnosticare i problemi e aumentare la consapevolezza della qualità. I modelli addestrati su set di dati annotati da umani possono apprendere pattern di valutazione della qualità. Sebbene la pratica dipenda dal dominio, i lavori precedenti nel campo della qualità dei dati forniscono basi a livelli di astrazione elevati e intermedi. Questi includono visioni orientate al consumatore, descrizioni delle dimensioni, formulazioni generiche di metriche e modellazione concettuale per separare l'intenzionalità dalle rappresentazioni dei dati. Questo lavoro esplora la definizione e la valutazione di un modello di qualità per i dati testuali. La nostra revisione della letteratura indica che l'uso di rappresentazioni e metodi validati dal dominio e la prospettiva delle predizioni del modello come stime di qualità sono fondamentali per comprendere e implementare metriche automatiche. Gli approcci includono modelli guidati dal dominio basati su regole, stocastici e parametrici, clustering per studiare gruppi di dati in termini di proprietà di qualità e reti neurali ad alte prestazioni in grado di estrarre caratteristiche rilevanti per i compiti. Modelli preaddestrati leggeri e librerie complete riducono i costi e le barriere di competenza. Per illustrare gli aspetti pratici, addestriamo classificatori basati su DistilBERT come metriche automatiche per le dimensioni comprensibili, prive di ripetizioni e grammaticali del dataset SEAHORSE. I nostri modelli mostrano un ROC AUC superiore all'88% con un rapporto di parametri di 0.12 rispetto al riferimento mT5-SEAHORSE. Mentre le metriche locali possono essere validate con test-split di dati distribuiti in modo identico, le affermazioni sulla generalizzazione richiedono valutazioni fuori distribuzione. Dati di alta qualità e modelli ad alte prestazioni formano un circolo virtuoso: dati di qualità consentono modelli migliori, che possono fungere da metriche automatiche per quantificare le proprietà di qualità.

Parole chiave: valutazione qualità dei dati, qualità basata su questionari, dati non strutturati, qualità soggettiva, qualità dei dati contestuale, metriche di qualità

Contents

Abstract	i
Abstract in lingua italiana	iii
Contents	v
Introduction	1
1 State of the art	5
1.1 Context-aware Data Quality	5
1.1.1 Definitions and views of data quality	5
1.1.2 Data quality dimensions	8
1.1.3 Issues of applying and measuring DQ dimensions	9
1.1.4 Diverging Taxonomies and the need for transparency	12
1.2 Subjectivity for Type-agnostic Data Quality	15
1.3 Data Quality in semi-structured Texts	18
1.3.1 Linguistic quality foundations	18
1.3.2 Model-based Text quality assessment	20
1.4 Questionnaire-based Data Quality	24
1.4.1 The Diagnostic Approach (compare assessment methods)	24
1.4.2 The Alternative Approach (compare aggregated results)	28
1.4.3 The Aim Quality IQ assessment methodology (AIMQ)	30
1.4.4 Participative dimension selection and improvement plans	31
1.5 Understanding Quality in complex data objects	33
1.5.1 The Heterogeneous Data Quality Methodology (HDQM)	33
1.5.2 Assessment of unstructured data	35
1.5.3 Approaches to the Analysis of Big Data	39
1.5.4 Clustering for Data Profiling	42
1.6 Data annotation	45

1.6.1	Gold and Silver standard datasets	45
1.6.2	Production of annotated datasets	46
1.6.3	Variability in annotations	47
1.7	Theoretical perspectives of model-based metrics	48
1.7.1	Automatic-metrics and Meta-metrics	48
1.7.2	Model predictions as data quality estimations	49
1.8	Principles of automatic DQ tools	51
2	Methodology	53
2.1	Introduction to the SEAHORSE dataset	54
2.2	Quality dimensions for non-structured datasets	54
2.3	Benchmark performance metrics	57
2.4	Strategies for handling class imbalance	59
2.4.1	Robust metrics for classification	59
2.4.2	Weighted Loss	60
2.5	DeLong test for ROC-AUC statistical difference	61
2.6	Model architecture selection	61
2.6.1	DistilBert	62
2.7	Hyperparameters for training	66
2.7.1	Considerations for BERT architecture models	67
2.7.2	Alternatives to fine-tuning	69
2.7.3	Comparison datasets	70
3	Implementation and results	73
3.1	Data preparation	73
3.2	Model Training	74
3.3	Model evaluation	78
3.3.1	Results for the SEAHORSE Test set	78
3.3.2	External validation	79
4	Conclusions and future developments	87
	Bibliography	91
A	Appendix A	117
A.1	Experimental hardware setting	117

A.2	Exploration phase	117
A.2.1	Hyperparameters	117
A.2.2	Model variants (Highest Val ROC AUC checkpoint)	118
A.3	Feature Extractor and LoRA (Additional epochs)	120
A.4	Learning rate search (Lowest Train Loss checkpoint)	123
A.5	Regularization (Highest Val ROC AUC checkpoint)	124
A.6	Model Architectures	126
A.7	Tuning of decision threshold in MRPC	128
A.8	Feature Extractor model results for MELA	129
B	Appendix B	131
B.1	Subjectivity in Philosophy: Consciousness	131
B.2	Subjectivity and Scientific Objectivity	132
B.3	Subjectivity in Business: Tacit knowledge	133
	List of Figures	135
	List of Tables	137
	Acknowledgements	139

Introduction

The progress of Machine Learning models depends on data that is sufficient in quantity and of high quality. While large datasets can be collected from organizational databases, the web, and specialized devices in fields like the Internet of Things and smart cities, ensuring data quality has motivated significant research efforts to provide data consumers with data fit for their intended uses. Formal theoretical models and actionable methodologies have been developed, and today, data cleaning is a standard step in many machine learning and data mining pipelines. Yet the relationship between data quality and machine learning has been far less explored in the opposite direction, particularly regarding quality measurement.

Modern systems must store and manage increasingly heterogeneous data, and different data types and domains that once required deep domain expertise have been successfully addressed through data-driven approaches. In particular, deep neural networks capable of extracting meaningful representations show state-of-the-art performance across a wide range of tasks. Given the success of Artificial Intelligence in addressing such a broad spectrum of problems, it is natural to consider the potential of quality measurement as a learnable task.

Our literature review identifies several pivotal elements for understanding the use of Machine Learning models as quality measurement tools: the concepts of automatic metrics and meta-metrics [1, 2], the use of classifier output probabilities as quality estimators [3, 4], and the establishment of theoretical requirements that data quality metrics must satisfy to be robust in high-stakes decisions [5]. Notably, generative AI research has devoted considerable attention to the quality evaluation of generated artifacts, but taking a performance-driven and benchmark-centered perspective, leaving the measurement of data quality dimensions largely underexplored.

This work explores the definition and assessment of a quality model for text data. Building on these conceptual foundations, we review methods used in Natural Language Processing for the text classification task, highlighting the central role of meaningful, domain-validated representations in the success of learned models. For example, while bag-of-

words approaches can be effective for tasks such as clustering, they lack the semantic richness achieved by deep neural network representations. Observing machine learning problems through a data quality lens opens promising opportunities for future research.

Data quality is a discipline characterized by careful, holistic attention to data-related issues spanning abstract theoretical foundations, technical aspects, tools and software implementations, and the processes and methodologies for applying this body of knowledge in real-world organizations. In this work, we address the technical aspect, guided by a cost-effective approach. In the field of natural language, we notice that using pretrained models and comprehensive libraries reduces cost and expertise barriers.

The subject dataset of our study is the SEAHORSE summarization dataset. We train DistilBERT-based classifiers as automatic-metrics for the Comprehensible, Repetition-free, and Grammatical quality dimensions. Our results illustrate both the potential and the risks of this approach: lightweight models can achieve high performance on unseen data, yet exposure to a different context or data distribution leads to noticeable performance drops.

We now outline the structure of this work:

Chapter 1 presents a review of data quality concepts and discusses proposed approaches to quality assessment for complex data objects.

Section 1.1 explores the foundational concepts of data quality and data quality dimensions, paying attention to the importance of context-awareness, flexibility, and transparency for confidence.

Section 1.2 explains the role of subjectivity for performing data quality assessment in unstructured data.

Section 1.3 covers the dimensions relevant to textual data quality and discusses model-based approaches used in classification tasks.

Section 1.4 addresses the use of questionnaires in data quality, by reviewing the components and methods of the AIMQ Methodology, as well as subsequent proposals.

Section 1.5 discusses the approaches proposed within the Data Mining and Big and Unstructured Data literature on how to understand and apply Data Quality in complex objects associated with human perception.

Section 1.6 gives an introduction to annotated datasets.

Section 1.7 covers the central concepts of Automatic-metrics and Meta-metrics, used

across domains to conceptualize the task of producing systems to replicate human assessments.

Section 1.8 explores the general use of Automatic Quality Tools in information systems.

Chapter 2 details the methodology we followed for training and evaluating lightweight automatic-metric models using SEAHORSE as a benchmark. We introduce the dataset characteristics and reference results for comparison. Notably, we observe that in quality-aware contexts, efforts to prevent the emergence of deficiencies can result in class imbalance, and explain weighted loss as a mechanism to leverage all available data without introducing bias towards the majority class. For model evaluation, we review performance metrics, highlighting the meaning of ROC-AUC as a metric representing ranking capacity and the use of the DeLong test for model comparison. We also identify datasets for external validation, and discuss the advantages of DistilBERT as a high-performance, low-latency, reduced-size model and the main aspects related to its training.

Chapter 3: presents the execution and results of our experiments. Our results exhibit elevated performance in relation to the reference benchmark values, but unsurprisingly, ranking capacity is reduced in different datasets. We discuss possible causes for such performance drops.

The repository with the code of our experiments can be found in:

<https://github.com/jortizsi/Metrics-for-Measuring-the-Quality-of-Unstructured-Data/>

1 | State of the art

1.1. Context-aware Data Quality

1.1.1. Definitions and views of data quality

The study of data quality is a broad, multifaceted field that involves tools, methodologies, and abstractions, requiring practitioners to adopt a holistic view of systems and information. Beyond only technical and scientific skills to engineer solutions, good data quality requires understanding how information moves and how it is used in organizations, and how stakeholders perceive it. It is widely known by IT professionals, from Software Engineers to Data Scientists, that ignoring data quality can originate in garbage-in, garbage-out issues resulting in low-quality products, services, and insights. In a time when information quantity and diversity continually grow, the success of operational and strategic decision support information systems in competitive organizations relies on high data quality standards.

An introduction to data quality should start by considering what data quality means to most of its practitioners and customers: broadly accepted today is the definition of data quality used in [6], which sees data quality as “fitness for use” by those who use the data, known as data consumers. To fully appreciate the implications of this concise definition, it is imperative to navigate the existing proposals for general quality and data quality conceptualizations and see their common points, subtleties, and nuances that contributed to the advance and construction of the current body of knowledge.

Data quality foundations and practices share a common base with the standard quality management applied in industry, manufacturing, and services. The following characterizations of general [7] and data-focused [8] quality illustrate this clearly: A product is defined as the output of any process, and its quality is achieved through specific built-in features or properties aimed at meeting customer needs, as well as by ensuring that the product is free from deficiencies that impair its fitness for use [7]; this fitness for use, translated for the case of data and data products, means that data must be suitable

to be used and successfully perform different roles in the operational, decision-making, planning, and strategic activities of organizations [8].

A complementary generic definition of quality is the one by [9], who sees quality as conformance to requirements, which are verified in products through measurements. These requirements, as discussed in [7, 9, 10], should align with customer needs and be updated over time, including both adding and removing requirements, and take provisions for a rigorous, precise, and realistic formulation as they describe key artifacts that encode consumer needs and guide development and production activities. In this context-based and customer-oriented view of quality as requirements compliance, those strategies pursuing products that exceed user expectations should carefully consider trade-offs and diminishing returns [9, 10].

In the same direction, for [10], Quality means satisfying the customer's overall needs and expectations regarding anticipated use. However, this general definition of quality does not encourage organizations to settle for the lowest satisfactory data quality standards; instead, they should adopt an integral policy that promotes long-term benefits and flexibility [10]. Applying this perspective, data quality is defined in [10] as fitness for all purposes to be made of the data, considering past and current uses and taking into consideration possible new uses that may arise in the future, hence envisioning quality as inherent and quality management as a proactive vital task.

A relevant work also aiming for high quality data that can be suitable for use in new contexts, and defending the idea that stored past data can create new opportunities and added value is [11]. The authors of [11] establish a comprehensive framework and foundations of data quality with a focus on Open Science and FAIR principles (Findable, Accessible, Interoperable, and Reusable) and review the meaning of data quality within these principles; recognizing the value of the identification in ISO 8000 of data quality as the degree to which a set of inherent characteristics of data fulfill requirements, the authors of [11] also consider, as of central importance, the concept of fitness for purpose: while a dataset is said to be fit-for-use if it is reliable and usable, to be fit-for-purpose further demands that it meets a set of non-functional requirements that bring a targeted utility by fulfilling specific customer needs for a known purpose [11]. Hence, FAIR principles encourage data managers to make provisions that enable possible future data consumers to assess, by means of complete and clear information, whether a dataset can serve their new purposes [11].

While the benefits of enhanced quality are clearly tangible to end consumers, the benefits of quality assurance are often hard to envision by organizations which may feel deterred

given the implied significant costs of adopting and maintaining robust quality management practices [7, 9]. According to [7, 9] these costs are actually advantageous investments because they contribute to reducing the costs of "rework" activities that become necessary after situations of nonconformance [7, 9]. It is fundamental for organizations to study past and current quality problems in order to understand their nature and trace their sources, and be able to design and implement future prevention actions [9]. Further persuasive benefits can be found in [10]: aiming for high data quality can support processes by providing better input materials and facilitating the interaction with data resources and data products, effectively removing obstacles to intellectual growth, allowing data to become a catalyst in learning organizations, helping them to eventually be capable of data-driven management thanks to adequate planning and early long-term strategic choices.

A description encompassing several of the previously discussed elements of data quality can be found in [12], where the authors explain that Data Quality is context dependent: Data is evaluated and considered as of high quality when it proves appropriate for a specific use within the fulfillment of a particular task. Furthermore, they describe that different users of same data can have different needs and perspectives, resulting in a multi-dimensional view that provides the base for the formulation of data quality requirements [12]. In Figure 1.1 these discussed views of data quality are presented sorted by their expected capacity to offer value across different contexts.

**Likely usable again by us
and others for diverse tasks**



- curated/managed to facilitate repurposing by others [11]
- fit for current and future internal use [10, 8]
- fit for original purpose [11] / satisfies effectively needs and expectations [10]
- complies with the original formulated requirements [9, 11]
- fit for original use [6, 7, 12] / has desirable features and is free from defects [7]
- reliable and usable [11]

Not usable now for the task

Figure 1.1: Views of Data quality

1.1.2. Data quality dimensions

A central concept in the body of knowledge of data quality is the data quality *dimension*. A data quality dimension is “a set of data quality attributes that represent a single aspect or construct of data quality” [6, p. 6]. An attribute corresponds to a specific characteristic of quality and serves as a measurement item for higher-level abstractions [6]. The attribute concept is explained in detail by [13] as follows:

The world is made of things that possess properties. A thing can be a composite-made of other things. Properties are represented as attributes which are characteristics assigned to things by humans, depending on purpose and experience. The values of the attributes at any given time comprise the state of the thing [13, p. 89]

The field of data quality devotes substantial effort to the identification and formal definition of quality dimensions, the analysis of their interrelationships, and their systematic organization into hierarchical structures. Such efforts are pivotal for having a complete theoretical understanding of the meaning of quality for data that can help practitioners to perform successful assessment and improvement projects.

One of the most influential dimensions proposals is the one presented in [6], whose development followed an empirical approach. The authors of [6] describe this type of approach:

An empirical approach to data quality analyzes data collected from data consumers to determine the characteristics they use to assess whether data are fit for use in their tasks. (...) it captures the voice of customers. Furthermore, it may reveal characteristics that researchers have not considered (...). The disadvantage is that the correctness or completeness of the results cannot be proven via fundamental principles. [6, p. 7]

Industry leaders and data consumers participated in the identification of relevant attributes, they were asked for importance ratings of dimensions, and their views on how to sort dimensions to organize categories were considered [6]. The resulting consumer centric taxonomy of quality dimensions into four groups, is explained briefly in [14]:

Intrinsic IQ implies that information has quality in its own right. Contextual IQ highlights the requirement that IQ must be considered within the context of the task at hand; it must be relevant, timely, complete, and appropriate in terms of amount, so as to add value. Representational and accessibility IQ emphasize the importance of computer systems that store and provide access to information; that is, the system must present information in such a way that it is interpretable, easy to understand, easy to manipulate, and is represented concisely and consistently; also, the system must be accessible but secure. [14, p. 135]

A notable advantage of the empirical approach is that building a theory of dimensions focused on listening to consumers ensures that the resulting taxonomy is useful and actionable; grounding it in past practice serves as an implicit validation criterion.

In contrast to their empirical method, the authors of [6] describe two alternative approaches: (1) The intuitive approach, based on expert judgments, and where the sum of views and agreement on what is important leads to the adoption of lists of dimensions, and (2) the theoretical approach, that sees data as the product of a manufacturing process and defines quality in terms of deficiencies [6]. An example of the theoretical type is the ontological approach [6], whose rationale is that:

The quality of data depends on the design and production processes involved in generating the data. To design for better quality, it is necessary first to understand what quality means and how it is measured. [13, p. 89]

The ontological approach understands quality under rigorous modeling and representation considerations that draw comparisons between the information system and the real world [6, 13]. It also differentiates an internal view that includes aspects of design, construction, and operation, from an external view that focuses on purpose and reduces the information system to a deployed black box that is used to create value [13]. Similarly, dimensions can be seen as data or system-related [13]. The ontological approach proposes four intrinsic data quality dimensions (correctness, unambiguous, completeness, and meaningfulness) and links them to deficiencies caused by design and operation failures [14, 13].

Extensive academic work has been published introducing lists of data quality dimensions, hierarchical structures, and practical recommendations for their use within data quality activities. A comprehensive overview of literature on data quality classifications is [15], while a detailed survey of data quality methodologies, including the specific dimensions they consider, can be found in [16].

A noteworthy additional contribution to data quality dimensions, proposed by [17], is the ethics cluster consisting of fairness, transparency, diversity, and data protection .

1.1.3. Issues of applying and measuring DQ dimensions

Numerous views and proposals on how data quality should be understood and how dimensions could be organized coexist, some of them aimed at particular domains or use cases, while others were created from a general perspective [15]. On this topic, [18, 19] present significant philosophical-conceptual contributions to explain why no single formal model of dimensions has prevailed over other proposals.

In the view of [18, 19], an important practical issue of using the high-level definition of data quality as “fitness for use” is that it neither specifies nor explains how to perform measures. Collections of dimensions and metrics have been proposed in the literature in rather general terms, but data quality practitioners quite often face situations where these allegedly “general” dimensions cannot be applied directly [18]. When a practitioner’s work follows a framework or list of relevant dimensions, but the actual use of such concepts brings no advantages, ground exists to support the idea that these dimensions were not truly use-independent, and in such cases, the professional assumes the responsibility of bridging a gap by translating abstract initial notions into operational definitions [18, 19]. Consequently, when professionals apply a methodology listing a fixed set of dimensions to be considered during the quality requirements elicitation, this often provides theoretical understanding but limited practical guidance [19].

A strict view of quality dimensions further proves inconvenient in cases where specialist knowledge is sufficient and obtains no gains from using abstract models [19]: when domain experts know beforehand what needs to be measured and in which terms to explain it to data quality professionals, developing an understanding of the case-specific concepts within the DQ dimensions used in a methodology can be challenging. Indeed, numerous domain-specific and context-aware metrics have been introduced and extensively validated in real-life practice because these metrics are always designed assuming a typical fixed structure of data and grounded in domain knowledge [18], being effective at capturing the underlying business rules and data semantics [19]. Unfortunately, while these domain-specific metrics are of great practical value, their constraints and assumptions make them too limited to offer insights on how to restate upper level definitions of dimensions, leaving unsolved the existence of a gap between abstract concepts and operational definitions [19].

One significant flaw of the view that data quality can be defined in a general-intrinsic way is an associated defense of the existence and validity of “absolute” objective measures and scales [18]. This position advocates for the use of general metrics suitable to evaluate data quality across different contexts, assigning to data quality practitioners the responsibility to identify which level of quality is “satisfactory” to address each specific problem and situation [18]. While intuitive, such a threshold-based definition of quality clearly rests on a simplistic assumption that “more is always better” and fails to acknowledge the complex dynamics that a multidimensional context-aware analysis could better handle, for example, related to quality-cost trade-offs and the local meaning of fitness for purpose [18]. For [18], an important problem lies in the fact that dimensions are determined, too, by relevance to the intended purpose. This relevance problem is illustrated for accuracy as follows: “the only completely accurate model of the system is the system itself. But (...)

you need to extract information about some aspects (...), and store it in a database" [18, pp. 284-285]. Similarly, [8] notes that data labeled as of the highest intrinsic accuracy is useless if data are in a format, structure, or terms that do not match the particular use case.

In this perspective of highly contextual data quality, [18] considers that what is really important is having awareness and transparency of the real state of quality by knowing the detailed characteristics and deficiencies of data. Quality metrics must be representative rather than artificially high, and the results of measurement activities must be presented to be easily understood by the data consumers who rely on the rigorous work of data quality practitioners, in particular for getting high quality data from assurance and improvement efforts, and getting precise estimations of the state of data quality from assessments [18]. This transparency is fundamental because well-informed decision making requires not only adequate data content, but also proper estimations of quality to build up a proper confidence that can avoid risks without missing opportunities [18].

In this view, metrics are seen as limited indicators or estimators useful to characterize a dimension [18, 19], and they are the input for an "interpretation" process that is dependent on purpose [18]. This estimation is done through a set of reasonability checks that use the target dataset together with trusted reference sources, and in order to promote high-confidence results, it becomes crucial to combine a variety of indicators [19] and prioritize any applicable specialized purpose-dependent metrics [18].

Noticing the difficulties associated to using fixed dimensions and the importance of using local metrics, [19] argues that a practical approach can offer better support to DQ practitioners than only orienting work following dimension sets. In this view, [19] proposes *Quality View patterns*, which are software components that implement the generic quality assessment patterns that characterize specific domains, working with specific types of data and using specific definitions of quality. A quality view pattern integrates transformation modules and displays an interface that receives an input dataset and produces one or more output datasets [19]. These transformations include assessment operations such as adding columns with DQ measures, and improvement operations, such as filtering data to satisfy some criterion or applying cleaning procedures [19]. The modular, domain adapted, and interface-based nature of Quality View Patterns makes them highly reusable [19].

These issues show that data quality metrics, methodologies and technical artifacts should be seen as instrumental elements for attaining fitness for purpose. Maximizing a metric without considering whether its application suits specific contexts is likely to consume resources and bring little to no benefits. This issue of high metrics performance that are

not representative of the real data quality state is one of the leading motivations of AIMQ [14](for a discussion of AIMQ refer to Section 1.4).

1.1.4. Diverging Taxonomies and the need for transparency

A strong case against the convergence of quality dimensions classifications is presented in [18], which argues that a root cause impeding convergence is that both dimensions and categories are presented in a natural way through intuitive but imprecise meaningful words, and therefore parts of dimensions can easily fall under different categories, in such a way that the results of efforts to build a one-to-one disjoint hierarchy are not consistent. Examples of dimensions whose classification is challenging due to including a wide variety of quality aspects are timeliness, completeness, and accuracy [19, 18].

Taking these issues into consideration, the author of [18] argues that to overcome the limitations of traditional strict hierarchical taxonomies it is convenient to use a relational model that can solve ambiguity by “mapping elements (...) for IQ improvement onto a space designed to represent what they are properties of.” [18, p. 299].

This proposal aims for clarity on the scope and granularity at which different assessment and improvement activities operate, making explicit what is being described through these reported dimensions and metrics — in other words, what they are being defined on. More precisely, the author explains that:

there are two kinds of things that might be bearers of properties (...): 1. Parts of the information system (...), there may be several that are useful to distinguish. 2. Relations between the information system and something external to it, its ‘context’ (...): (a) the relation (deployment) between the information system and the purpose of the user, and, (b) the relation (reference) between the information system and the external world [18, pp. 294-295].

Under such characterization, [18] revisits the traditional empirical hierarchical model and maps the intrinsic and contextual categories to items 1 and 2 respectively, noting that the remaining two traditional categories (representational and accessibility IQ) correspond to “features of the relationship between the information and the user” [18, p. 299].

Moreover, [18] explains that adopting an approach based on transparency and clarity about what is being evaluated endows computed metrics that would otherwise be mere proxies with meaning and further demonstrates through enumeration that, if ambiguity is not resolved, the levels of information a user can consider are numerous: (1) datum (2) object instance (3) object class (4) database (5) data source (6) information system as

static structure (7) information system dynamics accounting for improvement measures through time, and (8) relations associated with purpose between the information system and data consumers.

Two sources aligned with this view are [20, 21], who present solid theoretical foundations, and give practical but still flexible and high level guidance. Notably, contributions of works such as [21, 20, 22] introduce a middle ground between the multifaceted conceptual foundations of data quality (which draw from fields such as quality management, statistics, software engineering, data mining, knowledge representation and information systems management [20]) by clearly defining conceptual generic definitions of dimensions considered relevant across organizations, (including dimensions such as accuracy, completeness, consistency, believability, timeliness, accessibility and appropriate amount of data), together with several functional forms such as ratio, min, max and weighted average, that are available to define quantitative metrics needed for their practical use [21, 20, 22].

Even starting with these two important theoretical and practical foundational elements, working in data quality requires practitioners to develop skills for reading each organization context and properly defining the best intervention actions. Indeed, [21] notes that:

Often, the most difficult task is precisely defining a dimension, or the aspect of a dimension that relates to the company's specific application. Formulating the metric is straightforward once this task is complete. [21, p. 213]

A compelling analysis that gives central importance to transparency, context-awareness and a clear, intuitive but actionable vision of quality, as promoted in [18, 21, 20], can be found in [20] within the discussion of accuracy as driven by the quality concept of "free of error":

A concise description of what constitutes a data unit is required. Is it a field, a record, a table? Further, what exactly constitutes an error? The sense of the precision of the measurement must be specified. Quite possibly, in one case an incorrect character in a text string is tolerable, whereas in another circumstance all characters must be correct. These are distinctions that make the metric context-dependent and that demand unambiguous definition. We cannot stress enough the importance of developing definitions that span databases and stakeholders. [20, p. 55-56]

The divergence of dimension taxonomies must not be considered as a flaw in the data quality discipline. Quite the opposite, the availability of general and domain oriented

methodologies and frameworks shows its active efforts and success to understand the meaning of quality in different contexts and implement suitable management actions. Data quality success lies in giving each practitioner the concepts and tools for their work; a solution must be provided for each context, but each solution does not necessarily have to be applicable in different contexts to be considered a worthy asset of the discipline.

This highlights data quality — typically understood as fitness for use [6] and better described as fitness for purpose [11] — as multidimensional and context-aware. The characteristics and goals of organizations and consumers must be considered to represent different fitness definitions, problems, and optimization objectives calling for different methods for assessment and improvement.

1.2. Subjectivity for Type-agnostic Data Quality

Humans are both information processors and information creators [23, 24, 25, 26]. This ability to read and interpret reality can be used for data quality assessment in two ways:

1. As *recorded perceptions*, corresponding to the execution of dataset annotation projects and psychological standardized experiments intended to assign to data objects properties and features that are difficult to measure directly [27, 28, 29, 30]. In the case of generative AI, annotations are useful to evaluate system generated outputs [31]. From a broader data quality perspective, annotations are useful when they represent properties that determine fitness for use. Data annotation is also strongly related to the creation of metadata for digital resources, which has central importance for content management and data repurposing [32, 33, 34].

These recorded perceptions assume that individuals performing annotations have *expertise* [35]. This expertise is multifaceted [35]: it can be limited to the skill for translating sensory information to data following commissioner provided criteria, or it can rely on individuals applying domain-specific knowledge. Moreover, lived experiences and familiarity to different contexts constitute elements of expertise [35, 30]. This approach proves fundamental for problems characterized by a lack of formal scales or involving significant subjectivity given the nonexistence of a single reference ground truth or ideal [28, 30].

2. As *collected judgments*, representing individual views of the meaning, state, and importance of data quality within a specific context which are explored through interviews and questionnaires, and prove relevant for data quality requirements elicitation and for assessing the levels of quality dimensions by means of psychometric methods that allow the quantification of these judgments [14, 12, 16].

Learning from the perspectives of different stakeholders is a crucial element for addressing the challenge of interpreting, adapting, and applying Data Quality concepts and practices in specific contexts, to properly ensure the ideal of fitness for purpose [18, 19, 20, 14]. This approach measures the contextual value and utility of data for stakeholders [18, 19, 20, 14].

A notable work illustrating the subjective assessment of unstructured data is [28], whose authors explain that the systematic evaluation of videos and images can be carried out through psychological experiments presenting the data as stimuli in standardized settings and collecting evaluations using validated scales. Central to these procedures is the pos-

sibility to deliver either direct isolated evaluations, or evaluations supported by reference elements requiring comparisons among multiple displayed samples [28].

While traditional metrics rely on the production of definite verifiable measurements, which can be determined statically, or in relation to the performance achieved in a task (*a posteriori*), human evaluation is conducted intuitively through perception [30].

A noteworthy analogy on expertise is drawn in [30] using the Open and Closed World Assumptions of knowledge representation. In information systems, data can be seen in isolation or considered in relation to other elements. The system stores representations of objects and can integrate representations of other elements from trusted sources, or it may be aware of their existence in the real world [30]. In a similar manner, human evaluators use internal references shaped by personal experiences and knowledge [30].

A last matter involving human subjectivity and context-aware data-quality is modeling. Different factors and choices such as the level of abstraction, representation details, and the mechanisms provided for the interaction and visualization of data, affect how information is perceived, assessed and used [30].

Table 1.2 presents several implications of subjectivity in data quality. A more general characterization is outlined in Appendix B.

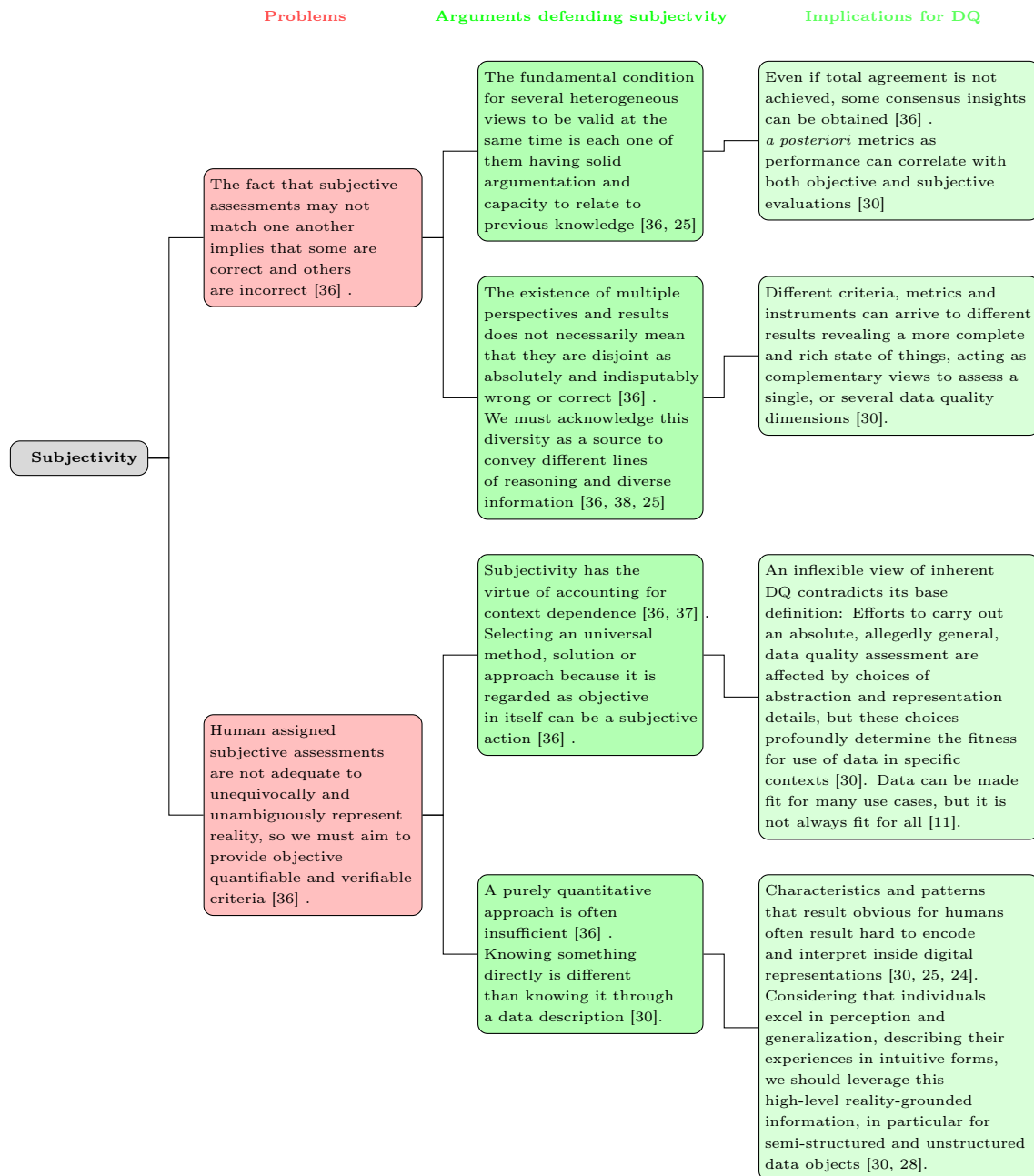


Figure 1.2: Problems of subjectivity

1.3. Data Quality in semi-structured Texts

1.3.1. Linguistic quality foundations

Natural language quality can be considered at different levels of granularity: word, sentence, text, and collection of texts [39]. This type of data shows unique characteristics that make the definition of intuitive quality properties straightforward but that at the same time require an extensive set of specific indicators to properly capture domain phenomena. A fundamental remark regarding text data, found in [39], is that although the data types used to represent text can store arbitrary sequences of characters, real world text data appears as semi-structured because even if texts cannot be decomposed to fit a fixed form of relational table, they present linguistic structures. According to [39], this structures are described in terms of five elements:

sentences (...) are characterized by a lexicon that considers words, a grammar that establishes a set of structural rules (...), and a semantics that provides the meaning (...). Further, rhetoric is (...) literary effect, and pragmatics is the way in which context contributes to meaning. (...) the term syntax is also used to refer to the rules [39, p. 62]

Natural language quality is founded on the body of knowledge of linguistics and as representative of human communication it is highly context dependent and multifaceted [39]. A noteworthy overview of general text quality dimensions can be found in [39], for which we present a summary in Table 1.1. Similar to structured data, dimensions can be decomposed into quality properties and observable phenomena to asses dimension constructs [13]. Notably, the assessment of the high level dimensions of comprehensibility and coherence proves complex given an intersection of factors such as communication goals, use of emotions and figurative language, temporal and spatial scope, or starting assumptions for elaborating new ideas [39]. For these dimensions the availability of human annotated datasets is a valuable element for validation of theoretical metrics [40].

Cluster	Dimension	Aspect measured	Implementation examples
Accuracy	<i>Lexical accuracy</i>	words Belong to a defined vocabulary	• Average of a closeness metric considering each text word and the considered vocabulary
	<i>Syntactic accuracy</i>	Satisfaction of language specific rules	• Italian grammar
Readability	<i>Readability</i>	Relative ease to read a text as estimated using statistical correlations determined by aspects such as word and sentence lengths, and complexity of individual words.	• Gunning Fog index using average words per sentence and relative frequency of complex words. • Automated readability index (ARI) using average characters per word and mean words per sentence.
	<i>Comprehensibility</i>	Target audience is able to understand the conveyed ideas and facts. In addition to inherent text features, writing must account for the prior knowledge and context of readers.	• Coh-Metrix Tool, providing indexes for linguistic features determining comprehensibility such as ambiguity, abstractness, and syntactic complexity.
Consistency	<i>Cohesion</i>	Use of linguistic mechanisms to connect different text units, establishing relations that aim at building a single representation of concepts.	• Local coreference, defined for a text consisting of n sentences as the ratio of adjacent sentences connected by a common argument noun divided by the number of total adjacent sentence pairs. A value of one indicates that the text is a continuous chain of linked sentences. • Global coreference, defined as the count of unrestricted sentence pairs that are linked by a reference divided by the number of total possible sentence pairs.
	<i>Coherence</i>	The reader is able to build a coherent mental model by combining text-provided explicit and implicit relationships, and personal skills and knowledge of the world. Requires composition strategies to guide readers towards author created structures of semantic constituents.	• Entity Grids, based on the weighting of entities by salience (importance in text estimated from frequency) and the use of a matrix encoding the syntactic role that each entity has in text sentences[40, 41]. A statistical transition model is used to model coreference patterns that are used to rank coherence [40, 41]. • Latent Semantic Analysis (LSA), which starts representing the sentences or units of texts in a term matrix and then reduces dimensionality using Singular Value Decomposition (SVD) [42]. The obtained representations of the sentences are semantically meaningful and can be used to compute estimations of coreference in terms of cosine similarity [42].
Accessibility	<i>Linguistic accessibility</i>	Absence of difficult words	• Given a vocabulary of words considered as basic, assume words not in the basic set as difficult, and compute the proportion of difficult words to total words.
	<i>Cultural accessibility</i>	Skills and capabilities needed for understanding the text.	• Percentage of population proficient in the publication language.

Table 1.1: Text quality dimensions (based on [39])

Designing linguistic-oriented methods, such as those appearing in Table 1.1 requires significant effort for expressing domain knowledge through statistical and computational methods that can be directly interpretable [39]. One particular limitation of many linguistic metrics is their need to consider the rules of each language, for example the comprehen-

sive english Coh-metrix Toolkit has motivated similar software implementations for other languages [39, 43, 44, 45]. However, works as [46] have explored multilingual grammar. Also notable, is the need for users to understand the meaning of indicators and identify a set that is representative of the quality aspects relevant for their goals [39, 45].

An important remark for text data is that in addition to inherent linguistic-type quality properties, very often key quality aspects are of referential nature [39, 47]. For example, [39] explains that laws are understood in relation to other laws and a stored snapshot of a law can vary in quality given real world judicial-system dynamics that can make it amended, superseded or retired. Similarly, tasks as summarization need referential quality definitions to assess summary quality in relation to original documents [48, 31].

Different tasks and sub-fields of NLP use specific quality dimensions. For example, [31, 48] describe several crucial quality aspects for the summarization task: relevance, corresponding to the inclusion of important information and the absence of redundancies, factuality, corresponding to the entailment of statements in the summary from the original source, and adequacy, representing meaning preservation such that information is lossless, undistorted, and unexpanded (often associated with the absence of system hallucinations). In this context of summarization, a useful terminology distinction is drawn with the categories of *language quality* dimensions, for text as seen in itself, and *content quality* dimensions, assessing a summary as seen in relation to the original document [49].

In many contexts, ethical aspects such as misinformation [50] and toxicity [51] are also relevant.

1.3.2. Model-based Text quality assessment

As an alternative to linguistic-guided metrics, a data driven approach can be used to train models better capable of managing semantics and showing state of the art performance in replicating human judgments [52, 53]. In general, the performance of NLP models is evaluated using benchmarks consisting of datasets, evaluation metrics and aggregation rules to compute comparable performance scores [54]. Benchmarks can measure performance in specific tasks or include tests for several tasks in order to inquire for wider model capabilities as Natural Language Understanding [55, 54]. A comprehensive review of modern NLP benchmarks is [56].

A good benchmark should be useful for both model developers and model users [57]. They must provide traceable data, be interpretable by defining covered aspects and use cases, and be representative of real world workloads [57]. Moreover, benchmarks must provide

aid for studying causality relations, anticipate risks and avoid biases [56, 58], and during their life cycle they should evolve to maintain utility and be a meeting point for discussing and collecting new knowledge, before being finally deprecated [57, 56].

Four model based approaches for text evaluation are identified by [48]: (1) Unsupervised Matching, based on measuring semantic equivalence between the text and a reference source, (2) Supervised Regression, which uses models trained to predict human judgments, (3) Supervised Ranking, using scoring functions, and (4) Text Generation, based on the idea that generative models when computing their outputs, assign higher probabilities for higher quality responses.

Different tasks and benchmarks can privilege different approaches. In the context of Data Quality, the idea of fitness for use easily aligns with the approach of supervised regression.

Following the approach described in [3, 4] for using class membership probability predictions as quality metrics [1], we now discuss the classification topic for texts. An in-depth comparative review on text classification research is [53], where three families of methods are identified. The first family represents texts as Bags-of-Words, and uses this vector representation to classify texts using approaches such as Support Vector Machines and simple Multi-Layer Perceptrons [53]. The second family decomposes texts into tokens that are used as the input representation in models, notably in transformers [53]. The third family includes those methods based on hierarchies and those that represent texts as graphs composed of nodes and edges that model relationships among information elements, which can be processed by dedicated Graph Neural Networks (GNNs) [59, 53]. Since [53] identifies the Transformer Deep Neural Network architecture as the state-of-the-art approach for text classification, we further explore this topic.

Transformers in Text Classification

The Transformer architecture constitutes one of the modern foundations to work with data such as text and images, extracting meaningful features and achieving high performance in downstream tasks [60]. Transformers are based on attention mechanisms, and they are typically released as pre-trained models that optimize a self-supervised training objective, which does not use explicit supervised examples with targets, but instead relies on consuming large quantities of data, which in the case of Natural Language Processing correspond to large text corpora [53]. For example, BERT pretraining optimizes two self-supervised objectives: the correct prediction of masked tokens within a sentence (Masked Language Modeling) and the prediction of whether a sentence pair occurs sequentially (Next Sentence Prediction) [61].

Regarding the commonly used distinction of Small Language Models (SMLs) and Large Language Models (LLMs), the authors of [53] note that better than considering the parameter count, the pre-training objective is more significant and straightforward for applying such categorization. They explain that:

The encoder-only and encoder-decoder SLMs (e. g., BERT and T5) are based on masked language modeling (...) while the decoder-only LLMs (e. g., GPTs) are pre-trained using the causal language modeling objective. [53, p. 4-5]

The authors of [53] identify three transformer architectures used in language classification: encoder-only, encoder-decoder, and decoder-only. These model architectures are explained by [53] as follows :

BERT and its follow-up approaches are encoder-only language models and are trained with the masked-language modeling objective. Their main purpose is encoding the text and performing a classification or regression task with a dedicated module (e.g., a linear output layer) on top of the encoder. The output layer is trained from scratch (i. e., from random initialization) during fine-tuning.

Encoder-decoder language models such as T5 cast all downstream tasks (...) into a text-to-text framework with task-specific prompt prefixes. The rationale is that this practice enables multi-task fine-tuning with the aim of similar tasks improving each other. Although these models generate text, they can also be used for classification tasks by interpreting generated tokens as class labels (...).

Decoder-only language models such as GPT-3 are trained with a left-to-right, i.e., causal language modeling, objective. This makes decoder-only models most suitable for text generation (...) and can also be used to carry out other downstream tasks, including text classification (...), specifying the task in natural language and providing a few examples in the prompt. This practice, known as in-context learning, does not require updating the model. [53, p. 10-11]

Even if decoder-only generative LLMs can be finetuned, in-context learning is often preferred because the model size makes additional training expensive [53]. Nonetheless, the recent advances for lighter models, such as DeepSeek provide a solid and more accessible entry to explore decoder-only models [62, 63].

The authors conclude that encoder-only models remain state-of-the-art in text classification [53]. This architecture is purely discriminative in the sense that it is not related to generating a token or response, but is built instead to provide a direct numerical output result using the encoder's semantically meaningful representation of the text with a problem-specific trained classifier head [53].

Limits of the generative approach

While encoder-only models prove superior in classification, generative models have advantages that motivate a further review of issues and opportunities. The Text generation approach is particularly interesting as it is highly flexible and does not require human annotations or additional training before deployment [48]. A first application of text quality based on generative models is the perplexity (PPL) measure, which considers the probability that an evaluator generative model would generate a considered text [64]. Nevertheless, such a measure has been proven to be unreliable, showing high variance for minor edits [64]. Better results have been obtained in referential contexts, for example in BARTScore, based on the encoder-decoder architecture [48]. BARTScore is used to assess quality in conditional text generation tasks such as summarization and translation, where a system produces as output a "hypothesis" text and quality is defined in relation to a given source text and optional gold reference outputs [48].

Also the use of LLMs such as GPT-4 and Llama 3 has been explored for text assessment within the task of generated response evaluation, but significant issues have been identified in research such as ranking inconsistencies arising from ordering choices, and the fact that only the largest variants show human aligned performance [65, 66]. The authors of [65, 66] note that the use of LLMs for assessment is viable but requires a methodical approach to yield reliable results.

1.4. Questionnaire-based Data Quality

An incremental line of research that considers individual views can be found in [20, 21, 67, 14], whose authors introduce the significant contributions of two comparative approaches, and the Aim Quality (AIMQ) Information Quality methodology, which constitute a solid foundation that organizations can use for incorporating subjectivity into data quality assessment and improvement initiatives. We review next these concepts, all sharing as practical foundation the application of a survey.

1.4.1. The Diagnostic Approach (compare assessment methods)

Step 1: Application of subjective and objective measurements

The organization considers and defines which data quality dimensions are the most valuable and then starts procedures to assess their quality using the following two methods that although being different in nature, should be expected to show consistent results [20]: (1) objective assessment through quantitative metrics developed ad-hoc or selected from academy or industry validated standards [20, 21], and (2) subjective assessment through the application of the Information Quality Assessment survey (IQA) [20]. This IQA survey is an instrument designed to capture “subjective assessments of data quality and the respondents’ knowledge of the data quality processes, programs, and tools that are in place” [20, p. 31]. A benefit of conducting this survey is that beyond serving for assessment, it also promotes quality awareness through its participatory dynamic [20].

Perceptions can differ significantly across stakeholder groups (data collectors, data custodians, and data consumers), based upon their personal concrete experience, knowledge, interactions and needs [20, 21]. In order to translate these diverse subjective views into quantitative data, the IQA survey uses a zero-to-ten Likert-type scale [20]. This scale is supported and has been validated in psychometric literature [68], defined as:

the field of study concerned with the theories and methods of psychological measurement (...) considered the engineering-level discipline associated with questionnaire and survey development [69, p. 140]

The scale is described in [68] as follows:

The Likert scale is a widely utilized rating scale for measuring unobservable constructs, such as attitudes, opinions, or perceptions through a structured response format. It consists of a series of statements or questions, each accompanied by a symmetric range of response options that allow respondents to indicate their level of agreement or disagreement. (...) numerous adaptations (...) measure frequency,

importance, satisfaction, and other perceptual dimensions. A Likert item refers to a single question with a Likert-type response format, while a Likert-type scale is a composite measure comprising several (typically 4 to 10) related Likert items designed to assess a broader construct. [68, pp. 1-2]

The creation of Likert-type questionnaires can follow guidelines such as those presented in [68], which we summarize in Table 1.2, however, the creation of a robust Likert-type instrument is a task requiring pilot studies and statistical validation, using indicators such as the Cronbach alpha to verify the consistency and reliability of different items designed to measure the same constructs [14, 70, 71]. The IQA survey has the advantage of having taken into consideration such a validation analysis [14].

Aspect	Guidelines
Instruction	Presented before an item or group of items, it explains the scaling system, its direction and the meaning of the options that the participant can select.
Layout	When designing printed or software delivered instruments, take into consideration that presenting the options in horizontal layout favors intermediate options because a wide visual range conveys the idea that side options are extreme. A vertical format can help reduce this central tendency bias to get sparse responses.
Items (questions or statements)	1. Be concise. Complex questions are harder to answer and may affect the precision of results. If an item deals with multiple ideas, then it is better to indagate the same topics through separated items. 2. Use neutral wording. Text must not give hints into the existence of an expected or ideal answer. If an item suggests that an option is considered socially desirable, best performing, or hypothesized as likely, respondents may be biased and results will not be representative of reality. 3. Include reverse written items. This compensates for acquiescence bias (a bias to agree and select positive answers). It can also motivate participant thoughtful reading. 4. Avoid double negative sentences. 5. Keep symmetry by presenting an equal number of positive and negative options. 6. Avoid including by default a neutral option. Unless an intermediate position has a significant meaning within the research or data utilization goals, aim for exclusively positive and negative options. This moves participants to take a side and impedes them to use the neutral option as a way to hide their true perceptions.
Composite measures (Scores)	Higher level constructs are evaluated using lists of several items, therefore the overall expected assessment error is reduced.
Missing data	Imputation can be applied. However, for cases where data is missing not at random it becomes fundamental to use robust methods, for example trained models

Table 1.2: General guidelines presented in [16] for Likert-scale based questionnaires

An example of a complete software-delivered IQA survey can be found in [20]. Similarly, [14] contains the valuable set of questions formulated to evaluate recognized standard data quality dimensions within the AIMQ methodology. Table 1.3 presents an overview of the IQA survey instrument based on these two sources. Topics are explored using Likert-response questions except where otherwise specified.

	Section focus	Topics
1	Characteristics of information	• Types of data (multiple choice questions). • Complexity of activities. • Position: Department and role. (multiple choice questions).
2	Dimensional quality	• Items to evaluate data quality dimensions. The AIMQ questionnaire consists of 65 questions implementing two strategies to support reliability: – Use 4 or 5 items per dimension. – Use of both positively phrased and reverse coded questions.
3	Respondent's knowledge regarding the organization context and environment	• Application of methodologies, practices, programs, standards, processes and tools. • Clarity of Responsibilities in DQ. • Participation in DQ activities. • Perceived competence to address DQ.
4	Background information	• Judgement on the value of information and views on past encounters with DQ problems (open questions). • Education and work experience (multiple choice questions).
5	Respondent's knowledge regarding Information collection	• Provenance: creators, sources, gathering process. • Awareness of problems. • Skill to identify problems. • Standard procedures. • Skill to solve common problems. • Skill to investigate new problems.
6	Respondent's knowledge regarding Information storage	• Processes to store and maintain information. • Responsible group. • Tools. • Comprehension of accessibility challenges. • Awareness of problems. • Skill to identify problems. • Standard procedures. • Skill to solve common problems. • Location of specific information. • Formats.
7	Respondent's knowledge regarding Information use	• Who are the users of information. • Skill to access information. • How to use information. • Which tasks use specific information. • Awareness of problems. • Skill to identify problems. • Standard procedures. • Skill to solve common problems. • Skill to investigate new problems.
8	Importance ratings	• Quality dimensions importance ratings in relation to the tasks executed by the respondent.

Table 1.3: Overview of the Information Quality Assessment (IQA) survey [20, 14]

Step 2: Comparative analysis of results

After the metric measurements and the IQA Survey answers have been collected, the observed agreements and conflicts in the results of the two evaluation methods are studied [20]. By classifying both data quality estimations as “High” or “Low” a 2x2 matrix is defined, whose quadrants help to guide the analysis [20]. Agreement as “High quality” is the goal state, and agreement as “Low quality” reveals a need to initiate improvement efforts [20]. Cases falling in mismatched quadrants call for an investigation of the causes leading to disagreement, taking into consideration the existence of any quality assurance mechanisms in the organization and the awareness that stakeholders have of their actual implementation [20].

The authors of [20] mention two insightful root-cause possibilities that are worth considering as likely causes for divergence: (A) When “High quality” valued metrics meet “Low quality” survey assessments, this can indicate an extended lack of confidence from stakeholders due to past data quality issues, showing wary views that persist even after significant improvements were achieved [20, 21]. In such cases training for successful data usage is an ideal strategy for the organization to overcome distrust and leverage its owned high-quality data [21]. And, (B) It can happen that the metrics are not aligned with the aspects that participants considered for delivering their subjective evaluations, and hence, assessments intended to estimate the quality of a specific dimension actually provide a measure of different things depending on the method [20].

Step 3: Data Improvement

By completing the comparative analysis, the organization is able to understand the initial divergences seen in results, and becomes aware of the ongoing data quality problems, their specific sources, and the ways in which different tasks are affected by reduced data quality [20]. At this point DQ practitioners can understand and diagnose DQ issues and have the responsibility to determine actions aimed at improvement [20]. While this comprehensive characterization of data certainly facilitates the selection and execution of improvement actions, further guidelines on how to proceed would be of significant value [72].

1.4.2. The Alternative Approach (compare aggregated results)

Investigations on the nature of data quality issues can consider qualitative insights from open questions, and on-site activities done by data quality practitioners, but a more straightforward option to study the results of the IQA survey is to perform aggregation operations at different levels of granularity [20, 21, 14]. These aggregations consider data as separated by organization and role, making visible the phenomenon of gaps [20, 21, 14]. Two types of gap are interconnected; If a company underperforms on quality metrics when compared to other competitive organizations, an investigation of which factors produce this situation can start by checking how different roles view the quality of data [20, 21, 14].

Benchmark Gap

Benchmark gap analysis compares the quality scores obtained by an organization after delivering the IQA survey with the results shared by other organizations, paying particular attention to those considered references for upholding best data quality practices [20, 21, 14]. Such comparisons should consider organizations belonging to the same domain or part of a related industry, and having a similar size and complexity [20, 21, 14].

The authors of [20, 21, 14] explain that this analysis considers the percentile plot of the scores distributions, focusing on three elements : (1) Gap size or distance (measured as the separating vertical line for a fixed x value), (2) gap area (defined between intervals), and (3) gap location (corresponding to the y-axis range defined by a superior and inferior value). These three elements are complementary; being situated below a reference organization can be described solely by distances, bringing evidence that improvement is possible, but understanding whether this improvement efforts constitute an urgency or an opportunity, and describing the exact patterns of perceptions, and confidence for high and low quality data, require considering location and area as well [20, 21, 14].

Quantile or Percentile Plot: In its general formulation, this type of plot can be drawn by positioning and connecting the points described by all the sorted observations of a variable and the quantile or percentile each represents [73]. The plot is useful to describe and compare data distributions, as it displays the range of values for the observations of each variable, and the overall magnitude of variation from minimum to maximum, and it also includes key values such as the median and quartiles, and facilitates spotting patterns, for example, distributions that are identical except for a shift [73]. Analysis supported by these plots can also be done by computing representative and evenly spaced quantiles [20, 21, 14]. In the benchmark gap, one dimension is considered, with the x-axis assigned to percentiles and the y-axis to quality score values [20, 21, 14].

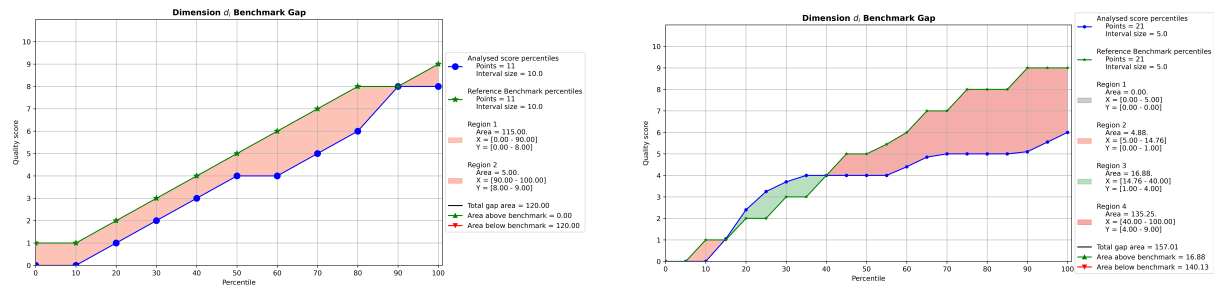


Figure 1.3: Benchmark Gaps (based on [20, 21, 14])

Role Gap

Before designing and executing plans for quality improvement, it is necessary to know with certainty the current state, requiring a consensus among stakeholders [20, 21, 14]. The analysis of role gaps has as its goal verifying whether the evaluation of quality by Information Systems professionals is aligned with that of data consumers [20, 21, 14]. The role gap plot for one dimension has several organizations on the x-axis, and for each one of them a vertical segment goes from the value corresponding to the mean score of the dimension assigned by data consumers to the mean score by IS professionals [20, 21, 14]. The gap segments are described by their size or length, their location in terms of y-axis coordinates, and their direction [20, 21, 14]. A positive gap means that the IS professionals ignore quality problems experienced by consumers, overestimating data quality [20, 21, 14].

A complementary visualization can be obtained using a profile plot displaying the evaluation of several dimensions across stakeholder categories within a single organization [20, 21]

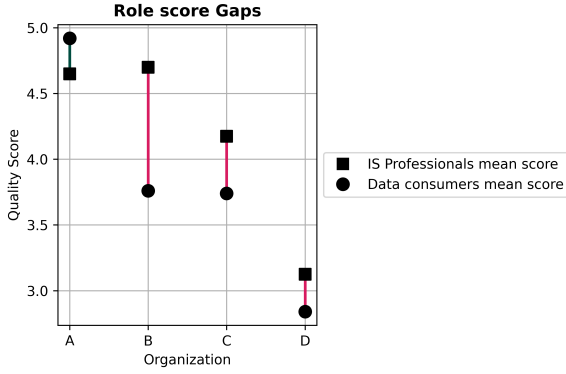


Figure 1.4: Role Gap Plot
(based on [20, 21, 14])

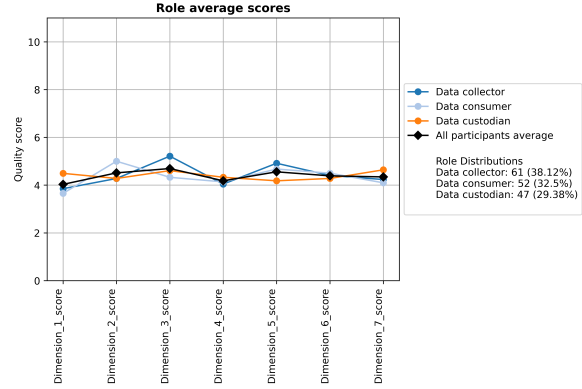


Figure 1.5: Role average scores
(based on [20, 21, 14])

1.4.3. The Aim Quality IQ assessment methodology (AIMQ)

AIMQ extends the previous comparative approaches, using their tools and concepts, and adding conceptual foundations and procedures to aggregate dimensions in a holistic view of quality. This methodology sees organizations as information production systems and conceives quality as value, in a vision of “affordable excellence” [14, 67].

The conceptual base is a 2x2 matrix called the Product and Service Performance model for Information Quality (PSP/IQ) [67]. Rows of the model consider the dual nature of information as a product and as a service [67]. In the product view, information can be produced, owned, stored, used, repurposed, and sold, and quality is defined by tangible product features [67]. In the service view, focus is on how data is transformed into information to be consumed, and quality is defined by intangible measures corresponding to latent product attributes important for effective use [67]. The columns present two views of quality as conforming to specifications, and meeting or exceeding consumer expectations [67]. Conformance to specifications applies data quality in operations using definitions and measurements to ensure products and services are free of deficiencies [67]. Meeting or exceeding consumer expectations is harder to measure and ensure, as it requires that information is actually useful and creates value for consumers whose needs are different from one another and change over time [67].

The resulting four quadrants, describe the possible combinations of how information is conceived (product or service), and how quality of information products is assessed (specifications or expectations) [14]. The model constitutes a framework that specifies which dimensions to use and how these are positioned in the matrix [14]. The application of AIMQ is explained by the creators as follows:

The PSP/IQ model consolidates the dimensions into four quadrants: sound, dependable, useful, and usable information (...). The IQA instrument measures IQ for each of the IQ dimensions. These measures are averaged to form measures for the four quadrants. The IQ Gap Analysis techniques assess the quality of an organization's information for each of the four quadrants. These gap assessments are the basis for focusing IQ improvement efforts (...). Analysis begins by analyzing a specific quadrant of the PSP/IQ model, such as usability. This continues for the other quadrants. After all four quadrants have been individually analyzed, the four quadrants can be compared to detect common patterns or to focus on the quadrant that most needs to be improved. [14, pp. 137-140]

1.4.4. Participative dimension selection and improvement plans

A work that reviews and points out shortcomings in AIMQ is [72]. The authors point out four issues [72]: (1) using a fixed list of dimensions expected to be meaningful in all contexts, (2) a lack of clear rules and justifications for developing individual items of the IQA questionnaire, (3) excluding independent experts from the considered stakeholder groups, and (4) being limited to assessment, missing valuable directions towards improvement. The authors of [72] present a more flexible methodology solving these problems.

In the proposal of [72, 74], organizations use dimensions that they have identified as relevant, either by expert analysis or based on a questionnaire that captures importance ratings. The used instrument presents clear definitions of the dimensions alongside the questions for their importance ratings to avoid ambiguity [74].

The methodology of [72] adds the role of Independent Experts to the stakeholder groups to consider. This is considered convenient because IS professionals in the organization may show biases [72]. Moreover, Independent experts have a technical knowledge that data consumers usually miss, making possible new comparisons with a group that is technically skilled but external to the organization [72].

During the creation of the assessment questionnaire (equivalent to the IQA survey of AIMQ), items are formulated using a simple rule that describes four questions [72]: a direct question that includes the dimension name, a reverse question that explores the negative perspective using the concept opposite to the evaluated dimension, a question using a synonym to reduce biases for specific words, and a question that includes the complete definition of the dimension.

The results of the assessment questionnaire are used by [72] to categorize the state of quality of the selected dimensions into three groups: those scoring below 4 are considered

weak, those between 4 and 7 are seen as intermediate, and those above 7 as strong. Improvement is prioritized for weak dimensions [72].

Regarding improvement, an approach supported by questionnaires is followed again to explore the perceptions that different stakeholder groups have in relation to possible improvement strategies [72]. For each dimension, personal evaluations of strategies are collected using Likert-scale rates where 0 is "irrelevant", and 10 is "vital" [72]. The questionnaire presented by [72] also asks whether each of the strategies is already in use, and encourages respondents to share any valuable improvement strategies that they are aware of, but that were not included. After the improvement surveys have been delivered, strategies widely recognized as vital become candidates for execution. However, in the case that answers from different stakeholder groups do not agree, it is necessary to investigate the reasons leading to this, in order to better choose how to allocate resources for those strategies most likely to be effective [72].

AIMQ [14] and its extension for dimension selection and improvement [72, 74] describe a valuable conceptual framework, principles, instruments, and analysis techniques worth considering during data quality interventions. These methods not only provide psychometrically robust measurements but also are highly context-aware and data type independent, showing diagnostic power that supports finding the sources of problems by listening to different stakeholders, and have a key instrumental value for quality enhancement by involving organization members in quality activities [14, 21, 67, 74, 72, 20].

1.5. Understanding Quality in complex data objects

1.5.1. The Heterogeneous Data Quality Methodology (HDQM)

Scientific and engineering advances in Data Quality and its related fields such as computer science, hardware infrastructure design, statistics, data mining, and systems modeling and architecture, have produced a wide collection of exceptional models, formulas, tools, definitions of relevant concepts, and technological devices and applications that result of immense value for DQ practitioners. A successful DQ intervention requires a systematic approach to apply this body of knowledge and deliver tangible benefits in organizations. To this aim, data quality methodologies are fundamental. According to [16, 75] a data quality methodology is:

a set of guidelines and techniques that, starting from input information describing a given application context, defines a rational process to assess and improve the quality of data. [16, p. 16:2]

Methodologies generally consist of three phases [16, 75]: State reconstruction, Assessment, and Improvement. Specific methodologies can focus or reduce their scope to cover only some phases, or they can be formulated in a modular-flexible way to leverage current knowledge of the organization and limit activities to what is really needed, reducing times and costs [16, 75]

The management of unstructured data, including data types characterized by their trivial subjective evaluation but complex automated evaluation, has become of central importance with the rise of Big Data, and significant attention has been placed on the construction of highly efficient algorithms and systems to process them [76]. Nonetheless, not only do Big Data contexts have to account for diverse data.

A substantial contribution to the understanding and effective work with data characterized by variety is [75], where the *Heterogeneous Data Quality Methodology (HDQM)* is introduced. The core of HDQM is a meta-model designed to understand quality in terms of conceptual representations [75]. The authors explain that to avoid the complexity and costs of previous methodologies, the meta-model was designed for a simple but robust abstraction and formalization of information use in the organization [75].

As an example of such highly complex approaches, [75] mentions the Information Product MAP (IP-MAP) used in TDQM. Also notable is the existence of extensive literature on Conceptual Modeling and concept-oriented content management (CCM), valuable to

develop more sophisticated and accurate models [77, 76].

An important remark on conceptual modeling, as explained by [76], is that when describing information systems in terms of concepts, distinction must be made between the technical and domain knowledge aspects of entities, and their intensional semantics (definitions and meaning) and extensional semantics (associated values and observations):

In a model, a concept names a phenomenon in the domain, and rules and constraints (...). A technical concept is a data type (“integer”) and might be given for a particular database. A domain concept captures domain knowledge (“typical human weight”) and can, therefore, not be predefined in a generic way. It carries some domain semantics that may be given as intensional semantics, i.e., by rules and constraints that data of a certain class have to fulfill, or by extensional semantics through the set of data that it is assigned to. (...) a possible answer to the question whether to start with a model (as databases do) or to start with data (as data analytics do): Instances and models co-evolve. [76, pp. 14-15].

HDQM has several features aimed for usability and flexibility [75]: it considers quality at different granularity levels, accounts for the role of data within processes, is open to both data-driven and process-driven improvement strategies, is cost-aware, and it does not impose specific dimensions. It is still worth mentioning that for [75], Accuracy and Currency are two significant dimensions given their importance in different contexts.

The meta-model elements are explained by the authors of [75] as follows:

The HDQM meta-model (...) represents the main types of organizational knowledge (...). A Process is an articulated sequence of activities (...). An Organizational Unit is a significant element of an organization that is involved in data production, use and processing; it is characterized by an internal structure and a set of internal rules. A ReSource (RS) is any information source (...) to represent some aspects of the reality of interest. (...) RSs can be seen as (...) sources of supply by which to get information (i.e., a resource) – or also as the origin itself of these data (i.e., a source) (...). A Conceptual Entity (CE) is any concept that refers to a single phenomenon (...) possible to abstract (...) independent of the specific way a single RS represents it [75, p. 66]

Dimensions in HDQM are considered both at ReSource level and at Concept Entity level by applying composition operations.

The authors of [75] note that even if the initial HDQM proposal was limited to relational data sources (either structured, semistructured or unstructured), involving in particular XML and document files, such a focus is convenient but not necessary, highlighting the

advantage of relational settings of having explicit concepts and aspects associated with them [75]. Therefore HDQM can be extended to a wider variety of settings and objects through "dimensions, metrics and improvement techniques tailored to specific type of data" [75, p. 62].

HDQM follows the general 3-phase structure of DQ methodologies [16, 75]. We summarize next the implementation of these steps within HDQM [75]:

1. State reconstruction: Contextual information is collected [16], conducting interviews and surveys to diagnose quality problems and identify the context-specific elements of the meta-model [75]. Focus is placed on the master-data and those data associated with quality problems [75]. In this step each ReSource is carefully analysed considering its designated intention, and reverse engineering is applied to produce conceptual schemas [75]. Subsequently, conflict resolution procedures are applied in order to derive a unified, integrated schema.[75].

2. Assessment: Involves two steps [75]: (1) ReSources ranking, where the relevance and scope of ReSources are quantified to get composition weights and guide improvement decisions, and (2) data quality measurement, which considers the identified problems to select meaningful dimensions and metrics, and computes the values.

3. Improvement: Performed in three steps [75]: (1) Data quality requirements definition, which sets target quality values at the Concept Entity level and propagates these requirements by defining individual ReSource target values, (2) Selection of improvement activities, based on the goal targets and the properties and types of involved ReSources, and (3) Selection and Evaluation of Improvement Processes, where focus ReSources and candidate improvement activities are placed in a matrix to study combination strategies in the light of their expected data quality improvement effects and execution costs.

A reasonable complaint against HDQM could argue that this methodology, by aiming for such high levels of abstraction, keeps a significant gap between theoretical background and operational definitions. However, such a view can be challenged by noting that (1) the provided generic definitions and operators are sufficient to elaborate context-specific metrics, and (2) methodologies attempting to propose detailed but context-independent dimensions and metrics often fail for some contexts, becoming useless or even confusing and detrimental to practitioners (see Subsection 1.1.3).

1.5.2. Assessment of unstructured data

Mining for value in diverse data

Competitive organizations assign high importance to extracting insights from data by applying solid procedures for Knowledge discovery and data mining [3]. Knowledge discovery is the "nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [78, p. 30]. Similar but more specific is the concept of data mining, defined in [79] as:

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques (...) find novel and useful patterns (...). They also provide the capability to predict the outcome of a future observation (...). Not all information discovery tasks are (...) data mining. Examples include queries (...) because such tasks can be accomplished through simple interactions with a database management system or an information retrieval system [79, p. 4]

These data mining processes can be applied to the wide diversity of data owned by organizations, creating challenges and opportunities beyond traditional relational and structured data, given the advances in the understanding and management of flexible models, unstructured data objects, and Big Data contexts. Companies working with Big Data must properly consider data Volume, Variety, Velocity, Veracity, and Value (known as 5 Vs) [80]. These characteristics increase the difficulty of applying established data quality approaches such as sampling, which is affected, given the unpredictability of shapes and values, in its task to obtain adequate subsets that can capture the properties and representation of the complete datasets processed in the information system [76]. Similarly, in Big Data settings, practices as anomaly detection, must resort to the available domain knowledge and organization information to understand the intensionality of data and derive some predictability [76]. In this way, Big Data requires practitioners to select and adapt adequate algorithms, infrastructures, and workflows [80, 76, 3].

Regarding variety, in recent years, numerous data mining techniques have been introduced for unstructured types of data, such as text, audio, images, and videos, but the success of their application largely depends on possessing high-quality data, which creates a need for appropriate assessment methods to build up confidence before decisions are taken [3]. In high-velocity applications, such as those integrating real-time data streams, the requirement for low-latency methods to process, store, and analyze data [80] is extended to these quality assessment methods, making it necessary to use time-efficient and scalable dedicated automatic tools and metrics [3].

A noteworthy work exploring the meaning of quality in Big Data and promoting the use of quality aware algorithms is [22]. For [22], two key factors of big data are being highly

source specific and unstructured and highlight that different data types need to use different appropriate quality dimensions [22]; for example in texts the relevant properties such as readability and ambiguity can be defined formally by considering syntaxis, and sets of lexical and semeantic features to describe rules and patterns [22]. On this topic a valuable source on text quality dimensions is [39]. Similarly, image quality can be described in terms of fidelity (match to an original), naturalness (match to viewer references) and usefulness (suitability for a task and gains from information use) [22].

The authors of [22] consider as a valuable foudation for understanding Big Data Quality the UNECE(United Nations Economic Commission for Europe) classification, which econcompasses three types of big data: (1) human sourced, consisting of the records of human experiences in forms such as texts, images, audio and videos, (2) Process-Mediated, corresponding to registers of business events appearing in operational and analytic systems and modeled in terms of domain entities and data items, and (3) machine generated, associated with sensor mesuarements and computer logs that digitalize world phenomena being present in contexts such as security, scientific research, smart cities and Internet of Things.

Quality Dimensions for Unstructured Data

A remarkable work on the quality assessment of unstructured data is [3], which contributes solid theoretical concepts on the meaning of dimensions and provides high practical value by identifying quantifiable indicators and their existing implementations. In [3], the quality of unstructured data objects is conceived as conformance to expectations and fitness for use as characterized in relation to their use in data mining pipelines.

This vision includes several steps involving one or more machine/algorithm data consumers and a final human data consumer. For illustration [3] describes the example of a pipeline, where each step takes as input the output of the previous step, starting with raw data as input and then it moves through a sequential process in a preprocessor, a classifier, and a visualizer (3 machine consumers) to deliver a final output for a user (human consumer).

Dimensions are defined theoretically in [3] by comparing the considered dataset and a group of "ideal" datasets, where "ideal" is defined from three points of view: data consumer and task (which are usually equivalent), and representation of the real world. Table 1.4 presents these dimensions identified by [3]. Similar to dimensions used for traditional structured data, here dimensions can be described intuitively and generically, but their application results even more context and purpose-dependent.

Helping to overcome the challenge of translating dimension definitions into operational terms (see Subsection 1.1.3), [3] presents a series of indicators in the range 0 to 1. A summary of indicators discussed in [3] is presented in Table 1.4, where we adapt their original table enumerating dimensions and indicators to contain summaries of explanations and examples from the same source.

Dimension	Focus	Ideal
<i>Interpretability</i>	Data consumer	Data satisfies expectations • Human consumer: expectations determined by personal knowledge, experiences, and objectives. • Machine consumer: expectations defined by rules, statistical models, properties of previously observed data (training data), types, taxonomies, formats and conventions.
<i>Relevancy</i>	Task	Data optimizes performance
<i>Accuracy</i>	Real world	Data is representative • Direct comparison: Conditioned on the availability of a gold-standard reference dataset containing the ground-truth values as annotated by human experts who evaluated the unstructured data objects. • Estimation: Rules to verify conformance to expected patterns. (Similar to the consistency dimension with semantic rules [81])

Table 1.4: Dimensions for unstructured data discussed in [3]

Dimension	Indicator	Aspect measured	Examples in text mining
Interpretability	<i>Similarity to training data</i>	Resemblance of new input data to data seen before, as quantified by formulas and algorithms designed to work with the particular data types.	Cosine Similarity and Greedy String Tiling
	<i>Confidence</i>	Classification probabilities, which are typically passed to max or threshold operators to obtain discrete class predictions and are considered directly as continuous scores.	Applying a part-of-Speech (POS) tagger and inspecting the probabilities used in decisions
	<i>Percentage of noisy instances</i>	Deviations from the expected distributions and properties of data objects • Human consumer: Divergence from the well-formed shapes, and easily understandable properties of data objects typically consumed in daily life and many work settings. • Machine consumer: Elements that appear as unexpected because they contrast with past seen training data that was assured with procedures of data selection, preprocessing and cleaning.	Elements such as grammatical errors and misspellings difficult readability for humans. Machine evaluations can consider the count of syntax parser errors, and the presence of unusual text lengths, unknown or rare words, and special characters or sequences as URLs
Relevancy	<i>Match to encoded information needs</i>	Individual estimations for the relevance of data objects, in the same way as information retrieval systems that filter and rank search results.	Frequency of keywords selected by a user, and domain specificity match, defined as the coverage of sets of central technical terms and specialized concepts. Implementations can use cosine similarity or related formulas
Accuracy	<i>Precision</i>	Instances assigned as belonging to a category actually belong to that class	Defined in relation to particular tasks and datasets
	<i>Accuracy</i>	Correct predictions	
	<i>Quality of gold annotations</i>	Inter-annotator agreement showing consistent judgments for the data when observed by different persons	

Table 1.5: Indicators for unstructured data (adapted from [3])

Confidence estimation and Calibration

An important remark on confidence is that even if many classification methods produce associated probability outputs for the predicted Class labels, guaranteeing that these probabilities are reliable is a specialized problem. This problem has significant theoretical implications in data quality (See Subsection 1.7.2).

In many critical domains, the estimation of prediction confidence is as important as the predictions themselves and this has motivated calibration research, which pursues the accurate estimation of the probability that a given prediction is correct [82, 83, 84, 85, 86]. These procedures can be applied in a model-specific manner or using post-hoc methods [85]. Investigations in this area also consider cases where model error is heterogeneous across the input space, driving more sophisticated approaches such as local calibration methods that account for subgroup-level performance variations [82, 83, 84, 85, 86]. Consequently, default model outputs can serve as a useful quality indicator but they should not be treated as indisputable ground truths.

The need for confidence estimation and calibration, and reviews of applicable methods in the field of NLP have been discussed in works such as [4, 87, 88]. A notable study of this topic for the Transformer architecture is [89], which concludes that pretrained models tend to be calibrated for in-domain use, but reliability is substantially decreased for out-of-domain observations, and highlights as effective techniques to enhance calibration temperature scaling and label smoothing. For this same problem, [90] proposes a regularization approach during fine-tuning.

1.5.3. Approaches to the Analysis of Big Data

A remarkable review of data quality in big data is [76], whose authors advocate for strategies that, while ensuring high performance, maintain only the minimally necessary level of complexity under the principle of parsimony (or Occam's razor). In [76], three approaches for the analysis of Big Data are identified, considering the certainty levels on data structures, and the associated conceptual and technical solutions to address them. A summary of this taxonomy is presented in Table 1.6.

Approach	Idea	Examples
<i>Educated Guess</i>	Form hypotheses of the data structure to generate predictions	• Hidden Markov models • Association rules
<i>Methods in the Wild</i>	Making no assumptions of data properties, inquire directly for its structure and quality	• Clustering • Dimensionality reduction
<i>Neural Networks</i>	Learn arbitrarily complex functions that involve the encoding of representations at different levels of abstraction, extracting relevant features from the data	• Self-Organized Map (SOM) • Deep Neural Networks (DNN)

Table 1.6: Analysis approaches for Big Data (adapted from [76])

When there is certainty about the data structure of objects, it is possible to leverage domain-specific knowledge in the form of rule definitions, parametric methods and stochastic models [76]. This can include the use of Data Mining and Machine Learning methods which have been adopted in specific domains, and that have a clear domain-specific interpretation and have been subject to theoretical and empirical validation [76]. Cases with uncertainty about data structures and distributions benefit from unsupervised approaches [76]. Lastly, incorporating the learning of meaningful features into task optimization through neural networks can achieve high performance [76].

Limitations of DNNs: Generalization and Shortcut-Learning

While Deep Neural Networks show the highest performance in different tasks, this should not motivate us to consider them as a first choice [76]. A fundamental reason to be wary of neural networks is their use of shortcuts to quickly learn parameters that can achieve high performance, ignoring the scientific principles and safeguards of parsimony and causality [76]. On this the authors of [76] note that:

DNNs always learn the easiest possible solution to a problem, but understanding which solutions are easy (...) requires disentangling the influence of structure (architecture), experience (training data), goal (loss function) and learning (optimisation), as well as a thorough understanding of the interactions between these factors. [91, pp. 17-18]

According to [91], shortcut learning creates solid ground for skepticism in Machine Intelligence, and proposes adapting *Morgan’s Canon* of psychology, which describes the principle of parsimony of psychological process theorized to explain animal behavior, to Machine Learning, adapting it in the form “Never attribute to high-level abilities that which can be adequately explained by shortcut learning” [91, p. 12]

This case of [91] against anthropomorphizing technological artifacts, not only rejects the ability for machines of thinking but also includes rejecting the capacity for neural networks of imitating the complex processes mediated by perception such as understanding texts or perceiving properties in images.

A related problem lies in the constraints imposed by available training data; the authors of [91] make a key observation on datasets when noting that the traditional single dataset split successfully serves models to learn and evaluate performance in unseen but yet *independent and identically distributed* data (i.i.d), but for truly trusting these systems it is necessary to test models on *out-of-distribution* data (o.o.d). The availability of popular benchmarks must not be a reason for overtrust in model correctness, making it necessary to carry out robust validation procedures and provisions to make sure data is representative of the world [91]. The authors of [91] hold that less attention should be given to performance, and more to interpretability and generalization capacities, arguing that “a dataset is useful only for as long as it is a good proxy for the ability one is actually interested in” [91, p. 11]

These considerations are central to all critical domains and should not be overlooked in general practice; in data quality, automatic-metrics should also be aware of these issues, given the instrumental value and central role of precise data quality assessments for the selection and improvement of the data that is used for informed decision making and the creation of quality products and services needed by society [5].

Indeed, [91] explains that the phenomenon of shortcuts has an important role in algorithmic fairness, as model shortcuts can replicate biases and amplify societal disparities by accessing directly, or through proxies, attributes of human diversity that should not determine the functionality of systems. This unfairness has been observed in the imitation of real-world discrimination dynamics encoded in data, and in systems showing different performance for underrepresented groups [91].

Post hoc Interpretability

Intrinsically interpretable models have the advantage of being easily modified, varied, improved and explained [92, 93]. However, the superior performance of complex but

less interpretable models has motivated research. Explaining model predictions (post-hoc explainability) is a widely studied approach [93, 92, 94, 95]. One of such methods applicable to objects such as text and images is LIME [96]. Regarding the field of Natural Language Processing, a comprehensive survey of post hoc interpretability methods for Neural Networks can be found in [93].

Nevertheless, several authors argue that post hoc interpretability provides only a limited view, and what constitutes a satisfactory explanation and the selection of which plausible hypotheses to consider or ignore as likely explanations often lacks objectivity [94, 95]

1.5.4. Clustering for Data Profiling

The authors of [76], when reviewing the complexity of managing Big Data in contexts such as smart cities, remark that:

with such an increasing amount of (...) unstructured (...) data, it is almost impossible to extract relevant information or detect quality and meaning (...). Unsupervised learning that intends to discover patterns in unlabeled and unstructured data or detect anomalies could be an algorithmic solution. [76, p. 12]

Observing and studying distributions of data scattered in representation spaces is a central task in Data Science. In this scenario, clustering methods employ similarity or distance metrics to partition data into meaningful groups in an unsupervised setting, without relying on labels [76, 97, 98, 99]. Clustering has two key advantages [98, 76]: first, it has not only one but several well-established algorithms (e.g., Hierarchical, k-means, DBSCAN, GMM-EM) that can be selected, some of which have low time complexity, and second, in cases where no labels are available, it allows to extract insights from the data without having to perform the resource-intensive tasks of analysis and annotation.

Members of each cluster are expected to be similar, showing cohesion, while different clusters should be clearly distinct and separated [76, 97, 98, 99]. These assignments can either be absolute or soft-probabilistic [98, 99]. The epistemic foundation for such groups rests on the assumption that essential facts are linked and can be observed in terms of contingent (not necessary, or variable) properties of data [26]. In this way, clustering can help to identify latent properties and get insights into structures and patterns [26]. As clustering relies on space representations and similarity metrics, it can be used to study any entity having data that satisfies these elements [76, 97].

Cases having clear numerical vector shapes often use directly euclidean distances, while domains working with complex data objects turn to their specialized representations and

metrics [76, 97, 99]. As examples, edit distance is used to measure text similarity, and Bag-of-words representations have been used successfully in Natural Language Processing and Computer Vision (by extending the textual idea of vocabulary, which allows to ignore nuances of spatial locations) [99, 100]. Some sources on the application of clustering in text, image and audio data types are [101, 100, 102, 103, 104, 105].

This flexibility and openness for specific implementations is explained by [76]:

characteristics in the training data need to be captured by the algorithm. Cluster algorithms are dedicated to specific data topologies, hence the need for data normalization. (...) the similarity metric needs to reflect the core attributes of each dataset [76, p. 8]

Straightforward uses of clustering algorithms in data quality, within data profiling and data cleaning activities, are outlier detection and noise removal. A complete explanation of these concepts can be found in [106]:

Outliers are samples which deviate extremely from other data samples. The process of detecting outliers is also known as anomaly detection (...) considered as the complement of the clustering process. Outliers are genuine objects in the data set that have unique characteristics that isolate them (...). While noisy data result from erroneous data that randomly appear in datasets. Outliers reveal that they have been generated by a mechanism different than the one that generated the remaining data. [106, p. 21]

In addition to outlier detection, [99] describes numerous advantages of considering cluster analysis in data quality: clustering can be used to study the dynamics of data acquisition and transformations, perform the detection of duplicates and inconsistencies, hypothesize data structures, and evaluate schemas and models. Moreover, it facilitates the detection of quality problems, and it can reveal useful knowledge and rules embedded in specific problems and datasets [99]. Clustering has also been applied to metadata [107].

Related to clustering but operating in supervised settings, k-Nearest Neighbor (KNN) algorithms leverage spatial representations and distance measures in datasets using known labels [98]. Although considered a lazy learner with frequently high inference cost, its conceptual simplicity and flexibility for regression, classification, and noise detection tasks make it a useful tool to consider [98, 108].

Evaluating the quality of a clustering partitioning is a topic that has been the subject of extensive investigation in the field of data mining. Evaluation strategies can rely on internal validation metrics based on cohesion and separation, such as Within-Cluster Sum of Squares (WSS), Between-Cluster Sum of Squares (BSS), and Silhouette Score (SS), or

they can use external validation metrics exploiting available knowledge of ground-truths, as in the Adjusted Rand Index (ARI) and Purity [98]. Similarly, for anomaly detection, simple statistical approaches can be employed, such as defining decision thresholds in terms of standard or median absolute deviations from central tendencies, or specialized algorithms can be used, such as Support Vector Machines (SVM), Isolation Forest, and Local Outlier Factor (LOF).

Although these cluster validation methods and anomaly detection algorithms are valuable for confident pattern recognition, reading these patterns for data quality assessment requires further knowledge; they serve as valuable methods for data profiling and cleaning, but because of their unsupervised nature, expert examination remains crucial to explain the patterns in the light of domain knowledge, and to describe groups in terms of their quality properties [76, 99, 109].

Therefore, within the assessment phase of DQ interventions [16], clustering stands out as a valuable method for data analysis that also shows some potential for bringing insights on requirements, critical areas, and processes, but whose systematic application requires investigation; notably on the problem of translating clustering indicators and concepts into data quality specific metrics that can be considered during measurement procedures and be analyzed to quantify the impact of improvement strategies.

1.6. Data annotation

1.6.1. Gold and Silver standard datasets

A fundamental condition to start the development of a quality assessment model is to have data representative of the useful distribution or actionable patterns it should learn. Two types of datasets widely used in the training of predictive models and characterized by their high quantity and high quality data are the silver and golden datasets.

In [110] gold standard datasets are defined as follows:

Gold standard corpora in NLP context are manually annotated collections of text. For high quality (...) multiple experts view the data independently and the inter-annotator agreement is computed to ensure quality. This makes the creation (...) a very costly process. The agreement of the annotators depends on the ambiguity of the data, skill of the annotators and the task at hand (...) because these determine not only the difficulty when choosing a tag but also the number of (wrong) choices (...). As NLP tasks largely depend on domain and intended results, proper gold standard corpora are required for each domain and task. [110, p. 2]

Each of these safeguard mechanisms can be loosened, instead of totally removed, to reduce costs while also preserving elements for reasonable trust [110, 111]. In this way, it is feasible to reduce the total annotated instances or the number of multiple annotations per instance, as well as employing annotators with lower domain expertise [110, 112]. In some tasks, crowd-sourced annotations delivered by non-expert contributors have been shown to constitute a cost-effective alternative, trading off expertise for quantity.[111].

An alternative or complementary approach is to use silver standard datasets, which strive to approximate the quality of the gold standard within limited resources, by initially employing multiple systems to automatically annotate large datasets, and subsequently apply a harmonization process to enhance consistency [110, 113]. The creation of reliable models using noisy annotations produced through methods such as non-expert crowd-sourcing, automatic labeling, and the application of heuristic rules is the subject matter of the Weakly Supervised Learning (WSL) paradigm [114, 115, 116]. Nevertheless, the authors of [114] remark that the production of models trained on noisy labels still relies heavily on having some clean trusted samples for validation, and that the convenience of a trade-off of quality for quantity in training data is often overestimated.

1.6.2. Production of annotated datasets

Data annotation projects can be conducted either internally or through external service providers, and in both cases, it is essential to implement a rigorously defined annotation process and a careful validation of the outcomes. Regarding technical needs, internal initiatives can leverage one of several software tools that have been developed to support annotation tasks, some of which are highly flexible [117]. As related to process and validation, comprehensive reviews of data quality concepts and management practices for annotation have been published, such as [118, 119]. Unsurprisingly, the domain-specific nature of data-powered tasks has been met in literature with the publication of specialized annotation guidelines, for example [27, 120] present methodological and validation contributions in the field of Natural Language.

These projects follow systematic procedures where the relevant properties are clearly defined and accuracy is assured by providing guidelines on how to annotate the data and how values are expected to be mapped [119, 118, 27, 121]. Subsequently, the quality of results is controlled through statistical analysis, in particular by checking consistency [119, 118].

A fundamental aspect that is domain independent is the need for diversity and fairness in the generation and use of datasets: producing a new dataset, and extracting its embedded knowledge for developing tools, can not be reduced to a functional pipeline of inputs and outputs, to be solely evaluated with performance metrics [122]; It involves complex ethical, psychological and sociological aspects that must be considered [122, 35]. Hence, when annotation tasks are outsourced, it is imperative to take provisions for an ethical AI supply chain [121]. It is ideal to select providers that document their practices and that are open to being audited and have engagement with the commissioning party [121].

This engagement can involve creating channels for anonymous complaints, promoting communication between workers, and offering support to properly explain what is needed, for them to understand what is expected, and how to better perform the job [121]. Moreover, not only fair compensation but also providing a reasonable time frame to complete the annotation tasks is important for worker well-being and annotation correctness [121].

In contrast, some providers should be avoided on suspicion of unethical practices [121]. For example, the practice of subcontracting limits the pay given to those workers actually doing the task, and some platforms can have structures and designs that enforce an asymmetrical relation, for example, by denying or difficulting the possibility to dispute the arbitrary rejection of a correctly delivered job [121].

In addition to transparency during the production process, the final dataset should also be transparent by including documentation and metadata, for example in the form of a Data card, to facilitate usage and repurposing [11, 123]. The goal of this artifacts is providing an in depth profile of data, which can also include ethical considerations as the provisions taken for ethical annotation, and the possible presence of toxicity patterns in sections of the dataset [121, 123, 32]. This efforts not only contribute to preserve the value of data over time, but also can support the effectiveness of social justice mechanisms, such as reputational damage for unfair organizations and products [121, 11].

1.6.3. Variability in annotations

Annotator agreement and correlation are some of the aspects considered in the statistical validation of the reliability of annotations, using indicators such as the percentage of agreement, Cohen's κ , Fleiss's κ , and Krippendorff's α [118]. These mathematical tools are important because the phenomenon of noise is frequently observed in projects using human-annotated data [124] [125].

Even if consistency is desirable, variability does not always represent a sign of poorly executed annotations [38]. Indeed, variation in human annotations has a very different interpretation from other types of data [38]; While finding variance in numerical and categorical data samples based on modeling and measuring real-world absolute ground-truths can indicate an alert of imprecise or unreliable data [38], the comprehensive view for human annotations is that:

variation can indeed be disagreement due to difficult cases, subjectivity or cases where multiple answers are plausible. (...) human label variation (HLV) provides rich information that should not be discarded [38, p. 1]

Such diversity can be leveraged by means such as consistency analysis and weighting, or a simpler, straightforward policy as majority vote can be applied [38, 111, 126]. Related research has explored cases where variability can be explained systematically by understanding what is important to annotators and the relation of assigned labels to the ground truths relevant as targets in a particular task [127]. In these cases, annotations can be seen as imperfect but meaningful proxy indicators for algorithms to successfully extract and learn needed features for task proficiency [127].

1.7. Theoretical perspectives of model-based metrics

1.7.1. Automatic-metrics and Meta-metrics

The performance of machines at tasks previously possible only through human perception has increased significantly in recent years. Similar to structured data, complex data objects are associated with specialised domains that use specific metrics and quality dimensions [22]. Two concepts central to the general theoretical understanding of model based quality assessment are automatic-metrics and meta-metrics.

Any tool intended to replicate the human evaluation results of data quality assessments will fall under the concept of automatic-metric [1], and the task of producing such tools will correspond to the problem of training a meta-metric proficient model [1]. A concise and complete characterization of automatic-metrics and meta-metrics is the one presented in [1]:

Conceptually, an automatic metric should emulate human judgments. Selecting an automatic metric typically entails generating a set of human judgments for a wide variety of outputs from a large number of different systems, and selecting the automatic metric that produces scores most similar to the human judgments. But how do we quantify similarity? To select the metric which produces output most similar to human judgements, we need a meta-metric to compare metric scores and human judgments. [1, p.1]

Automatic metric research is a field that has captivated wide interest across different tasks, in particular associated with generative AI systems where examples of such proposals are [47] in text generation, [128] in audio aesthetics, [129] in text-to-image generation, and [130], whose authors present a flexible approach intended to optimize the combination of existing metrics for vision and language tasks.

Important remarks on automatic and meta-evaluation metrics can be found in [49, 2], whose authors argue that meta-evaluation metrics can be helpful during the development of models because of their convenience to replace complex and expensive human annotation processes with simple automatic and cost-effective measurements. These same authors find it fundamental to always consider the specific context, and deployment environment, as well as who are the information consumers and the overall goal, role, or purpose for deploying the system, holding that the selection of quality indicator metrics should carefully consider even those metrics proposed to be used within a specific domain or group of interrelated tasks [49, 2].

Indeed, [2] advocates for an initial use of robust and stable "local" metrics, where combinations of metrics support confidence, and a posterior long-term plan for monitoring and adapting quality assessment is applied to consider emerging dynamics and foster persistent high quality. A strong example that illustrates the need for a user-centric context-aware approach is shown in [49] when remarking on how different categories of natural language data, such as scientific literature, clinical data, and daily news are used, consumed, and evaluated in very different ways because of their fundamentally different intended use and the characteristics defining the target readers. trust in domain interpretable methods.

The experiments of Subsection 3.3.2, where model performance is determined by training data, serve as an example of the importance of using automatic-metric models only as local metrics, requiring more evidence for generalization claims [91, 2].

1.7.2. Model predictions as data quality estimations

The concepts of automatic-metrics and meta-metrics constitute a key foundation for understanding Artificial Intelligence model performance in terms of quality estimations. A key contribution of [3, 4] is the idea of implementing quality metrics through the use of *confidence*, corresponding to the membership output probabilities of a classifier.

In addition to this perspective, a central issue is verifying in specific use cases whether these tools are fit to serve as robust quality metrics. On this issue, an exceptional contribution for data quality theory and practice can be found in [5], whose authors describe five requirements of data quality metrics: (1) the existence of minimum and maximum values; (2) values must be interval scaled, such that the distance between two measurements can be determined and has an associated meaning; (3) metric values and configuration parameters must follow the scientific criteria of objectivity (independence from external influences), reliability (accuracy and replicability) and validity (effective measurement of the target theoretical construct); (4) aggregation can be performed in a sound way to assess quality at different levels of granularity; and (5) the metric must be economically efficient, justifying the costs of configuration and use through expected returns.

For the case of a metric defined as the output probability that an instance belongs to the *satisfactory quality* class, computed by a binary classifier model, [3, 4], requirement (1) is satisfied as values range from zero to one. Regarding requirement (2) a validated trained model can be trusted to provide correct rankings assigning higher scores to better quality objects, but the meaning of the interval itself cannot be described as probabilistic unless proper calibration procedures have been applied. Requirement (3) requires careful analysis in each use case by considering the aspects that are the target of measurement, but in

principle the availability of frameworks and literature in Natural Language Processing, and the existence of benchmarks and validation methodologies aimed at testing model capacities and assessing effectiveness in new contexts, provide reasonable grounds to argue that in general applying model-based metrics can follow a scientific approach. While tuning model hyperparameters for a task or context can be challenging, performing a comprehensive search or following heuristic rules derived from empirical studies focused on specific models are evidence-supported options [131, 132]. Regarding requirement (4), aggregation is easy to perform and using min and max operators can provide a first insightful indicator of quality level bounds. Anyway, guaranteeing general soundness raises further considerations. Again, on this point confidence estimation and calibration constitute a mechanism to ensure that operations such as *average* and *weighted average* have a precise and consistent interpretation [85, 4, 5]. Lastly, on requirement (5), the use of lightweight models trained and applied for measurements using owned hardware has reduced time and expertise costs. However, the acquisition of training data and the related quality assurance activities can present additional costs.

The authors of [5] present the five requirements as fundamental to trust quality metrics for confident decision making. The importance of upholding all five requirements in critical tasks and decisions involving high-cost stakes is notable [5].

Consequently, the careful consideration of requirements (2) and (4) raises a key insight: even if training classifiers is a straightforward task, output membership probabilities provide a quality metric robust only at the instance level, for pairwise comparisons, and for bound estimation. Safe average aggregation requires a consistent interval [5]. Future research can explore the conceptual and technical aspects needed for a robust use of model-based automatic-metrics at higher levels of granularity.

1.8. Principles of automatic DQ tools

A review that clearly characterizes the role of data quality toolkits within information systems is [133]. Investments in DQ tools aim for improving quality and two paths exist to achieve this objective [133]: While the data-driven strategy considers the data correctness, fit to the specific information system, and its source, to subsequently implement targeted modifications on the data values, the process-driven strategy aims for a sustained, long-term increase in quality by solving the causes of deficiencies through the redesign of processes and the insertion of quality controls.

The quality of a data object is not static, and problems may rise at any time of its life cycle during different interactions [133]. Data problems can emerge at the integration of new data into the information system, after processes involving data manipulation, or they may arise naturally from quality decay dynamics where data eventually stops being an accurate representation of the real world as the modeled objects change [133].

The authors of [133] identify four important features provided in data quality toolkits: (1) Assess data quality, (2) highlight low-quality subsets, explaining their observed deficiencies in terms of patterns, statistical descriptions of processes, and visualizations as the Pareto chart, (3) prescribe and implement suitable quality enhancement actions, and (4) track data lineage so users can assess whether previously performed operations make data ready for their tasks.

An organization making efforts for quality improvement depends on robust temporal analysis to check whether these efforts are successful, and therefore, a necessary complement to an automatic data quality toolkit is a monitoring application [133]. The central component of DQ monitoring applications is a data quality repository that stores schema descriptions, metadata, and assessment results, and is used to perform continuous profiling for time-series analysis [133].

One key idea found in [133] is that the concept of unit test can be applied to data quality. The original concept is a fundamental part of Software Engineering as explained in [134]:

Unit testing verifies the functioning of software modules in isolation (...). Developers can use unit testing tools and frameworks to extend and create an automated testing environment. For example, the developer can code criteria into the test (...) to verify the unit's correctness under various data sets. Each test is implemented as an object, and a test runner runs the tests. Failed test cases are automatically flagged and reported during the test execution. [134, Chapter 4 p.15]

Adopting unit tests to data quality means implementing automatic processes to verify

user-defined DQ criteria [133]. The authors of [133] explain that adopting automatic DQ systems is beneficial for organizations of any size and complexity. In small organizations DQ tools provide a robust but easy way to assess and improve quality, reducing expertise barriers, while in large organizations, incorporating these tools becomes a key component to systematically manage and assure quality of complex data flows, often including big and unstructured data [133]. This is particularly convenient for those infrastructures designed to condense diverse data into data lakes [133].

The automatization of quality assessment of unstructured objects can also implemented following these guidelines.

2 | Methodology

Objective

The goal of this work is the definition and assessment of a quality model for text data. Our revision of literature indicates that adopting a perspective of predictions as quality estimations (see Subsection 1.7.2) and the use of validated representations and methods within the domain (see Subsection 1.3.2) are the central theoretical elements for model-based quality assessment for unstructured data. We now discuss practical aspects of the implementation of these tools. For these technical considerations we use the SEAHORSE summarization quality dataset [135]. Our aim is to train cost-effective lightweight models that can be used as automatic-metrics for the language dimensions of *comprehensible*, *repetition* and *grammar*, and which can serve as an example of measurement tools for unstructured data types strongly associated with human evaluation.

Chapter summary

We use as base model the case-sensitive multilingual version of DistilBERT operating with up to 512 token sequences. In order to do this, a revision of the DistilBERT and BERT family training approaches for hyperparameter tuning and layer management is conducted. The considered setting approaches for the base model are direct feature extraction, transfer learning by fine-tuning, and Parameter Efficient Fine-Tuning (PEFT) through Low Rank Adaptation (LoRA). For variations of the classification head module, the fusion of different sentence representations is explored. The relative performance differences among variants are verified to be significant by using the DeLong ROC-AUC statistical test with Holm correction. Given the class imbalance of the dataset, we make use of a weighted loss cross-entropy optimization target, together with the inspection of imbalance robust metrics. Finally, datasets available for external validation are identified. a review of the benchmark reference model mT5 is presented. we also discuss the role of AI models as data quality tools.

2.1. Introduction to the SEAHORSE dataset

The Seahorse dataset from the Natural Language Processing domain contains 96,645 instances composed of a text and its annotations for six text quality dimensions that are relevant for the specific task of summarization [135]. The dataset has the advantage of being multilingual, including texts in the languages English, German, Russian, Spanish, Turkish, and Vietnamese [135].

The texts correspond to article summaries taken from four different source datasets, whose topics are news (*XSum*, *XL-Sum*, and *ML-Sum*) and instructional guides (*WikiLingua* *WikiHow* articles) [135].

The variety of texts of SEAHORSE was obtained using two methods [135] in relation to the complete original articles: (1) Taking a human produced reference summary available in the source datasets and (2) producing a summary through one of eight automatic summarization models, each one of them having different complexity and validated performance to ensure a variety of observations even when summarizing a same base text.

The rows in the dataset contain these human or machine-generated summaries, together with the human-made annotations for the six quality dimensions, and an identifier to find the original, complete articles in the source datasets [135].

2.2. Quality dimensions for non-structured datasets

In the SEAHORSE dataset each instance has annotations for six quality dimensions. we report the definition dimensions as defined by the dataset creators in [135]:

Q1 comprehensible: The summary can be read and understood by the rater.

(If "No," the rest of the questions will be skipped.)

Q2 repetition: The summary is free of unnecessarily repeated information.

Q3 grammar: The summary is grammatically correct.

Q4 attribution: All the information in the summary is fully attributable to the source article (...).

Q5 main ideas: The summary captures the main idea(s) of the source article.

Q6 conciseness: The summary concisely represents the information in the source article. [135, p.3]

The possible values for annotations of these dimensions are "Yes", "No", and "Unsure", where "Yes" means that the text satisfies the quality dimension [135].

The authors of SEAHORSE explain that the dimension annotation was conducted consid-

ering dimensions in two groups [135]. In the procedure described by [135], the annotators were presented first with the isolated summary of a hidden article, and asked to annotate the satisfaction of the dimensions *comprehensible*, *repetition* and *grammar*. Afterwards, both the summary and the complete referenced article were made available, and annotators were asked to evaluate the satisfaction of the *attribution*, *main ideas* and *conciseness* dimensions [135]. This division of dimensions into two groups presents *comprehensible*, *repetition* and *grammar* as inherent language quality dimensions, and *attribution*, *main ideas* and *conciseness* as content referential dimensions.

These three referential dimensions can be associated with specific tasks studied in NLP research: *attribution* can be associated with Fact extraction and verification, [136] as well as with Natural Language Inference (NLI) and consistency [137], *main ideas* can be linked to Semantic Textual Similarity (STS) [138], and *conciseness* corresponds to Content Selection [139]. These research topics have been explored both directly with domain driven methods, and specialized datasets and benchmarks, and indirectly as a result of being part of general annotated datasets used for summarization research [31, 135].

From now on, with the purpose of removing ambiguity and better capturing that annotations indicate dimension satisfaction, we will rename the dimensions *comprehensible*, *repetition*, and *grammar*, as *comprehensible*, *repetition-free*, and *grammatical*, respectively. The scope of this work will consider these three "inherent" language quality dimensions for two reasons: (1) because these dimensions are relevant in a diversity of use cases beyond summarization, and (2) because the excluded referential dimensions are strongly associated with the very specific task of generative text evaluation.

Noteworthy in any case, is the fact that the document-summary relation can be described in terms of ground truth accuracy [39, 81], but considering that in the typical use cases of referential assessments both data objects are available, the used metrics for these referential dimensions have more a role of achieved task performance than a role of estimated quality for the data owned by an organization.

The estimation of accuracy in these cases is better aligned with the interdisciplinary Natural Language Processing and Artificial Intelligence research topics of Commonsense and Natural Language reasoning [140, 141], Factuality [142], and particularly with Retrieval Augmented Generation [143], whose application in data quality could be conceptualized as the revision of the content of text instances by virtue of a trusted knowledge base. These topics represent possible paths for future research in the data quality discipline.

SEAHORSE as Meta-metric

The SEAHORSE dataset release contains splits for Training, Validation, and Testing [135]. The creators explain that provisions were taken to avoid leakage of test samples, in order to not overestimate the performance of models integrating additional data from the source datasets [135]. SEAHORSE can be used as a meta-evaluation benchmark (See Section 1.7): It provides the resources for training and evaluating the quality of a scorer model, as well as the quality of the metric that it computes. A model trained on SEAHORSE and showing high performance in predicting human annotations, can be used confidently to evaluate the quality of text of generative summaries and the corresponding summarization models [135]. Since we are focusing on dimensions applicable in the general domain, the metrics we train are by design generally applicable. Nonetheless using a distribution of train data constrained to article summaries, causes our metric to be of local-type without any guarantees of preserved performance in new contexts [2, 91]. This, in the consumer-oriented view of unstructured data proposed by [3] means that we may expect low interpretation capabilities in our algorithmic metric consumer for data covering topics different to news and tutorial articles.

Procedure

Model training and validation considered the following stages:

1. Data preprocessing: Apply necessary data manipulation steps for adapting the dataset to the selected model type, preserving the benchmark splits.
2. Initial Exploration: For each dimension and model variant, train a classifier using default and conservative hyperparameter values and inspect validation set results. Select models for focusing subsequent efforts.
3. Hyperparameter tuning: Apply a model-specific evidence-supported heuristic. Following [132], we set the learning rate to minimize validation loss, and then consider regularization.
4. Use a statistical significance test to study model superiority. For ROC-AUC, we use DeLong tests with Holm correction.
5. Internal Evaluation: Evaluate models using the SEAHORSE test-splits. Compare results with those reported by the dataset authors using mT5.
6. External evaluation: Evaluate models using different datasets to study whether the metric is locally constrained or has wider generalization potential.

2.3. Benchmark performance metrics

The SEAHORSE authors use two metrics to measure the performance of models on the binary classification problem of predicting human annotations [135]: the Pearson Correlation Coefficient (ρ) and the Receiver Operating Characteristic – Area Under the Curve (*ROC-AUC*). These two metrics have the property of being threshold-independent [135]. Consequently, while tuning a higher or lower threshold to be used by the classifier to declare positive class can improve metrics such as *Accuracy*, *F1*, *TPR*, or *FPR*, the two selected comparison metrics for scoring SEAHORSE as a benchmark are unaffected by this tuning [135]. However, the authors of SEAHORSE indicate that tuning should be done when using the classifier in a specific use case, requiring data representative of the class distributions in a particular context [135].

Receiver Operating Characteristic - Area Under the Curve

The ROC curve is obtained by varying the threshold of a binary classifier to predict the positive class in the interval $[0,1]$ and plotting the points in the 2D space where the x -axis is the False Positive Rate and y -axis is the True Positive Rate [144]. The area under the ROC curve is a metric robust to class imbalance that lies in the interval $[0, 1]$, where values below 0.5 indicate worse-than-random performance and 0.5 indicates a discriminative power no better than random guessing [144].

The ROC-AUC metric has a solid probabilistic interpretation that makes it robust for evaluating classifiers independent of any specific decision threshold [145]: it measures the probability that, for two randomly selected instances — one positive and one negative — the positive instance receives a higher prediction score. This is equivalent to a Mann-Whitney U statistic, testing whether positive instances tend to receive higher scores than negative ones, without assuming any particular score threshold [145]. Hence, ROC-AUC evaluates the ranking capacity of classifiers [145].

We show next the definitions and formulas involved in the computation of this metric in Table 2.1, Equation (2.1), and Equation (2.2).

True label	Predicted label	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Table 2.1: Confusion matrix for binary classification.

$$\text{Sensitivity} = \text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{n_+} \quad (2.1)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{\text{FP}}{n_-}. \quad (2.2)$$

Pearson Correlation Coefficient ρ

The Pearson correlation coefficient measures the strength of the linear relationship between two quantitative variables X and Y [146]. We show the general definition in Equation (2.3), the formula for its estimation from samples of X and Y in Equation (2.4a), and an equivalent expression in terms of standardized values in Equation (2.4b).

Values lie in the interval $-1 \leq \rho \leq 1$, where 0 indicates no relationship, positive values indicate that variables tend to vary together, both falling below or above their respective means, and negative values indicate an association in the opposite direction [146].

When computing Pearson correlation, it is also important to perform a statistical test adequate for the data sample size and distribution, which ideally yields a near-zero p-value, estimating the probability that the correlation observed is due to chance [147].

$$\rho = \frac{\text{cov}(X, Y)}{s_X s_Y} \quad (2.3)$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2.4a)$$

$$= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_X} \right) \left(\frac{y_i - \bar{y}}{s_Y} \right) \quad (2.4b)$$

where x_i, y_i are paired observations, $\bar{x}, \bar{y}$ are sample means, n is the sample size, and s_X, s_Y denote the standard deviations of variables X, Y .

A threshold-dependent metric based on the Pearson Correlation Coefficient is the Matthews Correlation Coefficient (MCC), which measures the correlation between the true and predicted binary labels [148]. MCC ranges from -1 to 1 , where 1 indicates perfect prediction, 0 is equivalent to random guessing, and -1 represents inverse prediction.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (2.5)$$

2.4. Strategies for handling class imbalance

The three considered dimensions (*comprehensible*, *repetition-free*, and *grammatical*) present notable class imbalance. After importing and preprocessing the dataset, in all three cases, the training data has for each positive sample more than 7 negative samples.

When considering data objects in quality categories, class imbalance is often expected. The reason for this is that organizations with quality-aware processes invest in reducing deficiencies. Therefore, in operational databases, instances showing bad quality are expected to be exceptions and not the norm.

In the case of text, class imbalance can produce a bias in models towards the majority class [149]. Moreover, in imbalanced data problems, metrics such as accuracy are not reliable because under-performance on minority classes can be compensated by high performance in the majority class, producing deceptively high indicators. We discuss next methods to address class imbalance and additional performance metrics that we use in our analysis.

2.4.1. Robust metrics for classification

- **F1:** Corresponds to the harmonic mean of precision and recall (See Equations (2.1) and (2.6a)), being a better alternative to accuracy but biased towards the positive class because False Negatives are not considered [144, 150].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (2.6a)$$

$$\text{F1} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.6b)$$

- **Specificity / True Negative Rate:** Measures how well the classifier is able to correctly identify samples from the negative class [144]. The metric is useful in domains where class-imbalanced distributions characterize the typical observed data, requiring metrics that are not biased on majority classes. In such cases, learning to predict the minority class is critical, for example, in fraud detection and medical diagnosis.

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{\text{TN}}{n_-} \quad (2.7)$$

- **Youden's index / J-statistic:** This metric has values in the range $[0, 1]$ and is defined as the sum of the rates of correctly predicting a positive sample as positive and a negative sample as negative [150]. Given that the metric is built using rates, both classes are treated as equally important. The original intuitive formulation is presented in Equation 2.8a, while an equivalent formula often used for its computation due to using the same inputs as ROC-AUC is shown in Equation 2.8e.

$$J = \text{Sensitivity} + \text{Specificity} - 1 \quad (2.8a)$$

$$= \frac{\text{TP}}{\text{TP} + \text{FN}} + \left(\frac{\text{TN}}{\text{TN} + \text{FP}} - 1 \right) \quad (2.8b)$$

$$= \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN} - (\text{TN} + \text{FP})}{\text{TN} + \text{FP}} \quad (2.8c)$$

$$= \frac{\text{TP}}{\text{TP} + \text{FN}} - \frac{\text{FP}}{\text{TN} + \text{FP}} \quad (2.8d)$$

$$= \text{TPR} - \text{FPR} \quad (2.8e)$$

2.4.2. Weighted Loss

Techniques to manage class imbalance typically involve transforming the dataset or adapting the learning process [151, 152]. One option is to perform undersampling of the majority class, oversampling with replacement from the minority class, or creating new minority class instances by introducing perturbations that do not alter the class membership, a method known as Synthetic Minority Over-sampling (SMOTE) [151]. These SMOTE variations can be applied straightforwardly in numerical data using validated algorithms where a convex hull can define safe interpolation boundaries, and research has also been conducted in the field of Natural Language [153].

A second option is to use cost-sensitive learning, for example, by assigning different misclassification costs and giving observed instances an importance based on their class [152]. This approach is the one selected in the current work for model training because it increases the importance of the minority class while also preserving the diversity of data belonging to the majority class. More precisely, to scale the error for each sample depending on its class, the inverse frequency weights $\tilde{w}_k$ were computed, and then final weights w_k were obtained by applying normalization, as shown in Equation (2.9). We use the *Cross-Entropy Loss* as the base loss function.

$$\tilde{w}_k = \frac{1}{n_k}, \quad w_k = \frac{\tilde{w}_k}{\sum_{j=1}^K \tilde{w}_j} \quad (2.9)$$

2.5. DeLong test for ROC-AUC statistical difference

Before asserting that a model outperforms another in terms of ROC-AUC performance, it is important to account for possible variations in the metric that may arise purely by chance and not from actual superior prediction capabilities. For this verification, we use the DeLong non-parametric statistical test for correlated classifiers [154, 155]. This type of analysis is widely used in domains such as medicine [156, 157, 158]. Moreover, Holm correction is used to preserve confidence when multiple tests are performed [159, 160, 157]. This statistical comparison and validation is a fundamental part of Machine Learning model analysis that should not be overlooked [161].

The DeLong test is particularly convenient for our purposes, as it was specifically developed for the statistical comparison of Receiver Operating Characteristic (ROC) curves [154]; however, other types of tests, such as Student’s t-test, McNemar, and Wilcoxon can be used in tasks from the Natural Language Processing domain [162].

2.6. Model architecture selection

The analysis of unstructured data requires the application of methods that can convert data objects to standard meaningful representations fit to be processed by algorithms. Considering that Transformer encoder-only models are the state-of-the-art approach for text classification (see Subsection 1.3.2), we select a model designed with this architecture. We next discuss the reference model mT5 and our selected model DistilBERT.

The *mT5* reference results

The authors of SEAHORSE performed experiments with the dataset using models of the mT5 family, where each size configuration was fine-tuned and assessed on the task of predicting each of the six dimensions [135]. The mT5 model is presented in [163] and is described as the evolution of the earlier Text-to-Text Transfer Transformer (T5) model, trained on massive data and for multilingual tasks. It implements the Encoder-Decoder architecture, being suited for finetuning in tasks that can be translated into generation problems [53, 163]. The mT5 model has variants in several sizes [163]. The SEAHORSE authors use the Large and XXL variants having 1.2B and 13B parameters respectively [135]. The mT5 creators explain that the model is defined by:

a unified “text-to-text” format for all text-based NLP problems. This approach is natural for generative tasks (such as machine translation or abstractive summarization) (...). It is more unusual for classification tasks (...) trained to

output the literal text of the label (...) instead of a class index. The primary advantage of this approach is that it allows the use of exactly the same training objective (teacher-forced maximumlikelihood) for every task, which in practice means that a single set of hyperparameters can be used for effective fine-tuning on any downstream task. [163, p.2]

The text input for the SEAHORSE finetuned versions of mT5 uses a format where the model receives the complete article and summary identified by the prefixes “premise:” and “hypothesis:” [135].

Consistent with [53], the mT5 creators highlight that:

Since mT5 is a generative model, it can output arbitrary text predictions in a free form fashion. This is in contrast to “encoder-only” models like mBERT and XLM(-R) that make a prediction by either extracting it from the input or producing a class label. [163, p.7]

A noteworthy disclaimer expressed in the dataset document regarding the mT5 benchmark is that it is intended as a feasibility proof and not as an exhaustive aim for a best-performing model [135]. On this, they mention: "Note that our goal is not to train a state-of-the-art metric but rather to evaluate the utility of SEAHORSE as a resource to train and evaluate such metrics" [135, p.7].

The most interesting aspect of using mT5 is that, because it is a generative model, it can be trained to fulfill the two tasks of producing summaries and performing the evaluation of the quality of those summaries, on the only condition that suitable data is available for learning each of the tasks [135, 163, 53].

The use of mT5 can be beneficial to study and leverage correlations between dimensions, and learn complex patterns involving several data objects through a single structured prompt [135, 163, 53]. In contrast, BERT models follow a different approach for processing separated elements as their design and pretraining allow passing as input separate texts (called sentences) using a special [SEP] Token [61].

2.6.1. DistilBert

The central criterion for selecting a model in the current work is that it should offer computational efficiency while also showing a substantially high level of performance. This criterion aims to ensure utility and minimize adoption barriers for potential users. Considering these requirements and the high performance observed in encoder-only models, we select DistilBERT. The DistilBERT encoder-only architecture is presented in [164]

and implements the following principle to replicate the quality of the popular base BERT model:

Knowledge distillation is a compression technique in which a compact model - the student - is trained to reproduce the behaviour of a larger model - the teacher - or an ensemble of models. [164, p. 2]

The creators highlight that the base DistilBert model has 40% fewer parameters, retains 97% language understanding capabilities as measured through performance in the GLUE Benchmark, and achieves 60% faster inference [164]. The developers present this family to the community as light and fast, general-purpose, flexible pre-trained models suitable for training and fine-tuning, even under limited computational budgets, also remarking its capability to operate in real time on edge and mobile devices [164]. The authors also note that further benefits are derived from these technical results in reducing environmental impact and supporting a wider adoption of language models, helping the progress of applications [164].

Several variants of pretrained DistilBert models have been made publicly available in HuggingFace [164]. For the current work, we use the *distilbert-base-multilingual-cased* version, which is reported to have been trained on the concatenation of Wikipedia in 104 languages and is twice as fast at inference than the multilingual version of BERT [164]. This model has the advantage of upholding the multilingual diversity of SEAHORSE.

The need of efficient models in DQ

The role and importance of automatic metrics and meta-metrics for data quality differ greatly from how they are used and valued in Artificial Intelligence. Indeed, developing and applying state-of-the-art automatic metrics and meta-metrics, even if expensive, is seen in the field of AI as a worthy effort that can facilitate the development of novel systems by enabling quality evaluation of structured and unstructured data outputs in scenarios where the costs of human annotation are not affordable or the time latency associated with obtaining such annotations through rigorous procedures is detrimental to the interests of model commissioners [1].

While quality estimations for comparing data creation models are critical, given their role as objective functions to measure competitive systems, it is reasonable to think that DQ practitioners can pay a performance-usability trade-off; A DQ perspective goes beyond the creation of data, and these metrics would simply represent additional quality estimation tools for supporting assessment and improvement procedures. Moreover, DQ practice in real-life systems requires tools and concepts to be open to operate on both human and

machine-generated data under its founding principle of fitness for use.

Even if specialized hardware as GPU or TPU equipped computers, and dedicated organization owned or cloud-accessible AI workstations are typically available in companies working in the Industry of Artificial Intelligence, and even often in places dedicated to the general field of Information Technologies, the creation of models fit for low-resource devices at the cost of showing lower performance, remains a key aspect to make these systems available to a wider number of people and organizations. The task of finetuning a complex model for a specific context or deploying it as-is when considered fit already often implies specific knowledge and economic costs that can become an entry barrier.

Under these previous considerations, in this current work, without discouraging the development of powerful and novel tools imposing restrictive requirements but offering significant value, we advocate for the creation of light models and tools that can be adopted by data quality practitioners to perform assessments of subjective data quality dimensions. These models must uphold the statistical validation promise that the results produced are consistent with human evaluations of the dimensions. These tools should be accessible, fit to run in low-resource devices, and easy to integrate into DQ measurement pipelines.

In scenarios where the primary objective of the automatic assessment component is high performance at low inference latency but within lightweight size to run on resource-constrained hardware, and where computational resources and time do not constitute limiting factors for training, an alternative to fine-tuning a general-purpose pre-trained model can be the construction of a specialized model by distilling knowledge from a domain-proficient large model and leveraging a domain-specific corpus [164]. For example, [165] presents DistilBERT adapted for the Italian language, and [166] introduces several lightweight transformers, including DistilBERT, trained on biomedical data.

It is expected that creating and using such metrics for subjective quality dimensions will show the same issues that affect the repurposing of traditional metrics and taxonomies. To prevent this, it remains fundamental to communicate with clarity what is being qualified and at which granularity level (as discussed in Subsection 1.1.4). Some ways to do this could involve extending current standards and frameworks, or including rich metadata.

Tokenizer and casing

In text processing pipelines, tokenizers play a fundamental role as explained in [167]:

tokenization generally refers to the process of breaking up sequences of symbols into subsequences that can be represented as units (...). The importance of subword tokenization for language models (...) has grown (...) remaining the only major

step not fully integrated into widely used end-to-end neural models. (...) Tokenizers offer the ability to train language models over an open vocabulary [167, p. 1]

In practice, tokenizers provide mappings of natural language into the limited vocabulary sets of inputs that can be used by specific models [167], and frequently include preprocessing steps as "normalizing operations, such as lowercasing, stripping accents, removing punctuation, or uniformizing whitespaces" [167, p. 7]

Regarding the selected model differentiation of uppercase and lowercase characters, this aspect supports learning the grammar dimension due to rules such as capitalization at the start of paragraphs or proper nouns. Moreover, languages can use capitalization in different ways, for example, in German where all nouns are required to be capitalized. Even if the comprehensible and repetition-free dimensions could reasonably be learnt with a case-insensitive model, to date, the selected multilingual version of DistilBert has no publicly available pretrained uncased counterpart.

Even if we can proceed for all three dimensions using the mentioned case-sensitive model, it is important to remark that, in general, the aspect of casing has been shown to affect performance in different natural language tasks [168][169] [170].

It is also clear that casing plays a role in the real-world human readability of texts [169] [171], and that the use of uppercase can convey information on a special role of words [170]. One related practice that contributes to the construction of robust models is expanding the dataset by keeping all the initial samples and adding new derived instances with introduced casing alterations [170].

In the development of the current work, we use the library tokenizer that is paired for use with the original selected model, and therefore, the original casing of texts is preserved.

Sequence length

An important aspect for tokenization is that models are designed to process a specific maximum number of tokens. For DistilBert this limit is a sequence length of 512 tokens, which is sufficient for SEAHORSE. For tasks and datasets that involve longer sequences, a different model should be considered. In cases of extremely long texts, it becomes necessary to use specialized model architectures as [172] and [173], or to adapt a standard model following a chunking approach [174].

Considerations on data augmentation

Beyond simple casing alterations, in-depth reviews of data augmentation methods for training Language models have been published as [175] and [176], describing a diversity of transformations from simple perturbations to complex rewritings. However, a key issue in data augmentation for classification is the requirement that variations performed to obtain new instances from original data must preserve the class labels [176].

Regarding the augmentation of training data, while potentially beneficial, we do not apply it because guaranteeing a label consistency that can lead to better test performance metrics is non-trivial. Moreover, any attempt for augmentation should consider separately each of the three dimensions.

2.7. Hyperparameters for training

Guidelines for training DistilBert

A complete review of fine-tuning strategies for the DistilBert model in the task of text classification is [132]. In the document the authors study three core hyperparameters [132]: (1) Learning Rate, which scales the learning movements step size, affecting training speed and stability, (2) Batch Size, corresponding to the number of samples used for gradient estimation during each update, that influences memory usage and convergence, and (3) the Number of Epochs, indicating the number of passes or iterations over the complete training dataset.

The document [132] studies the effects of varying these hyperparameters when optimizing the metrics of Accuracy, F1 and Loss, a problem where each of these inputs and outputs cannot be seen in an isolated way, but instead is part of a system of complex interactions that can reveal cross-effects, tradeoffs, and inconsistencies. While minimizing train and validation loss is useful to combat underfitting and extract the highest model capacities, single performance metrics are preferred in early development [132]. By doing this careful joint analysis, the authors of [132] present their fine-tuning recommendations as follows:

- a) Initial Selection:
 - Batch Size and Epochs: Prioritize (...) to ensure a solid foundation, improving accuracy and F1.
- b) Final Adjustments:
 - Learning Rate: Carefully refine to capture incremental gains in loss and incremental performance in F1.
- c) Trade-Offs and Cross-Impact:

- **Joint Monitoring:** When optimizing one metric, (...) monitor the others (...).
 - **Regularization:** (...) prevent overfitting and ensure that reducing loss does not harm overall performance.
- d) **Interactions and Non-Linearities:**
- **Explore Interactions:** Adjust epochs and batch size jointly to optimize F1.
 - **Normalization and Regularization:** Reduce variability at extremes and enhance model robustness. [132, pp. 8-9]

Learning rate policies

Early warmup: Learning rate is increased progressively from zero to a target value, creating a period where learning happens by doing small, cautious steps [177]. It results beneficial, due to the high magnitude in numerical changes of the first steps associated with the randomness and expected reduced capabilities of initial weights [177].

Scheduled decay: The Learning rate is decreased from an initial value to allow for precise refinements in later learning steps, following a reduction policy, for example, as a linear or cosine function that gets to zero by the end of training [177] [178].

Regularization mechanisms

Weight Penalty: Model training uses optimizers that help the learning task, and these optimizers can include regularization in their design, as in the extensively used *AdamW* optimizer, which uses a weight decay parameter [179, 180].

Dropout: The architecture of BERT and DistilBert applies Dropout [61, 164], a technique that reduces overfitting based on the idea of randomly turning-off units, helping generalization by discouraging the co-adaptation of multiple neurons to learn instance-specific details [181]. Because each training step uses specific sets of active and disabled neurons, the final model can be seen as an average of many thinned networks [181].

2.7.1. Considerations for BERT architecture models

Transfer learning

Transformers are strongly linked to the transfer learning technique, which is based in "transfer parameters or knowledge from one trained model to another" [182, p. 1]. This broad principle includes an extensive taxonomy of problems, model architectures and lines of research as noted in [182]. For the current case of developing a binary classifier, the approach of sequential transfer learning is convenient [182]. In this approach, a pretrained

model can be fine-tuned, changing some or all of its parameters across layers to learn a new function that satisfies a new task, and it is also possible to extend the pretrained base model with adapter modules [182].

The training decision of how many layers from the pretrained base model to tune and how many to keep "frozen" has received interest as it deeply affects both performance and training-time complexity costs [183]. A study focused on the base and large versions of BERT and RoBERTa concluded that, in general, retraining the first half of these models is unnecessary, and in the case of large models, tuning only a fourth of the layers retains 90% quality [183]. The reason for this is a degree of over-parametrization and redundancy, and an architecture that captures different aspects in different layers [184] [183]. Being DistilBERT, a compressed small model, experimenting with tuning half the layers can be an efficient but conservative approach.

Sequence representation: CLS Token and Pooling methods

To perform classification, the classification head depends on receiving a meaningful representation of the input sequence.

By design, BERT produces a special vector known as the CLS token which can be directly used in classification [61] [185]:

The first token of every sequence is always a special classification token (*/CLS/*). The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks [61, p. 4]

Even if CLS provides good performance in many tasks, literature has explored alternative ways to get representative vectors [186]:

Transformer-based models produce a sequence of token-level embeddings, which are typically aggregated into a single vector that captures task-relevant information. This operation, commonly referred to as pooling, (...) directly affects the content of the final representation and thus the model's performance on downstream tasks [186, p. 2]

In this way instead of being limited to the CLS Token, research and practice show that multiple embedding vectors can be compressed by choosing only one or aggregating all with operators such as average, sum, or max [186]. The authors of [186] review pooling methods and explain that:

results support our theoretical analysis: no single pooling method is optimal across all tasks, and the best strategy depends on the task's contextual requirements. For

tasks requiring global context, such as classification and semantic similarity, global pooling methods like Average or Sum significantly outperform Last-token pooling. Conversely, Last-token pooling yields superior performance in next-token prediction tasks. These trends hold across models [186, p. 8]

Fusion Schemes

A classification head can achieve higher performance by incorporating multiple sequence representations through fusion schemes [187]. The four schemes presented in [187] can be summarized as follows: (1) *aggregate all*, which pools all tokens, treating the special CLS token and the word tokens equally, (2) *concatenate*, based on joining the CLS token with a pool vector computed from word tokens, (3) *sum*, which keeps separated CLS and a pooling of word tokens as two representations, connecting each one to a fully connected layer, and then sums the two outputs, and (4) *late*, where CLS and pooled word tokens stay further separated to compute individual classification scores.

2.7.2. Alternatives to fine-tuning

Feature extraction

Scientific works related to BERT are consistent in recognizing that different layers convey different text properties [61] [188] [189] [185] [190]. Key findings reported in [185] are the following:

Syntactic information is most prominent in the middle layers (...). The final layers (...) are the most task-specific (...). In fine-tuning, it explains why the final layers change the most (...), semantics is spread across the entire model [185, pp. 6-7]

Retraining both base model and classification head in fine-tuning shows solid performance due to learning features that matter in specific tasks [61] [188]. Nonetheless, it is possible to successfully use the base model original features in a diversity of problems [61] [188].

In this "feature-based" approach, we do not properly perform fine-tuning, but instead use the base model as a feature extractor [61] [188]. The core principle in this approach is to exploit the fact that distinct types and granularities of information are distributed across different network layers, by extracting and utilizing embeddings not only from the final layer but from multiple intermediate layers as well [61] [188]. The main advantage of such an approach is its lower computational cost, since there is no need to retrain base model parameters [61].

PEFT-LoRA

As an alternative to traditional fine-tuning, different Parameter Efficient Fine-Tuning (PEFT) methods have been proposed to facilitate the adaptation of pretrained models in new tasks by reducing the number of trainable parameters, lowering time and memory costs. [191]. One of such methods is *Low Rank Adaptation* (LoRA), where pretrained model parameters are frozen and training is redefined as a problem of optimizing injected rank decomposition matrices [192]. LoRA can reduce GPU memory requirements up to three times, preserving high performance, and not increasing inference time, given its solid mathematical foundation that makes it possible to merge LoRA-trained matrices into the base model weights [192]. LoRA is a validated method both for large and small language models, and it can even act as a regularization mechanism, valuable specially in cases of limited data [193] [194].

2.7.3. Comparison datasets

The SEAHORSE authors use for reference the Multilingual Summarization with Factual Consistency Evaluation (mFACE) dataset [135, 195, 196]. An important remark is that it is not always possible to map a direct identification of dimensions across datasets, making necessary to review the datasets and their label dimensions to find similarities [135].

Our search found the following datasets to be other possible sources for external validation:

- Corpus of Linguistic Acceptability (CoLA) [197, 55]
- Italian CoLA (ItaCola) [198]
- Norwegian CoLA (Nocola) [199]
- Multilingual Evaluation of Linguistic Acceptability (MELA) [200]
- Grammatical versus UnGrammatical (GUG) [201]
- JHU FLuency-Extended GUG corpus (JFLEG)[202]
- Benchmark of Linguistic Minimal Pairs for English (BLiMP) [203]
- Boilerplate and Redundancy Evaluation on Assorted Documents (BREAD) [204]
- PARaphrase identification based on Domain knowledgE (PARADE) [205]
- Quora Question Pairs (QQP)[55]
- Microsoft Research Paraphrase Corpus (MRPC) [55, 206]

- Cross-lingual Paraphrase Adversaries from Word Scrambling (PAWS-X) [207]

3 | Implementation and results

Tools: The main libraries used in this work were Pandas and HuggingFace Datasets for data manipulation, HuggingFace Transformers (PyTorch backend) and peft for model development, and statsmodels and MLstatkit for statistical significance tests.

3.1. Data preparation

Before training models we perform several preprocessing procedures:

1. Load the dataset, which is already divided into train, validation, and test splits.
2. Drop non-used columns and rename those relevant for readability: question1 to comprehensible, question2 to repetition_free, question3 to grammatical, summary to text, and worker_lang to language.
3. Remove rows with exact duplicates.
4. Check for rows with a null/None text field. There are no null texts.
5. Check for rows with null/None annotations. There are no None values for comprehensible, while for repetition-free and grammatical, given that only comprehensible texts were annotated for the other dimensions, the functional dependency $comprehensible = 0 \rightarrow (repetition_free = None \wedge grammatical = None)$ holds. No necessary modifications.
6. Replace No/Yes String target labels with Integers 0/1. Rows with Unsure ratings are dropped (the same policy is followed by dataset creators in their model training).
7. Because a pattern of blocks by language is observed we shuffle the data.
8. Filter the subsets of data with comprehensible=1, to be datasets for repetition-free and grammatical dimensions.
9. Inspect sequence lengths. In all subdatasets 512 Tokens is above percentile 98. Therefore DistilBERT is appropriate for handling the extension of texts. Longer instances are truncated at inference time.

The previous steps were applied to training and validation data. For the test set, the minimal required preprocessing steps (1 and 6) were applied in order to keep fidelity to the benchmark.

A noteworthy remark is that the SEAHORSE authors [135], in addition to the split benchmark dataset, also published a subset of data with multiple annotations, which they use in annotator-agreement analysis; In SEAHORSE, the authors are explicit in saying that this additional set of annotations must not be included in training because of possible test examples leakage. However, in general, when working with gold standard datasets that have multiple rater annotations, a better alternative to our step 3 policy should maximize estimated accuracy by using methods such as weighted or majority vote (See subsection 1.6.3).

3.2. Model Training

Environment configuration

We installed Nvidia CUDA to run experiments on the GPU device. Then, for reproducibility, libraries were set to use fixed random generation seeds and deterministic algorithms. Hardware specifications are reported in Appendix Section A.1.

Initial exploration

An initial phase of exploration to see whether any configuration appeared as superior for focusing later efforts was performed by training models with library defaults and conservative values for hyperparameters. In each experiment, the checkpoint for the epoch of highest validation ROC AUC was selected. This metric is threshold independent and allows comparison with the benchmark. We trained for 4 epochs and a batch size of 32, which showed to produce good gradient estimations and notable metrics within our computational constraints. Indeed, all of our trained predictors obtained ROC AUC values above 0.74. The considered models are explained in Table 3.1 and visualizations of the architectures are presented in Appendix Section A.6.

Approach	Description
Fine Tuning (library default)	The transformers library class <code>DistilBertForSequenceClassification</code> provides a simple straightforward model to fine-tune DistilBERT for downstream tasks. It consists of DistilBERT connected to one hidden layer and an output layer using the standard [CLS] token as representation. We consider tuning half of the transformer layers (3/6) as a baseline.
Feature extraction	Training of a classifier head that takes as input the [CLS] and average pooling sequence representations of the last 3 transformer layers. Advantages are low training cost and containing different higher and lower level features [61].
Concatenation and Sum fusion schemes	Fine Tuning of the last 2 or 3 transformer layers using classification heads that combine the [CLS] and Average Pooling representations of the input sequence as described in [187]
Low Rank Adaption (LoRA)	Efficient adaption of last 3 layers

Table 3.1: Models of exploration phase

For the three dimensions, we observed that while models involving fine-tuning stopped improving metrics after the second or third epoch, models known for having fewer trainable parameters and being more robust to overfitting (LoRA and Feature Extraction) showed improvements till the end of the fourth epoch but achieved notably lower performance. Interestingly, the decision to finetune only two layers rather than three, combined with the use of fusion schemes, did not result in notable performance gains. This may reflect overfitting, or it may indicate that two layers are already capable of learning the essential features for each of the three tasks.

We also inspected threshold-dependent metrics (associated to a default 0.5 threshold), and in all cases, a specificity lower than sensitivity reflected difficulties at learning to identify the minority class even after applying weighted loss to account for imbalance. This suggests a strong need to do a decision threshold tuning before deployment that considers the context-dependent prior distributions and different error costs. The numerical results and used hyperparameters for this phase are presented in Appendix Section A.2.

Hyperparameter tuning

After the initial exploration phase, we selected as candidates for further refinements the Concat Fusion model (2 trainable layers) for comprehensible and the Sum Fusion model (2 trainable layers) for repetition-free and Grammatical. With support on the findings from [132] on the fine-tuning of DistilBERT, we followed a practical heuristic consisting of first encouraging model fit to the training data by monitoring the loss, and subsequently addressing overfitting through regularization techniques.

These types of strategies, intended to produce high performance without the complexity of considering all the involved model settings interactions simultaneously, offer no guarantee of superior results across all use cases, and it remains fundamental for practitioners to follow strategies and best practices documented to work on their specific selected models. Moreover, although ranking metrics such as ROC-AUC often improve alongside

reductions in loss, no general theoretical guarantee establishes a strict correspondence [208, 209, 145]. In the current case, this search for hyperparameters did not yield any validation ROC AUC values that surpassed the previous overall bests. Experiment results are presented in Appendix Sections A.4 and A.5.

While in this work, given computational cost constraints, a phased heuristic strategy was preferred; the issue of escaping performance plateaus can be addressed with other approaches. As an alternative or complement to relying on model-specific literature for hyperparameter tuning, a more general and exhaustive search procedure can be followed, supported by tools like the Optuna Hyperparameter Optimization Framework, which can automatically manage the exploration of large hyperparameter spaces and offers advanced optimization functionalities, including pruning of unpromising trials, sampling methods, and parallel execution [131].

Training for more Epochs Feature Extraction and LoRA

Considering the observed continuous learning in the Feature Extraction and LoRA models, we performed the experiments of training the FE model for longer (12 epochs instead of 4) and training the LoRA model for 6 epochs with only two trainable layers and aggressive learning rate. These experiments, detailed in Appendix Section A.3, resulted in very limited improvement, being still outperformed by standard fine-tuning-based models.

Testing for significant difference in model performance

In the validation set, the highest ROC for the *Comprehensible* dimension was obtained by the 2-layer-Fine-Tuned Concat Fusion scheme model. For both the *Repetition* and *Grammatical* tasks, the best performance was achieved by the default model employing [CLS]-Token-based representation, using a single hidden layer head classifier and 3 trainable transformer layers. Since the validation set was used both for identifying the best training checkpoint and for model selection, performance estimates derived from it would be optimistically biased. We therefore evaluate on the held-out test set, which remains independent.

For our analysis, in addition to considering the best validation set performing models, we make sure that in each case, three particular cases are included: the baseline 3-layer fine-tuning, which can be easily deployed, the best performing 2-layer-fine-tuned fusion scheme, with lower time cost and leveraging two representation vectors, and the Feature Extraction model, characterized by a highly reduced number of trainable parameters. At considering these models, in all three dimensions, the highest to lowest ROC AUC

difference was of less than 0.03.

Table 3.2: Comprehensible Test results

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Fine tuning (3 layers)	0.831875	0.901116	0.969686	0.841603	0.733252	0.574855	0.873211	0.470622
Fine tuning (2 layers, Concat fusion)	0.828376	0.898693	0.971124	0.836317	0.747868	0.584185	0.871819	0.467516
Feature extraction	0.840897	0.907284	0.966074	0.855238	0.695493	0.550731	0.863720	0.465166

Table 3.3: Repetition Test results

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Fine tuning (3 layers)	0.940445	0.965457	0.982547	0.948952	0.879765	0.828717	0.966784	0.790262
Fine tuning (2 layers, Sum fusion)	0.942909	0.967016	0.980152	0.954228	0.862170	0.816398	0.965419	0.792378
Feature extraction	0.898978	0.940120	0.979004	0.904207	0.861681	0.765889	0.949850	0.699082

Table 3.4: Grammatical Test results

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Fine tuning (2 layers, Sum fusion)	0.820587	0.895482	0.945069	0.850839	0.537935	0.388774	0.774605	0.341288
Fine tuning (3 layers)	0.746633	0.843485	0.954202	0.755791	0.661070	0.416860	0.773207	0.313210
Feature extraction	0.787999	0.873900	0.944350	0.813232	0.552239	0.365471	0.746517	0.297221

To determine whether differences in ROC AUC were statistically significant, we performed DeLong tests with Holm correction for each of the three dimensions, at a significance level of $\alpha = 0.05$. In all three cases, the results showed that the low-cost Feature Extraction models were significantly outperformed by the fine-tuning-based approaches. However, among the fine-tuned models themselves, no statistically significant differences were found, ruling out any claims of general superiority.

Models	Adjusted p-value	Conclusion
FT, Concat Fusion	4.213×10^{-1}	H_0 : No significant difference
FT, FE	3.066×10^{-4}	H_a : Significant difference
Concat Fusion, FE	3.138×10^{-4}	H_a : Significant difference

Table 3.5: DeLong Test with Holm correction $\alpha = 0.05$ (*Comprehensible*)

Models	Adjusted p-value	Conclusion
FT, Sum Fusion	1.0900×10^{-1}	H_0 : No significant difference
FT, FE	4.2660×10^{-24}	H_a : Significant difference
Sum Fusion, FE	2.2983×10^{-25}	H_a : Significant difference

Table 3.6: DeLong Test with Holm correction $\alpha = 0.05$ (*Repetition free*)

Models	Adjusted p-value	Conclusion
FT, Sum Fusion	5.9350×10^{-1}	H_0 : No significant difference
FT, FE	1.5346×10^{-11}	H_a : Significant difference
Sum Fusion, FE	2.2662×10^{-12}	H_a : Significant difference

Table 3.7: DeLong Test with Holm correction $\alpha = 0.05$ (*Grammatical*)

3.3. Model evaluation

3.3.1. Results for the SEAHORSE Test set

To evaluate the effectiveness of our models as quality classifiers, we compute performance metrics on the test set provided by the SEAHORSE creators. These results are also compared with those reported in the original SEAHORSE dataset publication, corresponding to the fine-tuning of mT5 models [135].

For the *Comprehensible* and *repetition-free* dimensions, our least performing models (based on feature extraction) showed ROC AUC capacities above 95% in relation to the SEAHORSE authors' best performing mT5 variant, and for Grammatical 88%. Similarly, at considering our best ROC AUC test-set scoring models, we obtain 99%, 99%, and 92%. Regarding size, our model with more parameters has a ratio of 0.12 in parameter count to the lightest $mT5_L$ model, and of 0.01 to the larger $mT5_{XXL}$ version.

All ROC AUC values are above 0.74, which shows a good ranking power that can be leveraged by using the knowledge of data distributions and context costs and meaning of

the dimensions, for setting an adequate decision threshold, as in general in these models the samples for the positive class are highly likely to be assigned a higher score than those in the negative class. While the test split of SEAHORSE shows good performance, we evaluate the model with different datasets next for further validation.

Table 3.8: Comparison to mT5 (SEAHORSE Test set)

Dimension		<i>Comprehensible</i>		<i>Repetition free</i>		<i>Grammatical</i>	
Model	Size (Parameters)	ROC AUC	ρ Pearson correlation	ROC AUC	ρ Pearson correlation	ROC AUC	ρ Pearson correlation
DistilBERT Fine tuning (3 fine-tuned layers)	135.3M	0.873211	0.470622	0.966784	0.790262	0.773207	0.313210
DistilBERT Fine tuning (2 fine-tuned layers, Concat fusion)	135.9M	0.871819	0.467516	-	-	-	-
DistilBERT Fine tuning (2 fine-tuned layers, Sum fusion)	135.9M	-	-	0.965419	0.792378	0.774605	0.341288
DistilBERT Feature extraction	139.5M	0.863720	0.465166	0.949850	0.699082	0.746517	0.297221
<i>mt5_L</i>	1.2B	0.88	0.44	0.97	0.74	0.81	0.37
<i>mt5_{XXL}</i>	13B	0.90	0.52	0.98	0.86	0.84	0.45

3.3.2. External validation

Comprehensible dimension

mFACE: The *Multilingual Summarization with Factual Consistency Evaluation (mFACE)* dataset [195] is considered by the authors of [135] for evaluating mT5 after being finetuned using SEAHORSE. Both SEAHORSE and mFACE are datasets published by Google Research and DeepMind. This dataset has in common with SEAHORSE belonging to the text summarization domain, being human-annotated, having as topic news, and having multilingual data, but beyond the 5 languages of SEAHORSE, mFACE covers 45 languages [135, 195].

While mFACE explicitly uses the word "comprehensible" during the annotation process, SEAHORSE inquires about the dimension in terms of readability and making sense [135, 195]. The authors of SEAHORSE consider it adequate to identify these two columns, given their close meaning [135]. We now present our results together with the reference results reported for mT5 in [135]. Table 3.9 considers the subset of mFACE defined by the application of a filter to only keep samples from languages present on SEAHORSE, while Table 3.10 considers the entire mFACE dataset. The behaviour of performance in our DistilBERT models appears similar to the results of mT5 in that all show a drop of more than 0.1 in ROC AUC when evaluated on mFACE [135].

A further noteworthy observation for both the presented and referenced models is the reduced but preserved capacity to operate on texts in languages not included in SEAHORSE [135]. This phenomenon, where models exhibit strong performance in languages

beyond those seen during training, is the focus of cross-lingual transfer techniques [210]. It is also important to emphasize that, although the multilingual DistilBERT [164] model is pretrained on data from 104 languages, achieving solid performance on downstream tasks still requires linguistically diverse, task-specific training data to be able to adequately capture language-specific patterns.

Table 3.9: mFACE 5 common languages with SEAHORSE results (Comprehensible)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
DistilBERT Fine tuning (3 layers)	0.842857	0.913085	0.982319	0.852967	0.539823	0.392790	0.735731	0.215162
DistilBERT Fine tuning (2 layers, Concat fusion)	0.822286	0.900639	0.981204	0.832300	0.522124	0.354424	0.730868	0.205156
DistilBERT Feature extraction	0.820571	0.899648	0.980495	0.831119	0.504425	0.335544	0.726941	0.200867
$mt5_L$							0.77	0.14
$mt5_{XXL}$							0.73	0.09

Table 3.10: mFACE all 45 languages results (Comprehensible)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
DistilBERT Fine tuning (3 layers)	0.751270	0.853483	0.960680	0.767807	0.474986	0.242793	0.661992	0.138677
DistilBERT Fine tuning (2 layers, Concat fusion)	0.708127	0.823179	0.960763	0.720063	0.508713	0.228776	0.658755	0.133781
DistilBERT Feature extraction	0.701905	0.818481	0.961879	0.712291	0.528387	0.240678	0.656241	0.140323
$mt5_L$							0.68	0.13
$mt5_{XXL}$							0.70	0.15

Repetition free dimension

BREAD: The *Boilerplate and Redundancy Evaluation on Assorted Documents (BREAD)* dataset is a remarkable high quality source that spans 360 languages and has expert annotations [204]. As drawback, BREAD has only 2000 samples [204]. The BREAD texts are derived from the bigger MADLAD-400 dataset, whose topic corresponds to web-crawled general domain data. [211, 204]. The creators explain that this dataset is related to the web-scraping issue of distinguishing linguistic content from uninteresting data, and propose two benchmarks to be scored using the F1 metric[204]: (1) BREAD-REPEAT, consisting in separating natural text from repetitive boilerplate, and (2) BREAD-NOISY, consisting in separating natural text from general boilerplate and noise texts. For our case, the subset of data used in BREAD-REPEAT serves as a source for external validation.

The authors of the dataset present as solution to this task several language agnostic metrics based on ngram frequency distributions and remark that the use of neural network

methods in the study of underrepresented languages has the important shortcomings of difficult interpretability and need for high quantity and quality data [204]. This argument is aligned with the views of [76, 36] in aiming for simplicity and rejecting deceptively universal solutions.

Considering the reduced size of BREAD, and that none of its samples were seen before by our models, we decide to use the samples from both tune and test splits. Results are shown in Table 3.16. For reference, the best performing metric from the BREAD authors gets a F1 value of 0.943 for classifying new instances [204]. This reference results outperform our DistilBERT models (assuming a default 0.5 threshold) by more than 0.09 in F1.

Regarding our models, the superior performance of the feature extraction approach may be explained by its direct access to syntactic information encoded in middle layers [61, 185], which for the task definition and data distribution of BREAD-REPEAT appear more relevant than the higher-level semantic features that can be learnt in fine-tuned upper layers.

Table 3.11: BREAD results (Repetition)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
DistilBERT Feature extraction	0.851638	0.887482	0.937243	0.842738	0.871849	0.714587	0.936411	0.749494
DistilBERT Fine tuning (2 layers, Sum fusion)	0.826590	0.883420	0.828340	0.946346	0.554622	0.500968	0.879358	0.641187
DistilBERT Fine tuning (3 layers)	0.780347	0.853345	0.795364	0.920444	0.462185	0.382629	0.813980	0.522627

MRPC: *The Microsoft Research Paraphrase Corpus (MRPC)* contains 5801 english sentence pairs with human annotations for semantic equivalence [206]. These texts were obtained from internet available news articles [206]. An important remark made by the creators of MRPC is that the annotation of a sentence pair as paraphrase followed a loose criterion [206]. This decision was intended to increase textual diversity (in contrast to constraining to very similar worded sentences) but has a tradeoff in ignoring the presence of some different content and levels of detail [206].

Although it would be technically possible to pass sentence pairs to DistilBERT explicitly separated by the [SEP] token [61, 164], we do not do this because the Seahorse repetition dimension is defined for single texts. We therefore adopt a concatenation approach, treating sentence pairs as a single joined input. Considering these original tasks and

annotation definitions, we flip the labels to align the absence of semantically equivalent texts with the no repeated information criterion used in SEAHORSE [135].

We now present our rationale for why this transformation is valid for re-purposing the data. Joining two paraphrase sentences always produces a redundant single text. The converse, however, represents a necessary but not sufficient condition: a non-redundant single text can only arise from concatenating non-paraphrase sentences, yet concatenating non-paraphrase sentences may still yield a redundant text if one or both sentences contain repetitive elements individually, or if they share some overlapping information. Fortunately, the loose annotation criterion of MRPC, which easily declares sentence pairs as paraphrase, provides some support for the implication that non-paraphrase pairs are genuinely distinct, making the non-paraphrase to repetition-free mapping more reliable in practice.

To learn the original MRPC paraphrase classification task, we would require sentence pairs data, and it would be necessary to expose the model to the [SEP] token during training for it to learn the appropriate input structure and discrimination patterns. Therefore, while our transformation of MRPC enables evaluation in terms of the SEAHORSE redundancy definition, this pipeline would be inappropriate for the original sentence-pairs paraphrase classification task.

Table 3.12: MRPC results (Repetition)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
DistilBERT Feature extraction	0.698500	0.340746	0.601064	0.237770	0.923077	0.160847	0.702786	0.293978
DistilBERT Fine tuning (2 layers, Sum fusion)	0.675401	0.438748	0.506190	0.387165	0.815897	0.203062	0.677008	0.265046
DistilBERT Fine tuning (3 layers)	0.670057	0.422799	0.495406	0.368753	0.816923	0.185676	0.656483	0.250313

We observe in the ROC-AUC values that our models retain good ranking capacity; however, threshold-dependent metrics such as sensitivity yield imbalanced results at the default threshold. In this case, well-performing classifiers can be better adapted to a new use case - characterised by a specific data distribution and task definition - through the procedure of determining a more optimal decision threshold based on estimation in a data sample. In Appendix Section A.7 we show how using the validation split to find the threshold that minimizes the absolute difference between sensitivity and specificity can serve as an effective tuning strategy, yielding balanced metrics on the test set.

PAWS-X: The *Cross-lingual Paraphrase Adversaries from Word Scrambling (PAWS-X)* is a multilingual dataset for sentence pairs paraphrasing, derived from the earlier english-only *PAWS* dataset [212, 207]. The main feature of the dataset is including challenging examples characterized by high lexical overlap but different semantic meaning [207]. and we use for evaluation the development and test-sets, which are human-annotated and translated, in contrast to the training set that involves machine translation. To repurpose the dataset for the definition of the dimension from SEAHORSE of no repeated information, we follow the same concatenation approach we used for MRPC. In Table 3.13, we present the results for languages common to SEAHORSE (English, German, and Spanish) and in Table 3.14, the results considering all languages.

Table 3.13: PAWS-X 3 common languages with SEAHORSE results (Repetition)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
DistilBERT Fine tuning (2 layers, Sum fusion)	0.455358	0.140043	0.583796	0.079564	0.928571	0.008136	0.533732	0.022461
DistilBERT Fine tuning (3 layers)	0.455948	0.138566	0.589773	0.078506	0.931238	0.009744	0.531459	0.024834
DistilBERT Feature extraction	0.446084	0.038630	0.591928	0.019967	0.982667	0.002633	0.526131	0.018420

Table 3.14: PAWS-X 7 languages results (Repetition)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
DistilBERT Fine tuning (2 layers, Sum fusion)	0.464336	0.247295	0.571764	0.157765	0.850977	0.008742	0.526573	0.018646
DistilBERT Fine tuning (3 layers)	0.465348	0.239992	0.579261	0.151348	0.861359	0.012707	0.525606	0.023242
DistilBERT Feature extraction	0.449044	0.086659	0.574722	0.046863	0.956266	0.003129	0.523394	0.012872

The reduced ROC-AUC values, close to a random guess, indicate that ranking capacity is not preserved for this dataset. Taking into account the imbalance between sensitivity and specificity at the default 0.5 threshold, and the known data distribution, it is possible that our models are relying heavily on repeated words to classify instances as non-repetition-free, rather than capturing deeper semantic redundancy.

Regarding the original sentence-level paraphrase task, the authors of [207] achieve in their experiments with BERT-base (12 layers) accuracy above 0.70 when trained on English examples, and 0.83 when exposed to multilingual data. Although the task definitions differ significantly, the architectural similarity between BERT and DistilBERT and the good zero-shot performance on unseen languages suggests that good performance is attainable given sufficient high-quality data covering challenging examples. Considering that SEAHORSE was commissioned to annotate six quality dimensions and is sourced

from news and web-guides text, it is likely that its examples present low difficulty for repetition detection, which may explain the limited discriminative capacity observed.

Grammatical dimension

MELA: The *Multilingual Evaluation of Linguistic Acceptability (MELA)* [213] dataset integrates examples from different sources to create a benchmark to identify well-formed and syntactically acceptable texts. The score used for the MELA benchmark is MCC, and the best results reported by [213] are for XLM-R trained on the MELA training set with a score of 37.72 and for GPT-4o with a prompt using in-context examples with a score of 51.14.

We evaluate our models on the subset of common languages (English, German, Russian, and Spanish), and also on the complete MELA dataset, including all predefined splits. In both cases, the values for ROC AUC and Pearson correlation show that our models fail as grammar metrics in the MELA dataset.

Table 3.15: MELA results 3 common languages with SEAHORSE results(Grammatical)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
DistilBERT Fine tuning (2 layers, Sum fusion)	0.652101	0.777414	0.721067	0.843313	0.159378	0.002691	0.516648	0.013285
DistilBERT Fine tuning (3 layers)	0.598028	0.717975	0.725890	0.710231	0.308901	0.019131	0.516182	0.022934
DistilBERT Feature extraction	0.635741	0.762285	0.719338	0.810685	0.184940	-0.004375	0.506810	0.002019

Table 3.16: MELA results all 7 languages results (Grammatical)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
DistilBERT Fine tuning (2 layers, Sum fusion)	0.677893	0.797570	0.751219	0.850016	0.171003	0.021019	0.549429	0.045536
DistilBERT Feature extraction	0.660349	0.780569	0.753855	0.809246	0.221857	0.031103	0.547713	0.050555
DistilBERT Fine tuning (3 layers)	0.635421	0.755412	0.756647	0.754181	0.285678	0.039860	0.542532	0.052952

In light of these results, it arises the question of why our models performed well in the SEAHORSE dataset. Similar to our observed results in PAWS-X for the repetition dimension, significant reasons exist to consider the lack of better examples as one source of low performance. In particular, we must recall (see Section 2.1) that in SEAHORSE the texts of the dataset were produced without considering the specific dimensions that were the subject of annotation. In SEAHORSE, variability of texts was obtained by applying different summarisation methods to articles, and quality variation was assumed to emerge

given the intentional selection of different quality summarization methods [135]. For Linguistic Acceptability datasets, the case is very different as these datasets are typically curated with linguistic expertise and instances are either derived from a domain-specific corpus, or created using syntactic rules to deterministically exhibit meaningful patterns [198, 200].

SEAHORSE can certainly be considered as a high-quality dataset for learning the quality of article summaries, as its size and distribution of texts paired with human annotations properly represent the common variations in quality associated with the article summarisation task. However it results not fit for the purpose of training a linguistically competent grammar classifier. Even if a bigger model would likely be able to extract better features from SEAHORSE to generalize prediction of grammaticality, this low data quality give some ground to believe such improvements would not be significant. These remarks highlight the importance of giving attention to data quality in the practice of model development. To further explore this, we trained our feature extraction model using the training set of MELA. With this change, performance on the MELA test set largely improves, achieving 0.19 MCC and a ROC AUC value of 0.65, proving better ability to correctly rank texts. The complete results OF this experiment are presented in Appendix Section A.28.

Empirical notes on training-times and costs

Training times are influenced by multiple factors, including device components, GPU parallelization, GPU–CPU interactions, energy use, heat, memory bandwidth, and data spilling from VRAM to slower RAM. Consequently, analyzing hardware-related metrics would require repeated runs to reduce variance. Nevertheless, we provide comments on this topic as qualitative observations specific to the device used for the experiments. For the largest experimental setup (corresponding to the Comprehensible dimension which uses more data), training times were: 6h12min for fine-tuning with one hidden layer head (3 trainable layers), 1h25min for feature extraction, 2h00min for fine-tuning with a Sum-fusion head (2 trainable layers), and 2h01min for fine-tuning with a Concat-fusion head (2 trainable layers). This relative hierarchy of computational costs was consistently observed across other experiments.

Freezing transformer layers for fine-tuning or training only the classifier head in a Feature Extraction architecture were approaches that reduced training time while also preserving high performance. Even if performing a search for optimizing hyperparameters can be resource-intensive, in our case we observed that using typical community and library-suggested hyperparameter values brought solid performance from the start.

General and popular works on quality management, such as [9], have argued that by reducing deficiencies and leveraging improvement opportunities, organizations may ultimately find that “quality is free”; Also, the data quality discipline, in landmark works such as [20, 81] and specialized research, has emphasized data quality as a worthy investment for organizations. In this perspective, our results showing performances above 0.74 in ROC AUC, requiring less than an hour and a half of training, and ran in a medium-tier personal computer, can serve as evidence that creating a viable initial automatic metric is not necessarily an expensive task, and the associated technological and expertise entry barriers can be revealed as much more accessible and user-friendly than expected.

Moreover, adopting measurement procedures for new data objects and quality dimensions, and defining an interface for these types of metrics within the organization, should facilitate transitioning to more advanced tools in case research produces a new state-of-the-art publicly available model or the organization allocates resources for in-house production, aiming for returns from customization and higher performance.

While our work focused on Natural Language, other light-sized models aimed for data types strongly linked to human perception have been introduced. Some examples are DeiT [214] and TinyViT [215] in Computer vision, and DistilHuBERT [216] and MERT [217] for audio. Moreover, their size and design associated with faster inference times can ease measurement procedures and make them flexible and suitable for latency-sensitive domains. The use of these systems, paired with their conceptual understanding in quality terms, as data quality metrics (instead of mere regressors or classifiers), and following DQ methodologies, is also a possible investment worthy of venture. This serves as an example of tuning, adapting measurement metrics and understanding of dimensions and use cases and context.

4 | Conclusions and future developments

The main findings of this work can be summarized as follows:

- Data quality is highly contextual and domain dependent. In the case of unstructured data, quality is very often associated with specific tasks [3]. A significant contribution of the data quality discipline has been introducing high and intermediate level abstractions in the form of intuitive definitions of dimensions, generic metric formulations, and aggregation operators that can be implemented in each domain and context adapting them to align to the relevant data types, business rules and objectives [22, 20, 18, 19].
- The assessment of data quality supported in questionnaires has as its main advantages being data-type agnostic and context-aware. Even if data is not evaluated directly, the contextual value is adequately represented and is illustrative of the data fitness for use. These collected judgments prove valuable to determine the state of quality, gather key information to diagnose problems, and estimate the success of improvement actions, while also showing additional instrumental value from raising quality awareness in participating members of the organization [14, 21, 67, 74, 72, 20].
- Conceptual modeling separating intentionality or meaning from data types and values constitutes a central element for building actionable methodologies independent of data-formats and domain nuances [75].
- Workflows using unstructured data benefit from the perspective of machines as algorithmic or model data consumers that interpret and process data based on past experience encoded in previously observed data [3]. When models are trained on task specific datasets, and later deployed to process out-of-distribution data, no guarantees of sustained performance exist [91]. Metrics are to be assumed local unless extensive evidence is collected [2].

- When no labels are available, clustering can be a valuable approach to study data groups in terms of quality properties [76, 99, 109].
- Adopting a perspective of predictions as quality estimations [3] and the use of validated representations and methods within each target domain are the central theoretical elements for model-based quality assessment for unstructured data.
- The starting point to define an automatic quality metric is identifying attributes that are indicative of fitness for use. The aim of automatic-metrics is learning to imitate human judgment patterns for these attributes using large datasets [1]. Datasets and benchmarks for predicting human judgments constitute meta-metrics [1]. The quality of a model implementing an automatic-metric is evaluated using traditional classification performance metrics such as ROC-AUC and Pearson correlation, which indicate the effective capacity to score better quality objects higher, and show consistency with human evaluations. For using the trained model to measure data quality, probability predictions are considered as quality estimation scores [3]. While classification can consider multiple classes, a straightforward approach is creating a binary classifier to discriminate a *satisfactory quality* class [3, 4].
- While some model architectures may exhibit initial calibration for the distribution observed during training, ensuring a robust probabilistic interpretation requires confidence estimation and calibration procedures. This step is relevant for building a data quality metric that can be analyzed with sound aggregation [5]. A conservative approach can use min and max operations to assess quality bounds at higher levels of granularity.
- While data collection can be resource intensive, it is possible to find publicly available reliable data and validated proficient models associated with benchmarks defined for common tasks. The creation of silver datasets using automatic consistent annotations can also reduce costs, but reliable validation data is necessary.
- The availability of lightweight and low-latency pretrained models suitable for low resource environments, created from techniques such as knowledge distillation and the existence of comprehensive libraries providing implementations for training, modifying, tuning and experimenting with models are factors that reduce cost and expertise barriers for developing automatic quality metrics. The creation and deployment of an initial viable metric can be beneficial for adopting measurement procedures for new data objects and quality dimensions, and defining an interface that facilitates transitioning to more advanced tools.

- High-quality data and high-performance models form a virtuous circle: quality data enables better models, which can serve as automatic metrics to quantify quality properties.

The following aspects constitute potential directions for future work:

- The translation of clustering indicators and concepts into data quality specific metrics that can be considered during measurement procedures and be analyzed to quantify the impact of improvement strategies.
- The estimation of costs of using automatic-metrics for quality-measurement, including not only model training but also the deployment, and maintenance for operating into existing organization pipelines. Cost-benefit analysis for the use of these tools could provide guidance on the convenience of further investing on more powerful models, trained and deployed either locally or using cloud services.
- The execution of case studies for obtaining empirical validation that automatic-metrics adequately measure fitness for use. Such studies could identify factors facilitating and hindering adoption.
- The identification of best practices for conducting small-scale annotation activities, that preserve the rigor of larger-scale projects but are more easily actionable. Doing these activities internally can support the transformation of operational databases into quality patterns suitable for machine learning. These small-scale activities can produce golden datasets for validation. Complementarily, the establishment of a systematic approach for obtaining massive automatic silver datasets can address the data requirements of automatic metrics.
- The study of using automatic-metrics in dynamic contexts, which may require monitoring, periodic retraining and recalibration. This can consider formulating a data quality perspective for the concepts of continual learning [218], online learning [219] and concept drift [220].
- The comprehensive investigation of the role of confidence estimation and calibration in the aggregated analysis of automatic-metrics. This can include the articulation of a systematic calibration approach, or the identification of effective methods that formally satisfy the data quality metric requirements proposed in [5].
- The study of referential quality, defined over object pairs and relevant in tasks such as summarization and translation, shifting the focus from task performance to semantic — or ground truth [81]— accuracy.

- The investigation of the simultaneous use of multiple trained models, in ensemble schemes or as complementary automatic-metrics measuring the same dimension.
- The formulation of a theoretical and practical framework that, drawing from research in Commonsense and Natural Language reasoning [140, 141], Factuality [142], and Retrieval Augmented Generation [143], can proactively assess the information quality of stored data objects. For text, this can be conceptualized as the revision of the content of text instances by virtue of a trusted knowledge base. This can involve the use of agentic systems, or the application of scheduled quality measurement procedures. This could consider as reference benchmark the M^2 -TabFact verification task [221].

Bibliography

- [1] B. Thompson, N. Mathur, D. Deutsch, and H. Khayrallah, “Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.09598>
- [2] A. Deviyani and F. Diaz, “Contextual metric meta-evaluation by measuring local metric accuracy,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.19828>
- [3] C. Kiefer, “Assessing the quality of unstructured data: An initial overview,” in *Proceedings of the Lernen, Wissen, Daten, Analysen (LWDA) Conference 2016*, 2016, pp. 62–73. [Online]. Available: <https://ceur-ws.org/Vol-1670/paper-25.pdf>
- [4] S. Gandrabur, G. Foster, and G. Lapalme, “Confidence estimation for nlp applications,” *ACM Trans. Speech Lang. Process.*, vol. 3, no. 3, p. 1–29, Oct. 2006. [Online]. Available: <https://doi.org/10.1145/1177055.1177057>
- [5] B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz, “Requirements for data quality metrics,” *J. Data and Information Quality*, vol. 9, no. 2, Jan. 2018. [Online]. Available: <https://doi.org/10.1145/3148238>
- [6] R. Y. Wang and D. M. Strong, “Beyond accuracy: What data quality means to data consumers,” *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.
- [7] J. M. Juran, “Section 2: How to think about quality,” in *Juran’s Quality Handbook*, 5th ed., J. M. Juran and A. Blanton Godfrey, Eds. McGraw-Hill, 1998, pp. 26–43.
- [8] T. C. Redman, “Section 34: Second generation data quality systems,” in *Juran’s Quality Handbook*, 5th ed., J. M. Juran and A. Blanton Godfrey, Eds. McGraw-Hill, 1998, pp. 990–1003.
- [9] P. B. Crosby, *Quality Is Free: The Art of Making Quality Certain*. Penguin, 1980.
- [10] L. P. English, *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. Wiley, 1999.

- [11] C. Lacagnina, R. David, A. Nikiforova, M. E. Kuusniemi, C. Cappiello, O. Biehlmaier, c. Schubert, A. Bertino, H. Thiemann, and R. Dennis, “Towards a data quality framework for eosca,” *Zenodo*, 2023. [Online]. Available: <https://doi.org/10.5281/zenodo.7515815>
- [12] C. Guerra-García, A. Nikiforova, S. Jiménez, H. G. Perez-Gonzalez, M. Ramírez-Torres, and L. Ontañón-García, “Iso/iec 25012-based methodology for managing data quality requirements in the development of information systems: Towards data quality by design,” *Data & Knowledge Engineering*, vol. 145, p. 102152, 2023. [Online]. Available: <https://doi.org/10.1016/j.datak.2023.102152>
- [13] Y. Wand and R. Y. Wang, “Anchoring data quality dimensions in ontological foundations,” *Commun. ACM*, vol. 39, no. 11, p. 86–95, Nov. 1996. [Online]. Available: <https://doi.org/10.1145/240455.240479>
- [14] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, “Aimq: a methodology for information quality assessment,” *Information & Management*, vol. 40, pp. 133–146, 2002. [Online]. Available: <https://api.semanticscholar.org/CorpusID:263069687>
- [15] A. Haug, “Understanding the differences across data quality classifications: a literature review and guidelines for future research,” *Industrial Management & Data Systems*, vol. 121, no. 12, pp. 2651–2671, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0263557721000166>
- [16] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, “Methodologies for data quality assessment and improvement,” *ACM Comput. Surv.*, vol. 41, no. 3, Jul. 2009. [Online]. Available: <https://doi.org/10.1145/1541880.1541883>
- [17] D. Firmani, L. Tanca, and R. Torlone, “Ethical dimensions for data quality,” *J. Data and Information Quality*, vol. 12, no. 1, Dec. 2019. [Online]. Available: <https://doi.org/10.1145/3362121>
- [18] P. Illari, “Iq: Purpose and dimensions,” in *The Philosophy of Information Quality*, L. Floridi and P. Illari, Eds. Cham: Springer International Publishing, 2014, pp. 281–301. [Online]. Available: https://doi.org/10.1007/978-3-319-07121-3_14
- [19] S. M. Embury and P. Missier, “Forget dimensions: Define your information quality using quality view patterns,” in *The Philosophy of Information Quality*, L. Floridi and P. Illari, Eds. Cham: Springer International Publishing, 2014, pp. 25–41. [Online]. Available: https://doi.org/10.1007/978-3-319-07121-3_3

- [20] Y. W. Lee, L. L. Pipino, J. D. Funk, and R. Y. Wang, *Journey to Data Quality*. Cambridge: The MIT Press, 2006. [Online]. Available: <https://doi.org/10.7551/mitpress/4037.001.0001>
- [21] L. L. Pipino, Y. W. Lee, and R. Y. Wang, “Data quality assessment,” *Communications of the ACM*, vol. 45, no. 4, p. 211–218, Apr. 2002. [Online]. Available: <https://doi.org/10.1145/505248.506010>
- [22] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, “On the meaningfulness of “big data quality” (invited paper),” *Data Science and Engineering*, vol. 1, pp. 6–20, 2016. [Online]. Available: <https://doi.org/10.1007/s41019-015-0004-7>
- [23] J. Perry, “Subjectivity,” in *The Oxford Handbook of Philosophy of Mind*. Oxford University Press, 01 2009. [Online]. Available: <https://doi.org/10.1093/oxfordhb/9780199262618.003.0013>
- [24] H. P. Moravec, *Mind children : the future of robot and human intelligence*. Cambridge, Massachusetts: Harvard University Press, 1988.
- [25] R. K. Logan and M. Tandoc, “Thinking in patterns and the pattern of human thought as contrasted with ai data processing,” *Information*, vol. 9, no. 4, 2018. [Online]. Available: <https://www.mdpi.com/2078-2489/9/4/83>
- [26] J. Desai, D. Watson, V. Wang, M. Taddeo, and L. Floridi, “The epistemological foundations of data science: a critical review,” *Synthese: An International Journal for Epistemology, Methodology and Philosophy of Science*, no. 469, 2022. [Online]. Available: <https://doi.org/10.1007/s11229-022-03933-2>
- [27] F. Stollenwerk, J. Öhman, D. Petrelli, E. Wallerö, F. Olsson, C. Bengtsson, A. Horndahl, and G. Z. Gandler, “Text annotation handbook: A practical guide for machine learning projects,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.11780>
- [28] G. Ciocca, S. Corchs, F. Gasparini, C. Batini, and R. Schettini, *Quality of Images*. Cham: Springer International Publishing, 2016, pp. 113–135. [Online]. Available: https://doi.org/10.1007/978-3-319-24106-7_5
- [29] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, and H. Liu, “Large language models for data annotation and synthesis: A survey,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami,

- Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 930–957. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.54/>
- [30] C. Batini, M. Palmonari, and G. Viscusi, “Opening the closed world: A survey of information quality research in the wild,” in *The Philosophy of Information Quality*, L. Floridi and P. Illari, Eds. Cham: Springer International Publishing, 2014, pp. 43–73. [Online]. Available: https://doi.org/10.1007/978-3-319-07121-3_4
- [31] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, “SummEval: Re-evaluating summarization evaluation,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021. [Online]. Available: <https://aclanthology.org/2021.tacl-1.24/>
- [32] N. Carboni, R. David, F. de Jong, J. Ewbank, D. Fišer, B. Grüning, J. Thomas, and F. Schmidt-Tremmel, “On data annotation – from metadata to meaning – and how oscars and the science clusters contribute to it,” 2025. [Online]. Available: <https://doi.org/10.5281/zenodo.17224943>
- [33] M. Agosti and N. Ferro, “A formal model of annotations of digital content,” *ACM Trans. Inf. Syst.*, vol. 26, no. 1, p. 3–es, Nov. 2007. [Online]. Available: <https://doi.org/10.1145/1292591.1292594>
- [34] M. Wu, H. Brandhorst, M.-C. Marinescu, J. M. Lopez, M. Hlava, and J. Busch, “Automated metadata annotation: What is and is not possible with machine learning,” *Data Intelligence*, vol. 5, no. 1, pp. 122–138, 03 2023. [Online]. Available: https://doi.org/10.1162/dint_a_00162
- [35] M. Díaz, I. Kivlichan, R. Rosen, D. Baker, R. Amironesei, V. Prabhakaran, and R. Denton, “Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2342–2351. [Online]. Available: <https://doi.org/10.1145/3531146.3534647>
- [36] A. Gelman and C. Hennig, “Beyond subjective and objective in statistics,” *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 180, no. 4, pp. 967–1033, 08 2017. [Online]. Available: <https://doi.org/10.1111/rssa.12276>
- [37] I. Nonaka, “The knowledge-creating company,” *Harvard Business Review*, pp. 162–171, July-August 2007.

- [38] B. Plank, “The ‘problem’ of human label variation: On ground truth in data, modeling and evaluation,” 2022. [Online]. Available: <https://arxiv.org/abs/2211.02570>
- [39] C. Batini and M. Scannapieco, *Information Quality Dimensions for Maps and Texts*. Cham: Springer International Publishing, 2016, pp. 53–86. [Online]. Available: https://doi.org/10.1007/978-3-319-24106-7_3
- [40] R. Barzilay and M. Lapata, “Modeling local coherence: An entity-based approach,” in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, K. Knight, H. T. Ng, and K. Oflazer, Eds. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2005, pp. 141–148. [Online]. Available: <https://aclanthology.org/P05-1018/>
- [41] M. Elsner and E. Charniak, “Extending the entity grid with entity-specific features,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, D. Lin, Y. Matsumoto, and R. Mihalcea, Eds. Portland, Oregon, USA: Association for Computational Linguistics, Jun. 2011, pp. 125–129. [Online]. Available: <https://aclanthology.org/P11-2022/>
- [42] P. W. Foltz, W. Kintsch, and T. K. Landauer, “The measurement of textual coherence with latent semantic analysis,” *Discourse Processes*, vol. 25, no. 2-3, pp. 285–307, 1998. [Online]. Available: <https://doi.org/10.1080/01638539809545029>
- [43] A. C. Graesser, D. S. McNamara, and Z. Louwerse, Max M. and Cai, “Coh-metrix: Analysis of text on cohesion and language,” *Behavior Research Methods, Instruments, Computers*, vol. 36, no. 2, pp. 193–202, 2004. [Online]. Available: <https://doi.org/10.3758/BF03195564>
- [44] S. E. Leal, M. S. Duran, C. E. Scarton, N. S. Hartmann, and S. M. Aluísio, “Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese,” 2021. [Online]. Available: <https://arxiv.org/abs/2201.03445>
- [45] A. Quispesaravia, W. Perez, M. Sobrevilla Cabezudo, and F. Alva-Manchego, “Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odiijk, and S. Piperidis, Eds. Portorož, Slovenia: European Language

- Resources Association (ELRA), May 2016, pp. 4694–4698. [Online]. Available: <https://aclanthology.org/L16-1745/>
- [46] M. Santaholma, “Grammar sharing techniques for rule-based multilingual NLP systems,” in *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, J. Nivre, H.-J. Kaalep, K. Muischnek, and M. Koit, Eds. Tartu, Estonia: University of Tartu, Estonia, May 2007, pp. 253–260. [Online]. Available: <https://aclanthology.org/W07-2438/>
- [47] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” 2020. [Online]. Available: <https://arxiv.org/abs/1904.09675>
- [48] W. Yuan, G. Neubig, and P. Liu, “Bartscore: Evaluating generated text as text generation,” 2021. [Online]. Available: <https://arxiv.org/abs/2106.11520>
- [49] X. Dai, S. Karimi, and B. Fang, “A critical look at meta-evaluating summarisation evaluation metrics,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.19507>
- [50] P. Jayadharshini, C. Vasuki, L. Krishnasamy, J. Rakshitaa, N. Abarna, and N. Kannan, “Detecting and countering misinformation through nlp-based approach for fake news detection,” in *Intelligent Systems*, S. K. Udgate, S. Sethi, G. Ghinea, and S. K. Kuanar, Eds. Singapore: Springer Nature Singapore, 2025, pp. 101–113.
- [51] I. Gunasekara and I. Nejadgholi, “A review of standard text classification practices for multi-label toxicity identification of online content,” in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, D. Fišer, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, and J. Wernimont, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 21–25. [Online]. Available: <https://aclanthology.org/W18-5103/>
- [52] T. Xiao and J. Zhu, “Introduction to transformers: an nlp perspective,” 2023. [Online]. Available: <https://arxiv.org/abs/2311.17633>
- [53] L. Galke, A. Scherp, A. Diera, F. Karl, B. X. Lin, B. Khera, T. Meuser, and T. Singhal, “Are we really making much progress in text classification? a comparative review,” 2025. [Online]. Available: <https://arxiv.org/abs/2204.03954>
- [54] P. Colombo, N. Noiry, E. Irurozki, and S. Clemencon, “What are the best systems? new perspectives on nlp benchmarking,” 2022. [Online]. Available: <https://arxiv.org/abs/2202.03799>

- [55] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, “GLUE: A multi-task benchmark and analysis platform for natural language understanding,” in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, T. Linzen, G. Chrupała, and A. Alishahi, Eds. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. [Online]. Available: <https://aclanthology.org/W18-5446/>
- [56] S. Ni, G. Chen, S. Li, X. Chen, S. Li, B. Wang, Q. Wang, X. Wang, Y. Zhang, L. Fan, C. Li, R. Xu, L. Sun, and M. Yang, “A survey on large language model benchmarks,” 2025. [Online]. Available: <https://arxiv.org/abs/2508.15361>
- [57] O. Alonso and K. Church, “Evaluating the evaluations: A perspective on benchmarks,” *SIGIR Forum*, vol. 58, no. 2, p. 1–27, Mar. 2025. [Online]. Available: <https://doi.org/10.1145/3722449.3722467>
- [58] E. Montahaei, D. Alihosseini, and M. S. Baghshah, “Jointly measuring diversity and quality in text generation models,” 2019. [Online]. Available: <https://arxiv.org/abs/1904.03971>
- [59] J. H. Tanis, C. Giannella, and A. V. Mariano, “Introduction to graph neural networks: A starting point for machine learning engineers,” 2024. [Online]. Available: <https://arxiv.org/abs/2412.19419>
- [60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [61] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [62] D. Guo, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu *et al.*, “Deepseek-r1 incentivizing reasoning in llms through reinforcement learning,” 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2501.12948>
- [63] S. Mercer, S. Spillard, and D. P. Martin, “Brief analysis of deepseek r1 and its implications for generative ai,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.02523>
- [64] Y. Wang, J. Deng, A. Sun, and X. Meng, “Perplexity from plm is unreliable for evaluating text quality,” 2023. [Online]. Available: <https://arxiv.org/abs/2210.05892>

- [65] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, and Z. Sui, “Large language models are not fair evaluators,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.17926>
- [66] A. S. Thakur, K. Choudhary, V. S. Ramayapally, S. Vaidyanathan, and D. Hupkes, “Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges,” 2025. [Online]. Available: <https://arxiv.org/abs/2406.12624>
- [67] B. K. Kahn, D. M. Strong, and R. Y. Wang, “Information quality benchmarks: product and service performance,” *Commun. ACM*, vol. 45, no. 4, p. 184–192, Apr. 2002. [Online]. Available: <https://doi.org/10.1145/505248.506007>
- [68] M. Koo and S.-W. Yang, “Likert-type scale,” *Encyclopedia*, vol. 5, no. 1, 2025. [Online]. Available: <https://doi.org/10.3390/encyclopedia5010018>
- [69] M. A. Testa and D. C. Simonson, “Chapter 12 - the use of questionnaires and surveys,” in *Clinical and Translational Science*, second edition ed., D. Robertson and G. H. Williams, Eds. Academic Press, 2017, pp. 207–226. [Online]. Available: <https://doi.org/10.1016/B978-0-12-802101-9.00012-0>
- [70] R. F. DeVellis, “Inter-rater reliability,” in *Encyclopedia of Social Measurement*, K. Kempf-Leonard, Ed. New York: Elsevier, 2005, pp. 317–322. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B0123693985000955>
- [71] M. Tavakol and R. Dennick, “Making sense of cronbach’s alpha,” *International Journal of Medical Education*, vol. 2, pp. 53–55, 2011. [Online]. Available: <http://www.ijme.net/archive/2/cronbachs-alpha/>
- [72] R. Vaziri and M. Mohsenzadeh, “A questionnaire-based data quality methodology,” *International Journal of Database Management Systems*, vol. 4, pp. 55–68, 2012. [Online]. Available: <https://doi.org/10.5121/ijdms.2012.4204>
- [73] M. Bulmer, “Chapter 4: Quantiles,” in *A Portable Introduction to Data Analysis*. Brisbane: The University of Queensland, 2024, pp. 54–64. [Online]. Available: ["https://doi.org/10.14264/12dc235"](https://doi.org/10.14264/12dc235)
- [74] R. Vaziri and M. Mohsenzadeh, “Towards a practical “state reconstruction” for data quality methodologies: A customized list of dimensions,” in *Advances in Computer Science, Engineering & Applications*, D. C. Wyld, J. Zizka, and D. Nagamalai, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 825–835.

- [75] C. Batini, D. Barone, F. Cabitza, and S. Grega, “A data quality methodology for heterogeneous data,” *International Journal of Database Management Systems*, vol. 3, pp. 60–79, 02 2011. [Online]. Available: <https://doi.org/10.5121/ijdms.2011.3105>
- [76] B. Dresp-Langley, O. K. Ekseth, J. Fesl, S. Gohshi, M. Kurz, and H.-W. Sehring, “Occam’s razor for big data? on detecting quality in large unstructured datasets,” *Applied Sciences*, vol. 9, no. 15, 2019. [Online]. Available: <https://www.mdpi.com/2076-3417/9/15/3065>
- [77] A. T. Borgida, V. K. Chaudhri, P. Giorgini, and E. S. Yu, Eds., *Conceptual Modeling: Foundations and Applications: Essays in Honor of John Mylopoulos*. Springer Berlin, Heidelberg, 2009. [Online]. Available: <https://doi.org/10.1007/978-3-642-02463-4>
- [78] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “The kdd process for extracting useful knowledge from volumes of data,” *Commun. ACM*, vol. 39, no. 11, p. 27–34, Nov. 1996. [Online]. Available: <https://doi.org/10.1145/240455.240464>
- [79] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining (2nd Edition)*, 2nd ed. Pearson, 2018.
- [80] Ishwarappa and J. Anuradha, “A brief introduction on big data 5vs characteristics and hadoop technology,” *Procedia Computer Science*, vol. 48, pp. 319–324, 2015. [Online]. Available: <https://api.semanticscholar.org/CorpusID:60244553>
- [81] C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*, ser. Data-Centric Systems and Applications. Springer, 2006. [Online]. Available: <https://doi.org/10.1007/3-540-33173-5>
- [82] A. F. Machado, A. Charpentier, E. Flachaire, E. Gallic, and F. Hu, “From uncertainty to precision: Enhancing binary classifier performance through calibration,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.07790>
- [83] R. Luo, A. Bhatnagar, Y. Bai, S. Zhao, H. Wang, C. Xiong, S. Savarese, S. Ermon, E. Schmerling, and M. Pavone, “Local calibration: Metrics and recalibration,” 2022. [Online]. Available: <https://arxiv.org/abs/2102.10809>
- [84] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” 2017. [Online]. Available: <https://arxiv.org/abs/1706.04599>
- [85] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, “Classifier calibration: a survey on how to assess and improve predicted

- class probabilities,” *Machine Learning*, vol. 112, no. 9, p. 3211–3260, May 2023. [Online]. Available: <http://dx.doi.org/10.1007/s10994-023-06336-7>
- [86] N. Johnson, A. Cabrera, G. Plumb, and A. Talwalkar, “Where does my model underperform? a human evaluation of slice discovery algorithms,” *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 11, no. 1, p. 65–76, Nov. 2023. [Online]. Available: <http://dx.doi.org/10.1609/hcomp.v11i1.27548>
- [87] J. Geng, F. Cai, Y. Wang, H. Koepl, P. Nakov, and I. Gurevych, “A survey of confidence estimation and calibration in large language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2311.08298>
- [88] Z. Xia, J. Xu, Y. Zhang, and H. Liu, “A survey of uncertainty estimation methods on large language models,” in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 21 381–21 396. [Online]. Available: <https://aclanthology.org/2025.findings-acl.1101/>
- [89] S. Desai and G. Durrett, “Calibration of pre-trained transformers,” 2020. [Online]. Available: <https://arxiv.org/abs/2003.07892>
- [90] L. Kong, H. Jiang, Y. Zhuang, J. Lyu, T. Zhao, and C. Zhang, “Calibrated language model fine-tuning for in- and out-of-distribution data,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1326–1340. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.102/>
- [91] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, “Shortcut learning in deep neural networks,” *Nature Machine Intelligence*, vol. 2, no. 11, p. 665–673, Nov. 2020. [Online]. Available: <http://dx.doi.org/10.1038/s42256-020-00257-z>
- [92] M. Garouani, J. Mothe, A. Barhrhouj, and J. Aligon, “Investigating the duality of interpretability and explainability in machine learning,” in *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, Oct. 2024, p. 861–867. [Online]. Available: <http://dx.doi.org/10.1109/ICTAI62512.2024.00125>

- [93] A. Madsen, S. Reddy, and S. Chandar, “Post-hoc interpretability for neural nlp: A survey,” *ACM Comput. Surv.*, vol. 55, no. 8, Dec. 2022. [Online]. Available: <https://doi.org/10.1145/3546577>
- [94] Z. C. Lipton, “The mythos of model interpretability,” 2017. [Online]. Available: <https://arxiv.org/abs/1606.03490>
- [95] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” 2019. [Online]. Available: <https://arxiv.org/abs/1811.10154>
- [96] P. Knab, S. Marton, U. Schlegel, and C. Bartelt, “Which lime should i trust? concepts, challenges, and solutions,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.24365>
- [97] D. Fu and J. Fan, “Uniod: A universal model for outlier detection across diverse domains,” 2025. [Online]. Available: <https://arxiv.org/abs/2507.06624>
- [98] M. J. Zaki and W. J. Meira, “Part 3: Clustering,” in *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge: Cambridge University Press, 2020, pp. 332–466.
- [99] V. Wenz, A. Kesper, and G. Taentzer, “Clustering heterogeneous data values for data quality analysis,” *ACM Journal of Data and Information Quality*, vol. 15, no. 3, 2023. [Online]. Available: <https://doi.org/10.1145/3603710>
- [100] W. A. Qader, M. M. Ameen, and B. I. Ahmed, “An overview of bag of words;importance, implementation, applications, and challenges,” in *2019 International Engineering Conference (IEC)*, 2019, pp. 200–204.
- [101] M. H. Ahmed, S. Tiun, N. Omar, and N. S. Sani, “Short text clustering algorithms, application and challenges: A survey,” *Applied Sciences*, vol. 13, no. 1, 2023. [Online]. Available: <https://www.mdpi.com/2076-3417/13/1/342>
- [102] Q. Xu, H. Gu, and S. Ji, “Text clustering based on pre-trained models and autoencoders,” *Frontiers in Computational Neuroscience*, vol. Volume 17 - 2023, 2024. [Online]. Available: <https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2023.1334436>
- [103] X. Bai and S. Luo, *Combining Bag of Words Model and Information Theoretic Method for Image Clustering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 423–430. [Online]. Available: https://doi.org/10.1007/978-3-642-12990-2_48

- [104] H. Hou, S. Ding, C. Gui Cao, X. Xu, L. Guo, and X. Li, “A comprehensive survey of image clustering based on deep learning,” *Pattern Recognition*, vol. 172, p. 112590, 2026. [Online]. Available: <https://doi.org/10.1016/j.patcog.2025.112590>
- [105] B. Bergsma, M. Brzezinska, O. V. Yazyev, and M. Cernak, “Cluster-based pruning techniques for audio data,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.11922>
- [106] K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, “2 - data preprocessing,” in *Computational Learning Approaches to Data Analytics in Biomedical Applications*, K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht, and D. C. Wunsch II, Eds. Academic Press, 2020, pp. 7–27. [Online]. Available: <https://doi.org/10.1016/B978-0-12-814482-4.00002-4>
- [107] W. Hu, A. Zaveri, H. Qiu, and M. Dumontier, “Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata.” *BMC Bioinformatics*, vol. 18, no. 415, 2017. [Online]. Available: <https://doi.org/10.1186/s12859-017-1832-4>
- [108] H. Liu and S. Zhang, “Noisy data elimination using mutual k-nearest neighbor for classification mining,” *Journal of Systems and Software*, vol. 85, no. 5, pp. 1067–1074, 2012. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0164121211003049>
- [109] Z. Abedjan, L. Golab, F. Naumann, and T. Papenbrock, *Data Profiling*, ser. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2019. [Online]. Available: <https://doi.org/10.2200/S00878ED1V01Y201810DTM052>
- [110] L. Wißler, M. Almashraee, D. Monett, and A. Paschke, “The gold standard in corpus annotation,” in *5th IEEE Germany Student Conference*, 2014. [Online]. Available: <https://doi.org/10.13140/2.1.4316.3523>
- [111] R. Snow, B. O’Connor, D. Jurafsky, and A. Ng, “Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks,” in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, M. Lapata and H. T. Ng, Eds. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 254–263. [Online]. Available: <https://aclanthology.org/D08-1027/>
- [112] S. Nowak and S. Rüger, “How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation,” in *Proceedings of the International Conference on Multimedia Information Retrieval*, ser. MIR ’10.

- New York, NY, USA: Association for Computing Machinery, 2010, p. 557–566. [Online]. Available: <https://doi.org/10.1145/1743384.1743478>
- [113] D. Rebholz-Schuhmann, A. J. Jimeno-Yepes, E. van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, K. Tomanek, E. Beisswanger, and U. Hahn, “The calbc silver standard corpus for biomedical named entities - a study in harmonizing the contributions from four independent named entity taggers,” pp. 568–573, Jan. 2010. [Online]. Available: <http://hdl.handle.net/1765/103402>
- [114] D. Zhu, X. Shen, M. Mosbach, A. Stephan, and D. Klakow, “Weaker than you think: A critical look at weakly supervised learning,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.17442>
- [115] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, “Snorkel: rapid training data creation with weak supervision,” *Proc. VLDB Endow.*, vol. 11, no. 3, p. 269–282, Nov. 2017. [Online]. Available: <https://doi.org/10.14778/3157794.3157797>
- [116] S. Ibrahim, T. Nguyen, and X. Fu, “Deep learning from crowdsourced labels: Coupled cross-entropy minimization, identifiability, and regularization,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.03288>
- [117] E. Borisova, R. Abu Ahmad, L. Garcia-Castro, R. Usbeck, and G. Rehm, “Surveying the FAIRness of annotation tools: Difficult to find, difficult to reuse,” in *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, S. Henning and M. Stede, Eds. St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 29–45. [Online]. Available: <https://aclanthology.org/2024.law-1.4/>
- [118] J.-C. Klie, R. Eckart de Castilho, and I. Gurevych, “Analyzing dataset annotation quality management in the wild,” *Computational Linguistics*, vol. 50, no. 3, pp. 817–866, Sep. 2024. [Online]. Available: <https://aclanthology.org/2024.cl-3.1/>
- [119] J. Beck, “Quality aspects of annotated data,” *AStA Wirtschafts- und Sozialstatistisches Archiv*, vol. 17, no. 3, pp. 331–353, 2023. [Online]. Available: <https://hdl.handle.net/10419/313157>
- [120] R. Artstein and M. Poesio, “Inter-coder agreement for computational linguistics,” *Computational Linguistics*, vol. 34, no. 4, p. 555–596, Dec. 2008. [Online]. Available: <https://doi.org/10.1162/coli.07-034-R2>
- [121] M. Gonzalez-Cabello, A. Siddiq, C. J. Corbett, and C. Hu, “Fairness in crowdwork: Making the human ai supply chain more humane,” *Business*

- Horizons*, vol. 68, no. 5, pp. 645–657, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0007681324001265>
- [122] A. Smart, D. Wang, E. Monk, M. Díaz, A. Kasirzadeh, E. V. Liemt, and S. Schmer-Galunder, “Discipline and label: A weird genealogy and social theory of data annotation,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.06811>
- [123] M. Pushkarna, A. Zaldivar, and O. Kjartansson, “Data cards: Purposeful and transparent dataset documentation for responsible ai,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1776–1826. [Online]. Available: <https://doi.org/10.1145/3531146.3533231>
- [124] M. Kejriwal, H. Santos, K. Shen *et al.*, “A noise audit of human-labeled benchmarks for machine commonsense reasoning,” *Scientific Reports*, vol. 14, no. 8609, 2024. [Online]. Available: <https://doi.org/10.1038/s41598-024-58937-4>
- [125] E. Merdjanovska, A. Aynedinov, and A. Akbik, “Noisebench: Benchmarking the impact of real label noise on named entity recognition,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 18 182–18 198. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.1011/>
- [126] J.-Y. Antoine, J. Villaneau, and A. Lefeuvre, “Weighted Krippendorff’s alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation,” in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, S. Wintner, S. Goldwater, and S. Riezler, Eds. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 550–559. [Online]. Available: <https://aclanthology.org/E14-1058/>
- [127] I. Misra, C. L. Zitnick, M. Mitchell, and R. Girshick, “Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels,” 2016. [Online]. Available: <https://arxiv.org/abs/1512.06974>
- [128] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W.-N. Hsu, “Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.05139>

- [129] R.-C. Tu, Z.-A. Ma, T. Lan, Y. Zhao, H. Huang, and X.-L. Mao, “Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training, and meta-evaluation benchmark,” 2024. [Online]. Available: <https://arxiv.org/abs/2411.15488>
- [130] G. I. Winata, D. Anugraha, L. Susanto, G. Kuwanto, and D. T. Wijaya, “Metametrics: Calibrating metrics for generation tasks using human preferences,” 2025. [Online]. Available: <https://arxiv.org/abs/2410.02381>
- [131] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” 2019. [Online]. Available: <https://arxiv.org/abs/1907.10902>
- [132] G. Lorenzoni, I. Portugal, P. Alencar, and D. Cowan, “Exploring variability in fine-tuned models for text classification with distilbert,” 2024. [Online]. Available: <https://arxiv.org/abs/2501.00241>
- [133] D. Chandran and V. Gupta, “A short review of the literature on automatic data quality,” *Journal of Computer and Communications*, vol. 10, no. 5, pp. 55–73, 2022.
- [134] “Guide to the software engineering body of knowledge (swebok guide) version 4.0,” H. Washizaki, Ed. IEEE Computer Society, 2024. [Online]. Available: www.swebok.org
- [135] E. Clark, S. Rijhwani, S. Gehrmann, J. Maynez, R. Aharoni, V. Nikolaev, T. Sellam, A. Siddhant, D. Das, and A. P. Parikh, “Seahorse: A multilingual, multifaceted dataset for summarization evaluation,” 2023, GitHub Website: <https://github.com/google-research-datasets/seahorse>, HuggingFace page: <https://huggingface.co/datasets/SEACrowd/seahorse>. [Online]. Available: <https://arxiv.org/abs/2305.13194>
- [136] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: a large-scale dataset for fact extraction and VERification,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 809–819. [Online]. Available: <https://aclanthology.org/N18-1074/>
- [137] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, “Summac: Re-visiting nli-based models for inconsistency detection in summarization,” 2021. [Online]. Available: <https://arxiv.org/abs/2111.09525>

- [138] L. Kumar, N. S. Upadhye, and K. Piedy, “Advances and challenges in semantic textual similarity: A comprehensive survey,” 2025. [Online]. Available: <https://arxiv.org/abs/2601.03270>
- [139] A. Nenkova and R. Passonneau, “Evaluating content selection in summarization: The pyramid method,” in *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2 - May 7 2004, pp. 145–152. [Online]. Available: <https://aclanthology.org/N04-1019/>
- [140] E. Davis, “Benchmarks for automated commonsense reasoning: A survey,” *ACM Comput. Surv.*, vol. 56, no. 4, Oct. 2023. [Online]. Available: <https://doi.org/10.1145/3615355>
- [141] F. Yu, H. Zhang, P. Tiwari, and B. Wang, “Natural language reasoning, a survey,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.14725>
- [142] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, “FActScore: Fine-grained atomic evaluation of factual precision in long form text generation,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 076–12 100. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.741/>
- [143] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, “RAGAs: Automated evaluation of retrieval augmented generation,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, N. Aletras and O. De Clercq, Eds. St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 150–158. [Online]. Available: <https://aclanthology.org/2024.eacl-demo.16/>
- [144] M. J. Zaki and W. J. Meira, “Chapter 22: Classification assessment,” in *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge: Cambridge University Press, 2020, pp. 546–586.
- [145] J. Hanley, “The meaning and use of the area under a receiver operating characteristic (roc) curve,” *Radiology*, vol. 143, pp. 29–36, 05 1982.
- [146] M. Bulmer, “Chapter 7: Linear relationships,” in *A Portable Introduction to Data Analysis*. Brisbane: The University of Queensland, 2024, pp. 88–102. [Online]. Available: "<https://doi.org/10.14264/12dc235>"

- [147] H. Akoglu, “User’s guide to correlation coefficients,” *Turkish Journal of Emergency Medicine*, vol. 18, no. 3, pp. 91–93, 2018. [Online]. Available: <https://doi.org/10.1016/j.tjem.2018.08.001>
- [148] D. Chicco and G. Jurman, “The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, 2020. [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077436560&doi=10.1186%2fs12864-019-6413-7&partnerID=40&md5=c2e4ab5b979e336459e73b381fb6c84c>
- [149] X. Gao, D. Xie, Y. Zhang, Z. Wang, C. He, H. Yin, and C. C. Wentao Zhang, “A comprehensive survey on imbalanced data learning,” *CoRR*, vol. abs/2502.08960, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2502.08960>
- [150] D. Berrar, “Performance measures for binary classification,” in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. Oxford: Academic Press, 2019, pp. 546–560. [Online]. Available: <https://doi.org/10.1016/B978-0-12-809633-8.20351-8>
- [151] A. More, “Survey of resampling techniques for improving classification performance in unbalanced datasets,” 2016. [Online]. Available: <https://arxiv.org/abs/1608.06048>
- [152] Z. Qin, C. Zhang, T. Wang, and S. Zhang, “Cost sensitive classification in data mining,” in *Advanced Data Mining and Applications*, L. Cao, Y. Feng, and J. Zhong, Eds. Springer, 2010, pp. 1–11.
- [153] M. Bystronński, M. Hołysz, G. Piotrowski, N. V. Chawla, and T. Kajdanowicz, “Smotext: Smote meets large language models,” 2025. [Online]. Available: <https://arxiv.org/abs/2505.13434>
- [154] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, “Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach,” *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988. [Online]. Available: <http://www.jstor.org/stable/2531595>
- [155] X. Sun and W. Xu, “Fast implementation of delong’s algorithm for comparing the areas under correlated receiver operating characteristic curves,” *IEEE Signal Processing Letters*, vol. 21, pp. 1389–1393, 2014. [Online]. Available: <https://api.semanticscholar.org/CorpusID:8218719>

- [156] E. H. Weissler, Z. J., S. Lippmann, S. Rusincovitch, R. Henao, and W. S. Jones, “Use of natural language processing to improve identification of patients with peripheral artery disease,” *Circulation: Cardiovascular Interventions*, vol. 13, no. 10, 2020. [Online]. Available: <https://doi.org/10.1161/CIRCINTERVENTIONS.120.009447>
- [157] F. Nasriddinov, R. Kocielnik, A. Anandkumar, and A. J. Hung, “Generating natural-language surgical feedback: From structured representation to domain-grounded evaluation,” 2025. [Online]. Available: <https://arxiv.org/abs/2511.15159>
- [158] M. A. Clapp, E. Kim, K. E. James, R. H. Perlis, A. J. Kaimal, J. McCoy, Thomas H., and S. R. Easter, “Comparison of natural language processing of clinical notes with a validated risk-stratification tool to predict severe maternal morbidity,” *JAMA Network Open*, vol. 5, no. 10, pp. e2234924–e2234924, 10 2022. [Online]. Available: <https://doi.org/10.1001/jamanetworkopen.2022.34924>
- [159] M. Aickin and G. H., “Adjusting for multiple testing when reporting research results: the bonferroni vs holm methods,” *American Journal of Public Health*, vol. 86, no. 5, pp. 726–728, 1996. [Online]. Available: <https://doi.org/10.2105/AJPH.86.5.726>
- [160] K. E. Eichstaedt, K. Kovatch, and D. A. Maroof, “A less conservative method to adjust for familywise error rate in neuropsychological research: The holm’s sequential bonferroni procedure,” *NeuroRehabilitation*, vol. 32, no. 3, pp. 693–696, 2013, pMID: 23648625. [Online]. Available: <https://journals.sagepub.com/doi/abs/10.3233/NRE-130893>
- [161] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Nature Scientific Reports*, vol. 14, no. 6086, 2024. [Online]. Available: <https://doi.org/10.1038/s41598-024-56706-x>
- [162] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, “The hitchhiker’s guide to testing statistical significance in natural language processing,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1383–1392. [Online]. Available: <https://aclanthology.org/P18-1128/>
- [163] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, “mt5: A massively multilingual pre-trained text-to-text transformer,” 2021. [Online]. Available: <https://arxiv.org/abs/2010.11934>
- [164] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” 2020, huggingFace page of

- multilingual model: <https://huggingface.co/distilbert/distilbert-base-multilingual-cased>. [Online]. Available: <https://doi.org/10.48550/arXiv.1910.01108>
- [165] M. Muffo and E. Bertino, “Bertino: an italian distilbert model,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.18121>
- [166] O. Rohanian, M. Nouriborji, S. Kouchaki, and D. A. Clifton, “On the effectiveness of compact biomedical transformers,” 2022. [Online]. Available: <https://arxiv.org/abs/2209.03182>
- [167] J. L. Gastaldi, J. Terilla, L. Malagutti, B. DuSell, T. Vieira, and R. Cotterell, “The foundations of tokenization: Statistical and computational concerns,” 2025. [Online]. Available: <https://arxiv.org/abs/2407.11606>
- [168] T. Etchegoyhen and H. Gete, “To case or not to case: Evaluating casing methods for neural machine translation,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 3752–3760. [Online]. Available: <https://aclanthology.org/2020.lrec-1.463/>
- [169] C. Paik and K. V. D. Wense, “The effectiveness of uncased tokenization for clinical notes,” in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 14 986–14 992. [Online]. Available: <https://aclanthology.org/2025.findings-acl.775/>
- [170] A. Chizhikova, V. Konovalov, and M. Burtsev, “Multilingual case-insensitive named entity recognition,” in *Advances in Neural Computation, Machine Learning, and Cognitive Research VI*, B. Kryzhanovsky, W. Dunin-Barkowski, V. Redko, and Y. Tiumentsev, Eds. Cham: Springer International Publishing, 2023, pp. 448–454.
- [171] R. Phillips, “Why is lower case better? some data from a search task,” *Applied Ergonomics*, vol. 10, no. 4, pp. 211–214, 1979. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0003687079902126>
- [172] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” 2020. [Online]. Available: <https://arxiv.org/abs/2004.05150>

- [173] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, “Big bird: Transformers for longer sequences,” 2021. [Online]. Available: <https://arxiv.org/abs/2007.14062>
- [174] A. Jaiswal and E. Milios, “Breaking the token barrier: Chunking and convolution for efficient long text classification with bert,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.20558>
- [175] Y. Chai, H. Xie, and J. S. Qin, “Text data augmentation for large language models: A comprehensive survey of methods, challenges, and opportunities,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.18845>
- [176] M. Bayer, M.-A. Kaufhold, and C. Reuter, “A survey on data augmentation for text classification,” vol. 55, no. 7, Dec. 2022. [Online]. Available: <https://doi.org/10.1145/3544558>
- [177] D. S. Kalra and M. Barkeshli, “Why warmup the learning rate? underlying mechanisms and improvements,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.09405>
- [178] M. Li, E. Yumer, and D. Ramanan, “Budgeted training: Rethinking deep neural network training under resource constraints,” 2020. [Online]. Available: <https://arxiv.org/abs/1905.04753>
- [179] R. Abdulkadirov, P. Lyakhov, and N. Nagornov, “Survey of optimization algorithms in modern neural networks,” *Mathematics*, vol. 11, no. 11, 2023. [Online]. Available: <https://www.mdpi.com/2227-7390/11/11/2466>
- [180] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2019. [Online]. Available: <https://arxiv.org/abs/1711.05101>
- [181] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: <http://jmlr.org/papers/v15/srivastava14a.html>
- [182] Z. Alyafeai, M. S. AlShaibani, and I. Ahmad, “A survey on transfer learning in natural language processing,” 2020. [Online]. Available: <https://arxiv.org/abs/2007.04239>
- [183] J. Lee, R. Tang, and J. Lin, “What would elsa do? freezing layers during transformer fine-tuning,” 2019. [Online]. Available: <https://arxiv.org/abs/1911.03090>

- [184] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, “Revealing the dark secrets of bert,” 2019. [Online]. Available: <https://arxiv.org/abs/1908.08593>
- [185] A. Rogers, O. Kovaleva, and A. Rumshisky, “A primer in bertology: What we know about how bert works,” 2020. [Online]. Available: <https://arxiv.org/abs/2002.12327>
- [186] S. Ennadir, L. Zólyomi, O. Smirnov, T. Wang, J. Pertoft, F. Cornell, and L. Cao, “Pool me wisely: On the effect of pooling in transformer-based models,” 2025. [Online]. Available: <https://arxiv.org/abs/2510.03339>
- [187] J. Xie, R. Zeng, Q. Wang, Z. Zhou, and P. Li, “Sot: Delving deeper into classification head for transformer,” 2021. [Online]. Available: <https://arxiv.org/abs/2104.10935>
- [188] W. de Vries, A. van Cranenburgh, and M. Nissim, “What’s so special about BERT’s layers? a closer look at the NLP pipeline in monolingual and multilingual models,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4339–4350. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.389/>
- [189] D. Oh, Y. Kim, H. Lee, H. H. Huang, and H. Lim, “Don’t judge a language model by its last layer: Contrastive learning with layer-wise attention pooling,” in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 4585–4592. [Online]. Available: <https://aclanthology.org/2022.coling-1.405/>
- [190] I. Tenney, D. Das, and E. Pavlick, “Bert rediscovers the classical nlp pipeline,” 2019. [Online]. Available: <https://arxiv.org/abs/1905.05950>
- [191] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, “Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment,” 2023. [Online]. Available: <https://arxiv.org/abs/2312.12148>
- [192] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” 2021. [Online]. Available: <https://arxiv.org/abs/2106.09685>
- [193] F. Corradini, M. Leonesi, and M. Piangerelli, “State of the art and future directions of small language models: A systematic review,” *Big Data and Cognitive Computing*,

- vol. 9, no. 7, 2025. [Online]. Available: <https://www.mdpi.com/2504-2289/9/7/189>
- [194] S. Nwaiwu, “Parameter-efficient fine-tuning for low-resource text classification: a comparative study of lora, ia3, and reft,” *Frontiers in Big Data*, 2025. [Online]. Available: <https://doi.org/10.3389/fdata.2025.1677331>
- [195] R. Aharoni, S. Narayan, J. Maynez, J. Herzig, E. Clark, and M. Lapata, “Multilingual summarization with factual consistency evaluation,” in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 3562–3591. [Online]. Available: <https://aclanthology.org/2023.findings-acl.220/>
- [196] T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar, “XL-sum: Large-scale multilingual abstractive summarization for 44 languages,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4693–4703. [Online]. Available: <https://aclanthology.org/2021.findings-acl.413>
- [197] A. Warstadt, A. Singh, and S. R. Bowman, “Neural network acceptability judgments,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 625–641, 2019. [Online]. Available: <https://aclanthology.org/Q19-1040/>
- [198] D. Trotta, R. Guarasci, E. Leonardelli, and S. Tonelli, “Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus,” in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 2929–2940. [Online]. Available: <https://aclanthology.org/2021.findings-emnlp.250/>
- [199] M. Jentoft and D. Samuel, “NoCoLA: The Norwegian corpus of linguistic acceptability,” in *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, T. Alumäe and M. Fishel, Eds. Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 610–617. [Online]. Available: <https://aclanthology.org/2023.nodalida-1.60/>
- [200] Z. Zhang, Y. Liu, W. Huang, J. Mao, R. Wang, and H. Hu, “MELA: Multilingual evaluation of linguistic acceptability,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association

- for Computational Linguistics, Aug. 2024, pp. 2658–2674. [Online]. Available: <https://aclanthology.org/2024.acl-long.146/>
- [201] M. Heilman, A. Cahill, N. Madnani, M. Lopez, M. Mulholland, and J. Tetreault, “Predicting grammaticality on an ordinal scale,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, K. Toutanova and H. Wu, Eds. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 174–180. [Online]. Available: <https://aclanthology.org/P14-2029/>
- [202] C. Napoles, K. Sakaguchi, and J. Tetreault, “JFLEG: A fluency corpus and benchmark for grammatical error correction,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, M. Lapata, P. Blunsom, and A. Koller, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 229–234. [Online]. Available: <https://aclanthology.org/E17-2037/>
- [203] A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman, “BLiMP: The benchmark of linguistic minimal pairs for English,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 377–392, 2020. [Online]. Available: <https://aclanthology.org/2020.tacl-1.25/>
- [204] I. Caswell, L. Wang, and I. Papadimitriou, “Separating the wheat from the chaff with BREAD: An open-source benchmark and metrics to detect redundancy in text,” in *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, S. Gehrmann, A. Wang, J. Sedoc, E. Clark, K. Dhole, K. R. Chandu, E. Santus, and H. Sedghamiz, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 324–338, GitHub website: <https://github.com/toizzy/bread>. [Online]. Available: <https://aclanthology.org/2023.gem-1.27/>
- [205] Y. He, Z. Wang, Y. Zhang, R. Huang, and J. Caverlee, “PARADE: A New Dataset for Paraphrase Identification Requiring Computer Science Domain Knowledge,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 7572–7582. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.611/>
- [206] W. B. Dolan and C. Brockett, “Automatically constructing a corpus of sentential paraphrases,” in *Proceedings of the Third International Workshop on Paraphrasing*

- (*IWP2005*), 2005. [Online]. Available: <https://aclanthology.org/I05-5002/>
- [207] Y. Yang, Y. Zhang, C. Tar, and J. Baldridge, “Paws-x: A cross-lingual adversarial dataset for paraphrase identification,” 2019. [Online]. Available: <https://arxiv.org/abs/1908.11828>
- [208] A. Buja, W. Stuetzle, and Y. Shen, “Loss functions for binary class probability estimation and classification: Structure and applications,” 2005. [Online]. Available: <https://api.semanticscholar.org/CorpusID:5925076>
- [209] J. Bröcker, “Reliability, sufficiency, and the decomposition of proper scores,” *Quarterly Journal of the Royal Meteorological Society*, vol. 135, no. 643, p. 1512–1519, Jul. 2009. [Online]. Available: <http://dx.doi.org/10.1002/qj.456>
- [210] J. Eronen, M. Ptaszynski, and F. Masui, “Zero-shot cross-lingual transfer language selection using linguistic similarity,” *Information Processing Management*, vol. 60, no. 3, p. 103250, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S030645732200351X>
- [211] S. Kudugunta, I. Caswell, B. Zhang, X. Garcia, C. A. Choquette-Choo, K. Lee, D. Xin, A. Kusupati, R. Stella, A. Bapna, and O. Firat, “Madlad-400: A multilingual and document-level large audited dataset,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.04662>
- [212] Y. Zhang, J. Baldridge, and L. He, “Paws: Paraphrase adversaries from word scrambling,” 2019. [Online]. Available: <https://arxiv.org/abs/1904.01130>
- [213] Z. Zhang, Y. Liu, W. Huang, J. Mao, R. Wang, and H. Hu, “MELA: Multilingual evaluation of linguistic acceptability,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2658–2674. [Online]. Available: <https://aclanthology.org/2024.acl-long.146/>
- [214] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers distillation through attention,” 2021. [Online]. Available: <https://arxiv.org/abs/2012.12877>
- [215] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan, “Tinyvit: Fast pretraining distillation for small vision transformers,” 2022. [Online]. Available: <https://arxiv.org/abs/2207.10666>

- [216] H.-J. Chang, S. wen Yang, and H. yi Lee, “Distillhubert: Speech representation learning by layer-wise distillation of hidden-unit bert,” 2022. [Online]. Available: <https://arxiv.org/abs/2110.01900>
- [217] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Z. Wang, Y. Guo, and J. Fu, “Mert: Acoustic music understanding model with large-scale self-supervised training,” 2024. [Online]. Available: <https://arxiv.org/abs/2306.00107>
- [218] L. Wang, X. Zhang, H. Su, and J. Zhu, “A comprehensive survey of continual learning: Theory, method and application,” vol. 46, no. 8, p. 5362–5383, Aug. 2024. [Online]. Available: <https://doi.org/10.1109/TPAMI.2024.3367329>
- [219] S. C. H. Hoi, D. Sahoo, J. Lu, and P. Zhao, “Online learning: A comprehensive survey,” 2018. [Online]. Available: <https://arxiv.org/abs/1802.02871>
- [220] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, “Learning under concept drift: A review,” *IEEE Transactions on Knowledge and Data Engineering*, p. 1–1, 2018. [Online]. Available: <http://dx.doi.org/10.1109/TKDE.2018.2876857>
- [221] M. Zhou, L. Zhang, S. Horng, M. Chen, K.-H. Huang, and S.-F. Chang, “M²-TabFact: Multi-document multi-modal fact verification with visual and textual representations of tabular data,” in *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 26 239–26 256. [Online]. Available: <https://aclanthology.org/2025.findings-acl.1345/>
- [222] N. Kühl, M. Goutier, L. Baier, C. Wolff, and D. Martin, “Human vs. supervised machine learning: Who learns patterns faster?” *Cognitive Systems Research*, vol. 76, pp. 78–92, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1389041722000419>

A | Appendix A

A.1. Experimental hardware setting

Experiments were run in the following hardware:

- Processor: Intel Core i5-10300H CPU (2.5 GHz)
- GPU: NVIDIA GeForce GTX 1650
(4096MB GDDR6 VRAM / 896 CUDA Cores / 1.5 GHz)
- RAM: 16GB

A.2. Exploration phase

A.2.1. Hyperparameters

$Learning\ rate = 5 \times 10^{-5}$ $LR\ scheduler = Linear$ $warmup\ ratio = 0.1$ $Epochs = 4$ $weight\ decay = 0.01$
$Dropout$ 0.1 in all components of customized models. 0.1 at embeddings and transformer layers, and 0.2 at classification head for baseline [CLS] 3 trainable layers FT (default values).

Table A.1: Hyperparameters (exploration phase)

$rank\ (adapter\ size) = 8$ $\alpha\ (scaling\ factor) = 32$
$Dropout = 0.05$
Target modules in last 3 layers: attention projection layers $q_lin, k_lin, v_lin, out_lin$

Table A.2: LoRA hyperparameters (exploration phase)

A.2.2. Model variants (Highest Val ROC AUC checkpoint)

Table A.3: Comprehensible validation

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Fine tuning (2 layers, Concat fusion)	0.830467	0.898968	0.972178	0.836012	0.779310	0.615323	<u>0.893640</u>	0.511287
Fine tuning (2 layers, Sum fusion)	0.841372	0.906202	0.971217	0.849346	0.767816	0.617162	0.893181	0.517096
Fine tuning (3 layers, Sum fusion)	0.853850	0.914417	0.969295	0.865421	0.747126	0.612547	0.891907	0.525362
Fine tuning (3 layers)	0.838786	0.904604	0.970315	0.847227	0.760920	0.608147	0.891833	0.512972
Fine tuning (3 layers, Concat fusion)	0.842383	0.907029	0.969383	0.852212	0.751724	0.603936	0.890516	0.515109
Feature extraction	0.848117	0.910925	0.967236	0.860810	0.731034	0.591844	0.880155	0.510129
LoRA	0.840247	0.905863	0.967044	0.851963	0.732184	0.584147	0.877124	0.502349

Table A.4: Comprehensible test

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Fine tuning (3 layers, Sum fusion)	0.851176	0.913604	0.968636	0.864488	0.716200	0.580688	0.873249	0.487223
Fine tuning (3 layers)	0.831875	0.901116	0.969686	0.841603	0.733252	0.574855	0.873211	0.470622
Fine tuning (3 layers, Concat fusion)	0.839584	0.906130	0.969400	0.850613	0.727771	0.578384	0.872199	0.475442
Fine tuning (2 layers, Concat fusion)	0.828376	0.898693	0.971124	0.836317	0.747868	0.584185	0.871819	0.467516
Fine tuning (2 layers, Sum fusion)	0.837179	0.904606	0.969115	0.848150	0.725944	0.574094	0.871297	0.472540
Feature extraction	0.840897	0.907284	0.966074	0.855238	0.695493	0.550731	0.863720	0.465166
LoRA	0.837671	0.905226	0.965938	0.851694	0.695493	0.547187	0.861730	0.457138

Table A.5: Repetition validation

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Fine tuning (3 layers)	0.942804	0.967160	0.982841	0.951972	0.872432	0.824404	<u>0.967125</u>	0.789181
Fine tuning (3 layers, Concat fusion)	0.939564	0.965151	0.985184	0.945915	0.890811	0.836726	0.966972	0.782302
Fine tuning (2 layers, Sum fusion)	0.943427	0.967558	0.982013	0.953521	0.865946	0.819467	0.965721	0.788698
Fine tuning (3 layers, Sum fusion)	0.947040	0.969680	0.982507	0.957183	0.869189	0.826372	0.965663	0.799804
LoRA	0.937695	0.964122	0.982738	0.946197	0.872432	0.818630	0.965574	0.773904
Fine tuning (2 layers, Concat fusion)	0.942181	0.966781	0.983110	0.950986	0.874595	0.825581	0.964464	0.783736
Feature extraction	0.903676	0.943498	0.980702	0.909014	0.862703	0.771717	0.954781	0.702853

Table A.6: Repetition test

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Fine tuning (3 layers)	0.940445	0.965457	0.982547	0.948952	0.879765	0.828717	0.966784	0.790262
Fine tuning (3 layers, Concat fusion)	0.935877	0.962665	0.983625	0.942579	0.888074	0.830653	0.966105	0.779658
Fine tuning (3 layers, Sum fusion)	0.943870	0.967540	0.981664	0.953817	0.872923	0.826739	0.965436	0.798041
Fine tuning (2 layers, Sum fusion)	0.942909	0.967016	0.980152	0.954228	0.862170	0.816398	0.965419	0.792378
Fine tuning (2 layers, Concat fusion)	0.939844	0.965118	0.981919	0.948883	0.875367	0.824250	0.964669	0.787625
LoRA	0.933474	0.961282	0.981782	0.941620	0.875367	0.816986	0.963054	0.774806
Feature extraction	0.898978	0.940120	0.979004	0.904207	0.861681	0.765889	0.949850	0.699082

Table A.7: Grammatical validation

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Fine tuning (3 layers)	0.758006	0.850799	0.954984	0.767110	0.676580	0.443690	<u>0.791060</u>	0.345272
Fine tuning (3 layers, Concat fusion)	0.774206	0.862394	0.954286	0.786644	0.662949	0.449594	0.789607	0.347544
Fine tuning (3 layers, Sum fusion)	0.755639	0.849258	0.953894	0.765309	0.669145	0.434454	0.787096	0.343364
Fine tuning (2 layers, Sum fusion)	0.825171	0.898060	0.944232	0.856193	0.547708	0.403900	0.786017	0.357828
Fine tuning (2 layers, Concat fusion)	0.820685	0.895017	0.945292	0.849820	0.560099	0.409919	0.785987	0.357227
Feature extraction	0.792523	0.876566	0.942752	0.819063	0.555143	0.374206	0.758198	0.315123
LoRA	0.801495	0.883014	0.939522	0.832918	0.520446	0.353364	0.751581	0.310772

Table A.8: Grammatical test

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Fine tuning (3 layers, Sum fusion)	0.745130	0.842586	0.952961	0.755125	0.651741	0.406866	0.775082	0.319161
Fine tuning (2 layers, Sum fusion)	0.820587	0.895482	0.945069	0.850839	0.537935	0.388774	0.774605	0.341288
Fine tuning (3 layers, Concat fusion)	0.763468	0.855781	0.952505	0.776890	0.638060	0.414950	0.773708	0.319438
Fine tuning (3 layers)	0.746633	0.843485	0.954202	0.755791	0.661070	0.416860	0.773207	0.313210
Fine tuning (2 layers, Concat fusion)	0.817761	0.893466	0.946600	0.845980	0.554104	0.400084	0.772599	0.335533
LoRA	0.798461	0.881278	0.941787	0.828075	0.521766	0.349841	0.748229	0.299813
Feature extraction	0.787999	0.873900	0.944350	0.813232	0.552239	0.365471	0.746517	0.297221

A.3. Feature Extractor and LoRA (Additional epochs)

Feature Extraction

Table A.9: Comprehensible validation (Feature Extraction 12 and 4 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	val_loss	train_loss
Feature extraction (12 epochs)	0.857111	0.916738	0.966436	0.871900	0.720690	0.592590	0.888474	0.526380	0.436638	0.367900
Feature extraction (4 epochs)	0.848117	0.910925	0.967236	0.860810	0.731034	0.591844	0.880155	0.510129	0.437742	0.408400

Table A.10: Comprehensible test (Feature Extraction 12 and 4 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Feature extraction (12 epochs)	0.852214	0.914503	0.965858	0.868333	0.688794	0.557127	0.868393	0.476984
Feature extraction (4 epochs)	0.840897	0.907284	0.966074	0.855238	0.695493	0.550731	0.863720	0.465166

Table A.11: Repetition validation (Feature Extraction 12 and 4 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	val_loss	train_loss
Feature extraction (12 epochs)	0.915888	0.950970	0.981851	0.921972	0.869189	0.791161	0.958625	0.722394	0.277764	0.247400
Feature extraction (4 epochs)	0.903676	0.943498	0.980702	0.909014	0.862703	0.771717	0.954781	0.702853	0.280482	0.281200

Table A.12: Repetition test (Feature Extraction 12 and 4 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Feature extraction (12 epochs)	0.908413	0.946019	0.979176	0.915034	0.861193	0.776226	0.954084	0.718872
Feature extraction (4 epochs)	0.898978	0.940120	0.979004	0.904207	0.861681	0.765889	0.949850	0.699082

Table A.13: Grammatical validation (Feature Extraction 12 and 4 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	val_loss	train_loss
Feature extraction (12 epochs)	0.798879	0.880710	0.943916	0.825436	0.561338	0.386775	0.765510	0.324637	0.598411	0.524500
Feature extraction (4 epochs)	0.792523	0.876566	0.942752	0.819063	0.555143	0.374206	0.758198	0.315123	0.589364	0.564600

Table A.14: Grammatical test (Feature Extraction 12 and 4 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
Feature extraction (12 epochs)	0.794974	0.878579	0.944640	0.821153	0.550373	0.371526	0.747376	0.299238
Feature extraction (4 epochs)	0.787999	0.873900	0.944350	0.813232	0.552239	0.365471	0.746517	0.297221

LoRA

Table A.15: Comprehensible validation (LoRA 4 and 6 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	val_loss	train_loss
LoRA 6 epochs 2 Layers LR= 3×10^{-4}	0.861945	0.919937	0.964720	0.879128	0.703448	0.582576	0.887112	0.527942	0.446859	0.429000
LoRA 4 epochs 3 Layers LR= 5×10^{-5}	0.840247	0.905863	0.967044	0.851963	0.732184	0.584147	0.877124	0.502349	0.448897	0.418500

Table A.16: Comprehensible test (LoRA 4 and 6 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
LoRA 6 epochs 2 Layers LR= 3×10^{-4}	0.857682	0.918101	0.963991	0.876382	0.668088	0.544469	0.868709	0.482329
LoRA 4 epochs 3 Layers LR= 5×10^{-5}	0.837671	0.905226	0.965938	0.851694	0.695493	0.547187	0.861730	0.457138

Table A.17: Repetition validation (LoRA 4 and 6 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	val_loss	train_loss
LoRA 4 epochs 3 Layers LR= 5×10^{-5}	0.937695	0.964122	0.982738	0.946197	0.872432	0.818630	0.965574	0.773904	0.251182	0.257300
LoRA 6 epochs 2 Layers LR= 3×10^{-4}	0.941184	0.966228	0.981966	0.950986	0.865946	0.816932	0.965303	0.782056	0.251700	0.260715

Table A.18: Repetition test (LoRA 4 and 6 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
LoRA 6 epochs 2 Layers LR= 3×10^{-4}	0.937740	0.963875	0.981326	0.947033	0.871457	0.818490	0.964189	0.783386
LoRA 4 epochs 3 Layers LR= 5×10^{-5}	0.933474	0.961282	0.981782	0.941620	0.875367	0.816986	0.963054	0.774806

Table A.19: Grammatical validation (LoRA 4 and 6 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	val_loss	train_loss
LoRA 6 epochs 2 Layers LR= 3×10^{-4}	0.776698	0.864653	0.950515	0.793017	0.630731	0.423749	0.778113	0.334573	0.571281	0.542700
LoRA 4 epochs 3 Layers LR= 5×10^{-5}	0.801495	0.883014	0.939522	0.832918	0.520446	0.353364	0.751581	0.310772	0.598121	0.571200

Table A.20: Grammatical test (LoRA 4 and 6 epochs)

Model	Accuracy	F1	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation
LoRA 6 epochs 2 Layers LR= 3×10^{-4}	0.766114	0.857946	0.950405	0.781882	0.618781	0.400663	0.765559	0.311214
LoRA 4 epochs 3 Layers LR= 5×10^{-5}	0.798461	0.881278	0.941787	0.828075	0.521766	0.349841	0.748229	0.299813

A.4. Learning rate search (Lowest Train Loss checkpoint)

Table A.21: Learning rate selection (Comprehensible Concat fusion last 2 layers)

LR	LR scheduler	warmup ratio	train_loss	val_loss	Accuracy	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	F1
1×10^{-5}	linear	0.05	0.399000	0.450443	0.846543	0.967697	0.858567	0.735632	0.594199	0.881245	0.511457	0.909871
3×10^{-5}	cosine	0.05	0.345000	0.450598	0.854750	0.969199	0.866542	0.745977	0.612519	0.892385	0.523668	0.915000
5×10^{-5}	linear	0.1	0.332200	0.479402	0.871276	0.965242	0.889346	0.704598	0.593943	0.892094	0.535608	0.925741
1×10^{-4}	cosine	0.05	0.319100	0.502448	0.871276	0.963113	0.891464	0.685057	0.576522	0.890909	0.532229	0.925904
1.5×10^{-4}	cosine	0.05	0.326900	0.501624	0.873300	0.965201	0.891713	0.703448	0.595162	0.888367	0.531883	0.927003
3×10^{-4}	cosine	0.05	0.373600	0.512036	0.862619	0.965385	0.879252	0.709195	0.588448	0.880094	0.517053	0.920308

Table A.22: Learning rate selection (Repetition Sum fusion last 2 layers)

LR	LR scheduler	warmup ratio	train_loss	val_loss	Accuracy	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	F1
1×10^{-5}	linear	0.05	0.253500	0.263448	0.932835	0.982924	0.940423	0.874595	0.815017	0.962942	0.760245	0.961203
5×10^{-5}	linear	0.1	0.193900	0.294793	0.947165	0.981813	0.958028	0.863784	0.821812	0.962250	0.789183	0.969775
7.5×10^{-5}	cosine	0.05	0.164200	0.332924	0.948411	0.977708	0.963662	0.831351	0.795013	0.960347	0.791702	0.970634
1×10^{-4}	linear	0.05	0.156500	0.351475	0.947913	0.978653	0.962113	0.838919	0.801032	0.958179	0.788993	0.970313
3×10^{-4}	linear	0.05	0.196000	0.313245	0.951402	0.978875	0.965915	0.840000	0.805915	0.960601	0.796819	0.972352

Table A.23: Learning rate selection (Grammatical Sum fusion last 2 layers)

LR	LR scheduler	warmup ratio	train_loss	val_loss	Accuracy	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	F1
1×10^{-5}	linear	0.1	0.475700	0.606159	0.814704	0.945161	0.842893	0.562577	0.405470	0.782092	0.349396	0.891102
7.5×10^{-5}	linear	0.05	0.472900	0.633520	0.816698	0.945297	0.845109	0.562577	0.407687	0.776248	0.343356	0.892400
1×10^{-4}	linear	0.05	0.439800	0.690948	0.827414	0.943304	0.859795	0.537794	0.397589	0.763121	0.341717	0.899616
1.5×10^{-4}	cosine	0.05	0.448200	0.671876	0.828037	0.944428	0.859379	0.547708	0.407087	0.772653	0.349608	0.899898
3×10^{-4}	cosine	0.05	0.554700	0.614722	0.827290	0.940350	0.862704	0.510533	0.373237	0.761143	0.325977	0.899855

A.5. Regularization (Highest Val ROC AUC checkpoint)

Table A.24: Regularization (Comprehensible Concat Fusion 2 layers, LR = 1×10^{-4})

Dropout	weight decay	train_loss	val_loss	Accuracy	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	F1
<i>Embeddings = 0.05</i> <i>Transformer = 0.05</i> <i>Head = 0.1</i>	0.005	0.403000	0.416307	0.829455	0.972553	0.834517	0.782759	0.617276	0.891234	0.511756	0.898263
<i>Embeddings = 0.075</i> <i>Transformer = 0.075</i> <i>Head = 0.25</i>	0.05	0.415000	0.410779	0.828780	0.975015	0.831526	0.803448	0.634975	0.893530	0.511533	0.897572
<i>Embeddings = 0.1</i> <i>Transformer = 0.1</i> <i>Head = 0.1</i>	0.01	0.413000	0.413037	0.833614	0.972563	0.839252	0.781609	0.620862	0.892478	0.516519	0.901003
<i>Embeddings = 0.15</i> <i>Transformer = 0.1</i> <i>Head = 0.3</i>	0.005	0.334700	0.492255	0.862170	0.966004	0.878131	0.714943	0.593073	0.890974	0.525359	0.919974
<i>Embeddings = 0.1</i> <i>Transformer = 0.2</i> <i>Head = 0.4</i>	0.05	0.340900	0.483709	0.868016	0.965991	0.884860	0.712644	0.597503	0.892266	0.532610	0.923647
<i>Embeddings = 0.2</i> <i>Transformer = 0.25</i> <i>Head = 0.4</i>	0.1	0.430900	0.419473	0.819337	0.974845	0.820935	0.804598	0.625532	0.890647	0.501665	0.891294

Table A.25: Regularization (Repetition Sum Fusion 2 layers, LR = 1×10^{-4})

Dropout	weight decay	train_loss	val_loss	Accuracy	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	F1
<i>Embeddings = 0.05</i> <i>Transformer = 0.05</i> <i>Head = 0.2</i>	0.05	0.227400	0.242805	0.947539	0.982936	0.957324	0.872432	0.829756	0.965403	0.799363	0.969961
<i>Embeddings = 0.075</i> <i>Transformer = 0.075</i> <i>Head = 0.25</i>	0.1	0.237300	0.248423	0.945171	0.982469	0.955070	0.869189	0.824260	0.966313	0.795471	0.968576
<i>Embeddings = 0.1</i> <i>Transformer = 0.1</i> <i>Head = 0.1</i>	0.01	0.238500	0.246234	0.942555	0.983399	0.951127	0.876757	0.827884	0.965790	0.788465	0.966994
<i>Embeddings = 0.15</i> <i>Transformer = 0.15</i> <i>Head = 0.4</i>	0.01	0.307400	0.251641	0.933707	0.983083	0.941268	0.875676	0.816943	0.964389	0.766775	0.961721
<i>Embeddings = 0.2</i> <i>Transformer = 0.2</i> <i>Head = 0.4</i>	0.005	0.235600	0.268907	0.918505	0.986124	0.920845	0.900541	0.821386	0.965335	0.730790	0.952367
<i>Embeddings = 0.25</i> <i>Transformer = 0.25</i> <i>Head = 0.25</i>	0.1	0.245800	0.280104	0.907165	0.986675	0.907324	0.905946	0.813270	0.962755	0.707124	0.945337
<i>Embeddings = 0.3</i> <i>Transformer = 0.3</i> <i>Head = 0.3</i>	0.05	0.283600	0.317385	0.852212	0.989084	0.842254	0.928649	0.770902	0.960274	0.631763	0.909782

Table A.26: Regularization (Grammatical Sum Fusion 2 layers, LR = 1×10^{-4})

Dropout	weight decay	train_loss	val_loss	Accuracy	Precision	Sensitivity (Recall, TPR)	Specificity (TNR)	Youden's J statistic	ROC AUC	ρ Pearson correlation	F1
<i>Embeddings</i> = 0.015 <i>Transformer</i> = 0.015 <i>Head</i> = 0.25	0.15	0.550400	0.559535	0.752523	0.953537	0.761984	0.667906	0.429890	0.785506	0.336712	0.847066
<i>Embeddings</i> = 0.025 <i>Transformer</i> = 0.025 <i>Head</i> = 0.025	0.15	0.548200	0.558731	0.766106	0.952550	0.778748	0.653036	0.431784	0.786270	0.339951	0.856925
<i>Embeddings</i> = 0.05 <i>Transformer</i> = 0.05 <i>Head</i> = 0.05	0.01	0.552900	0.559376	0.756636	0.953176	0.767110	0.662949	0.430059	0.786065	0.336644	0.850081
<i>Embeddings</i> = 0.075 <i>Transformer</i> = 0.075 <i>Head</i> = 0.25	0.1	0.557400	0.556742	0.756262	0.952996	0.766833	0.661710	0.428543	0.787007	0.341517	0.849839
<i>Embeddings</i> = 0.1 <i>Transformer</i> = 0.1 <i>Head</i> = 0.01	0.05	0.564000	0.559408	0.758380	0.953756	0.768634	0.666667	0.435301	0.783955	0.337804	0.851247
<i>Embeddings</i> = 0.15 <i>Transformer</i> = 0.15 <i>Head</i> = 0.15	0.03	0.521300	0.610433	0.829283	0.941825	0.863536	0.522924	0.386460	0.786600	0.361328	0.900983
<i>Embeddings</i> = 0.2 <i>Transformer</i> = 0.2 <i>Head</i> = 0.3	0.05	0.511500	0.626379	0.821059	0.946280	0.849266	0.568773	0.418039	0.781895	0.352626	0.895152

A.6. Model Architectures

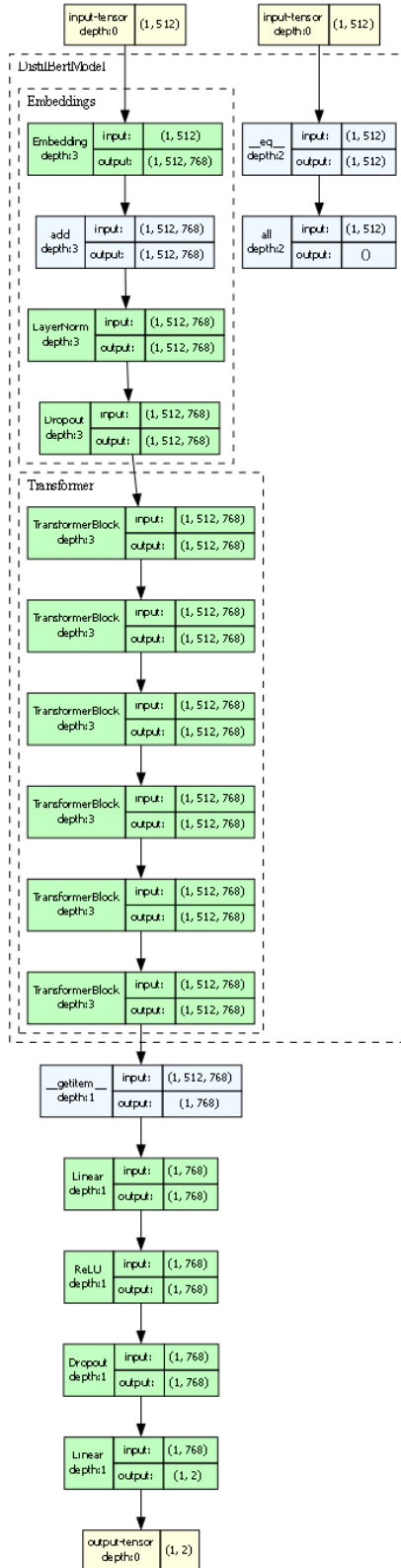


Figure A.1: HF DistilBERT Sequence classifier (Torchview)

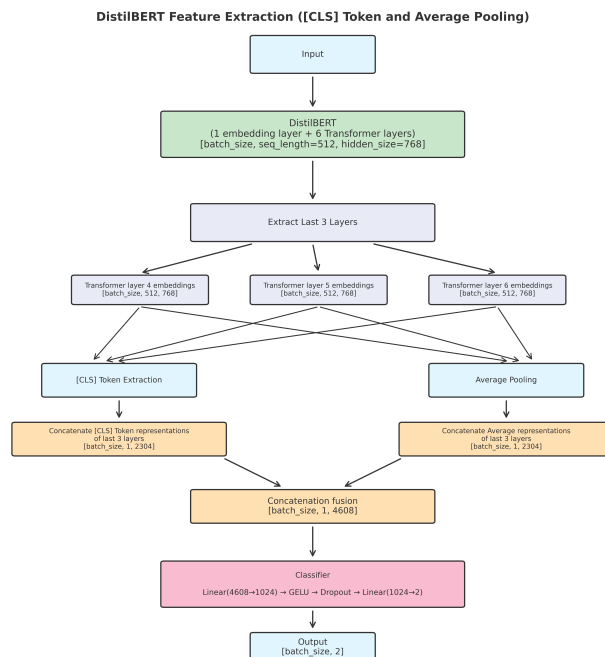


Figure A.2: DistilBERT Feature Extractor

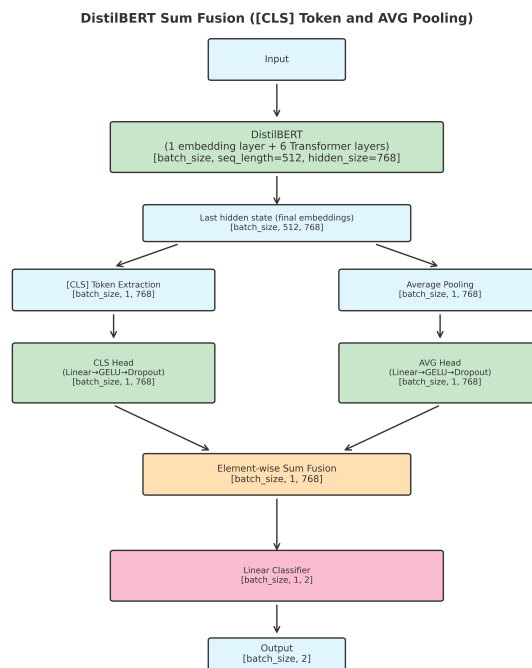


Figure A.3: Sum Fusion

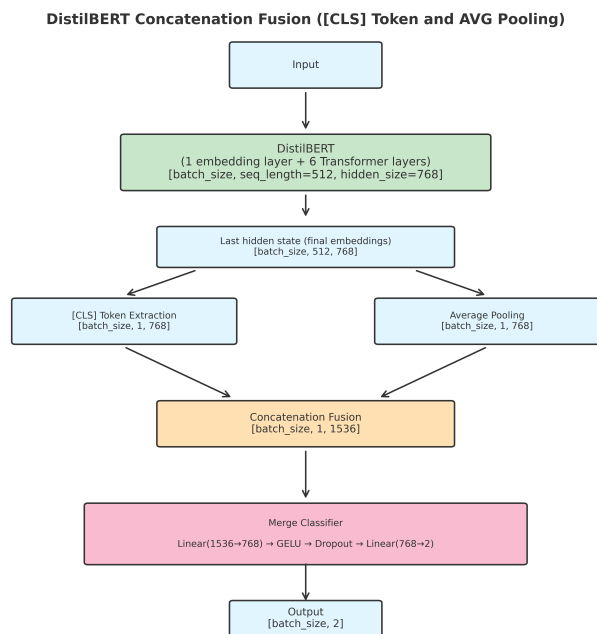


Figure A.4: Concat Fusion

A.7. Tuning of decision threshold in MRPC

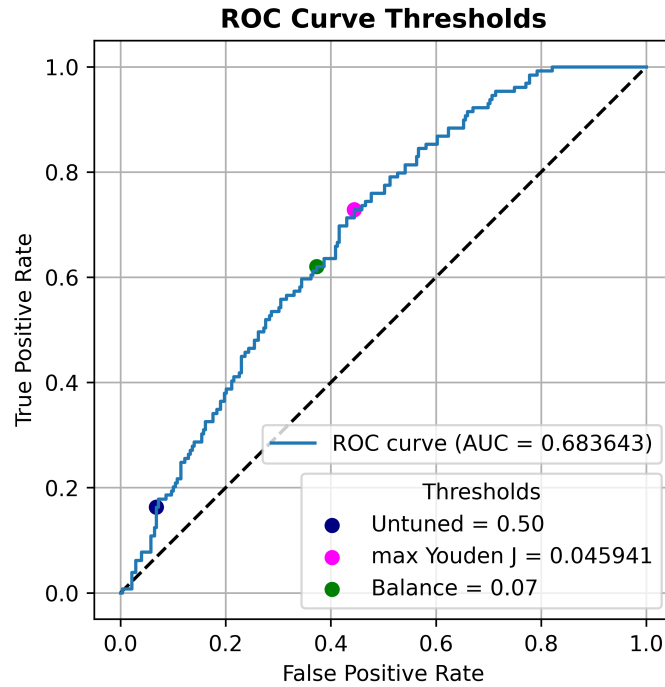


Figure A.5: MRPC validation set tuning of ROC AUC

TN = 720 FP = 427 FN = 175 TP = 403		
Accuracy	: 0.6510	1123 / 1725
F1 Score	: 0.5724	
Precision	: 0.4855	403 / 830
Sensitivity (Recall, TPR)	: 0.6972	403 / 578
Specificity (TNR)	: 0.6277	720 / 1147
Youden's J statistic	: 0.3250	
ROC AUC	: 0.7053	
Pearson correlation	: r = 0.3061	p = 9.6844×10 ⁻³⁹

Table A.27: MRPC test set results after tuning threshold using validation set

A.8. Feature Extractor model results for MELA

TN = 1322 FP = 703 FN = 2370 TP = 3026		
Accuracy	:	0.5859 4348 / 7421
F1 Score	:	0.6632
Precision	:	0.8115 3026 / 3729
Sensitivity (Recall, TPR)	:	0.5608 3026 / 5396
Specificity (TNR)	:	0.6528 1322 / 2025
Youden's J statistic	:	0.2136
ROC AUC	:	0.6468
Pearson correlation	:	$r = 0.2299$ $p = 1.4291 \times 10^{-89}$
MCC	:	0.1903

Table A.28: Feature extraction DistilBERT trained on the MELA training set

B | Appendix B

B.1. Subjectivity in Philosophy: Consciousness

The topic of how human beings developed and exhibit extraordinary perception and information processing capabilities has been the subject of large multidisciplinary research in the fields of philosophy of the mind, epistemology, cognitive science, and evolutionary biology. A comprehensive philosophical overview of subjectivity can be found in [23], where the author characterizes subjectivity as referring to an individual subject, and in particular to mental and physical states associated with consciousness. A fundamental aspect of subjectivity, explained by [23], is that individuals derive information from perceptions to construct beliefs and concepts. According to [23], perceptions are:

the recognition and interpretation of sensory stimuli, (...) we compare, we remember, we classify, we anticipate, we predict, (...). Perceptions are sensations associated with ideas and thoughts [23, p. 224]

The comparison of human and machine perception is a topic that has captivated researchers since the early days of Artificial Intelligence [24]. The issue is stated clearly in *Moravec's paradox*:

it is comparatively easy to make computers exhibit adult-level performance in solving problems (...), and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility. (...) we make the difficult look easy. (...) the most powerful aspects of thought are unconscious, inaccessible to mental introspection, and thus difficult to set down formally. [24, pp. 15-16]

A particularly engaging discussion on what makes human cognition so special is found in [25], whose authors argue that humans, in addition to *inductive reasoning* (based on generalizing observations), and *deductive reasoning* (based on the use of axioms to derive conclusions), have an invaluable capacity for *abductive reasoning* based on exploring minimally complex hypotheses and structures to explain phenomena [25, 26]. This type of reasoning is distinguished by its relation to randomness, imagination, and creativity [25]. Equally noteworthy is the low number of examples or data that humans need to

consume to be able to structure reality [222, 25].

For [25], the creation of information is paradoxical in nature: it appears in the form of notable differences during the process of explaining the world in terms of patterns that describe similarities. In this view, humans are natural individual and collective creators of information by virtue of their capacity to recognize and restructure patterns [25].

B.2. Subjectivity and Scientific Objectivity

A remarkable work reappraising subjectivity is [36], where a proposal is presented for its decomposition and reframing as “awareness of multiple perspectives and context dependence” [36, p. 967]. The authors hold that “multiple perspectives correspond to multiple lines of reasoning” [36, p. 992] and that “different people bring different information and different viewpoints, (...) they will use scientific results in different ways. (...) subjectivity can be productive and helpful” [36, p. 975]. Although focused on statistics, the theoretical and philosophical arguments presented in [36] are easily extended to related fields in science and engineering, as are several of its practical insights. In particular, statistics presents a duality in foundations and application as researchers explore complementary and competing axioms, hypotheses, and proofs proposing methods, estimators and models that often are used in practical settings with loose scientific criteria but achieving notable context-specific success [36]. Thus, data quality and statistics share the concern of building solid abstract theoretical foundations, but flexibility is widespread in practice and even appears to be a beneficial factor [36].

The authors of [36] argue that objectivity is ascertained by observer agreement that results from a consensus dynamic based on a sustained clear communication that includes but is not limited to the measurements and artifacts of science such as formal languages and models [36]. Indeed, for [36], the essence of objectivity lies in being able to provide clear arguments and connect them to existing knowledge. Their work also opposes the use of “objective” as “unquestionable” and argues that a naive effort to “appear objective” can be detrimental [36]. Of central importance is the acknowledgment of the shortcomings and risks of arbitrarily invoking assumptions and deploying solutions regarded as universally valid, noting that this widespread habit illustrates how deeply subjectivity pervades our shared understanding of reality, suggesting that the world is far more dependent on individual perspectives than one might initially assume [36].

Furthermore, [36] argues that a quantitative approach often proves insufficient to describe reality in a way that captures the entirety of what is observed, presenting as examples the challenges that disciplines such as political science, macroeconomics and psychology

have endured to propose metrics for topics whose complex nature makes them directly unmeasurable, defending the idea that limitations of quantification should be recognized.

B.3. Subjectivity in Business: Tacit knowledge

In addition to the previously collected philosophical arguments, we can find complementary subjectivity conceptualizations in the field of business management. The importance of subjectivity in the form of tacit knowledge is well explained in [37]. The author of [37] explains that organizations need proactive knowledge management to sustain continuous innovation and stay competitive, making the having and creating actionable knowledge a fundamental priority for success. The source of this new actionable knowledge can often be found inside the same members of the organization, either from a single individual characterized by personal knowledge and experience or at the upper team's level, where diversity within a shared context allows a team to pool, examine and integrate information under a collective perspective [37].

While explicit knowledge is formal and systematic, including elements such as followed structured procedures, stored quantifiable data, guiding principles and performance metrics, tacit knowledge is highly personal [37]. Moreover, the insights of tacit knowledge are linked to the context where individuals work, the activities they participate in, and the tasks they perform [37]. Another key point defended by [37] is that “experts” holding this valuable tacit knowledge can be found in any part of the organization, at any department or level of hierarchy.

Two key components of tacit knowledge are identified by [37]: (1) procedural *know-how* knowledge, which is informal in nature but valuable for technical-practical use because it allows a person to be skilled in a task, and (2) a *cognitive perception* of the world comprising implicit mental models, beliefs and views, which, despite their complexity and specificity, are not available in written or explicit forms, being simply understood and taken for granted by the people using them. On this matter, [8] explains that:

Enterprises strive to convert tacit “knowledge” into data. (...) “data” are the raw material for “information”, which in turn are the raw materials for “knowledge”. (...) knowledge created or developed by an individual or group must eventually become structured data so others can apply that knowledge. [8, ch. 34, p. 4]

The author of [37] describes that to leverage tacit knowledge, organizations should implement a “spiral of knowledge”, based on processes of “articulation” where tacit knowledge becomes explicit, and “internalization” where explicit knowledge is used to extend the

knowledge base of individuals. These business domain conceptualizations of explicit and tacit knowledge align to the views of [36, 30], in recognizing that a holistic approach can overcome the limitations of using exclusively quantifiable data representations, powering data quality enhancement through diversity and context awareness.

List of Figures

- 1.1 Views of Data quality 7
- 1.2 Problems of subjectivity 17
- 1.3 Benchmark Gaps 29
- 1.4 Role Gap Plot 30
- 1.5 Role average scores 30

- A.1 HF DistilBERT Sequence classifier (Torchview) 126
- A.2 DistilBERT Feature Extractor 126
- A.3 Sum Fusion 127
- A.4 Concat Fusion 127
- A.5 MRPC validation set tuning of ROC AUC 128

List of Tables

1.1	Text quality dimensions	19
1.2	General guidelines presented in [16] for Likert-scale based questionnaires .	25
1.3	Overview of the Information Quality Assessment (IQA) survey [20, 14]	26
1.4	Dimensions for unstructured data	38
1.5	Indicators for unstructured data	38
1.6	Analysis approaches for Big Data (adapted from [76])	40
2.1	Confusion matrix for binary classification.	57
3.1	Models of exploration phase	75
3.2	Comprehensible Test results	77
3.3	Repetition Test results	77
3.4	Grammatical Test results	77
3.5	DeLong Test with Holm correction $\alpha = 0.05$ (<i>Comprehensible</i>)	78
3.6	DeLong Test with Holm correction $\alpha = 0.05$ (<i>Repetition free</i>)	78
3.7	DeLong Test with Holm correction $\alpha = 0.05$ (<i>Grammatical</i>)	78
3.8	Comparison to mT5 (SEAHORSE Test set)	79
3.9	mFACE 5 common languages with SEAHORSE results (Comprehensible)	80
3.10	mFACE all 45 languages results (Comprehensible)	80
3.11	BREAD results (Repetition)	81
3.12	MRPC results (Repetition)	82
3.13	PAWS-X 3 common languages with SEAHORSE results (Repetition) . . .	83
3.14	PAWS-X 7 languages results (Repetition)	83
3.15	MELA results 3 common languages with SEAHORSE results (Grammatical)	84
3.16	MELA results all 7 languages results (Grammatical)	84
A.1	Hyperparameters (exploration phase)	117
A.2	LoRA hyperparameters (exploration phase)	117
A.3	Comprehensible validation	118
A.4	Comprehensible test	118
A.5	Repetition validation	119

A.6 Repetition test	119
A.7 Grammatical validation	119
A.8 Grammatical test	120
A.9 Comprehensible validation (Feature Extraction 12 and 4 epochs)	120
A.10 Comprehensible test (Feature Extraction 12 and 4 epochs)	120
A.11 Repetition validation (Feature Extraction 12 and 4 epochs)	120
A.12 Repetition test (Feature Extraction 12 and 4 epochs)	121
A.13 Grammatical validation (Feature Extraction 12 and 4 epochs)	121
A.14 Grammatical test (Feature Extraction 12 and 4 epochs)	121
A.15 Comprehensible validation (LoRA 4 and 6 epochs)	121
A.16 Comprehensible test (LoRA 4 and 6 epochs)	122
A.17 Repetition validation (LoRA 4 and 6 epochs)	122
A.18 Repetition test (LoRA 4 and 6 epochs)	122
A.19 Grammatical validation (LoRA 4 and 6 epochs)	122
A.20 Grammatical test (LoRA 4 and 6 epochs)	123
A.21 Learning rate selection (Comprehensible Concat fusion last 2 layers)	123
A.22 Learning rate selection (Repetition Sum fusion last 2 layers)	123
A.23 Learning rate selection (Grammatical Sum fusion last 2 layers)	123
A.24 Regularization (Comprehensible Concat Fusion 2 layers, $LR = 1 \times 10^{-4}$)	124
A.25 Regularization (Repetition Sum Fusion 2 layers, $LR = 1 \times 10^{-4}$)	124
A.26 Regularization (Grammatical Sum Fusion 2 layers, $LR = 1 \times 10^{-4}$)	125
A.27 MRPC test set results after tuning threshold using validation set	128
A.28 Feature extraction DistilBERT trained on the MELA training set	129

Acknowledgements

I would like to thank my parents, Liliana and Francisco, for their unconditional love and support, and for motivating me to pursue my dreams.

As a double degree student, I would like to thank my two alma mater institutions, Universidad Nacional de Colombia and Politecnico di Milano, for giving me the opportunity to be part of such an enriching academic, cultural, and personal experience.

I would like to thank all the professors who taught me, and whose dedication, expertise, and passion for teaching have provided me with invaluable knowledge and skills.

I would especially like to thank my supervisor, Professor Cinzia Cappiello, for her guidance in the writing of this work and for her teaching in the Data Quality course, which motivated me to pursue the topic of this thesis.

