Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Context-independent measurement of text data quality

MASTER THESIS

Presented in fulfillment of the degree of Master of Science in Business Intelligence and Process Management at Berlin School of Economics and Law

Supervisors: Prof. Dr. Diana Hristova
diana.hristova@hwr-berlin.de
Prof. Dr. Roland Müller
roland.mueller@hwr-berlin.de

Author: María Hubaldina Lozano González
1944932

Date of Submission: January 19th, 2024

Word Count: 12,144

M. Sc. Business Intelligence and Process Management

Hochschule für Wirtschaft und Recht Berlin | Berlin School of Economics and Law

Abstract

This paper focuses on developing and applying the Context-Independent Measurement of Textual Data Quality. Context-independent refers to text features unrelated to the text's interpretation or meaning, such as grammar. The Context-Independent Measurement of Textual Data Quality is performed by using ten defined and standardized metrics, which can be summed in three dimensions (Accuracy, Completeness, and Consistency) due to the alignment of requirements to create data quality metrics. The thesis includes a case study in which the proposed measurement is implemented, and the resulting data are analyzed to discern the effects of improved Data Quality compared to the performance of a Sentiment Analysis model.

Keywords: Data quality, Text Data, Context-independent Measurement, Data Quality Dimensions, Data Quality Metrics

Sworn Declaration

I now formally declare that I have written the submitted Master's Thesis without anyone else's assistance. Wherever I have drawn on literature or other sources, either in direct quotes or paraphrasing such material, I have referenced the original author or authors and the source where it appeared. I am aware that quotations, or close paraphrasing, from books, magazines, newspapers, the internet, or other sources, which are not marked as such, will be considered an attempt at deception and that the thesis will be graded with a fail. I have informed the examiners and the board of examiners that I have submitted the dissertation, entirely or partly, for other examination purposes.

Berlin, Germany January 19th, 2024

_____

María Hubaldina Lozano González

Table of Contents

List of Figures

List of tables

List of equations

List of abbreviations

DQ Data Quality

DQD Data Quality Dimensions

NLP Natural Language Processing

NLTK Natural Language Toolkit

POS Part Of Speech

TDQ Text Data Quality

1. Introduction

1.1 Motivation

In the current era marked by rapid technological advancements and an increasing reliance on data-driven decision-making, the significance of Data Quality (DQ) has emerged as a pivotal determinant of success for technological innovation projects. However, one of the foremost challenges DQ practitioners face is the nuanced task of defining DQ, mainly as data usage evolves and our dependency on it grows (Sebastian-Coleman, 2012, p. 29). This challenge is paramount and underscores the critical need for a meticulous examination and evaluation of DQ.

Within this context, text data represents a valuable reservoir of information. However, their unstructured nature makes them susceptible to DQ variations, presenting a distinct set of challenges. The impact of high DQ extends beyond mere data management, delivering tangible business value in the form of more informed and faster decisions, increased revenues, reduced costs, and enhanced compliance with legal and regulatory requirements, among other benefits (Gudivada et al., 2017, p. 1). Unfortunately, some companies overlook this reality, leading to failed projects that result in revenue losses and erode customer trust.

DQ introduces several challenges, encompassing the management of diverse data sources with multiple types and complex structures, thus amplifying the complexity of data management (Cai & Zhu, 2015, p. 4). The substantial volume of data further complicates the measurement of its DQ within a reasonable timeframe. Additionally, the rapid pace of data change necessitates advanced processing technologies to meet more demanding requirements.

Recognizing that challenges and problems hinder practical data usage, negatively impacting results and conclusions, it becomes imperative to scrutinize data before employing analysis-oriented tools. If the required quality is not guaranteed, improving DQ becomes crucial

(Oliveira et al., 2005, p. 3). Within this framework, the definition of Text Data Quality (TDQ) indicators holds immense importance. Creating metrics to evaluate these indicators, combining them, and quantifying them into a quality dimension score is pivotal (Taleb et al., 2018, p. 6). This scoring mechanism facilitates the identification of problems and serves as a foundation for enhancing TDQ.

1.2 Research Gap

Research in the field of DQ for text is still in its early stages, although quality measurement in TQD is already an integral part of some research (Kiefer, 2019, p. 2; Cai & Zhu, 2015, p. 4). Well-founded DQ metrics are essential in determining how decision-makers should rely on the underlying data values (Henrich et al. 2017, p. 2).

Kiefer (2019, p. 4) proposes nine indicators to be evaluated for TDQ. In this case, Kiefer mentioned two critical gaps in her paper: (1) the normalized DQ metrics, i.e., transform the percentages and raw numbers into metrics [0,1], and then (2) be able to integrate the DQ into a single measurement. Normalized DQ metrics are crucial for consistency, comparability, interpretability, and integration within a complete framework. It assures that metrics from various sources or indicators share a standard scale, facilitating meaningful comparisons and simplifying their interpretation. Closing this gap leads to enhanced DQ assessments, robust frameworks for analysis, efficient resource allocation, and benchmarking, supporting informed decision-making and effective data management practices. Integrating DQ metrics into a measurement is essential because it allows a straightforward interpretation of results, generating an efficient comparative analysis.

Some of the requirements for the DQ metric suggested by Henrich et al. (2017, p. 5) will be implemented to normalize our metrics: range, scale interval, interpretation, and aggregation. In

addition to being part of data normalization, the aggregation will also be a critical factor in creating the Context-Independent Measurement of TDQ.

DQ requires light and well-defined measurements that can run parallel during each phase. These processes include DQ management, monitoring, and control, the main objective of which is to track any changes that may improve or degrade DQ (Taleb et al. 2015, p. 4). In this thesis, the study case will measure DQ and try to improve it with pre-processing tools such as NLP.

1.3 Research Questions

The goal of this master's thesis is to measure and evaluate the context-independent measurement of TDQ, including a study case where we implement the proposed measurement and, as a result, analyze the effects of DQ improvement. To achieve this goal, the overall research question is:

RQ: *How can we measure context-independent quality in text data?*

To achieve the proposed goal, in addition to the main research question, this study has four research sub-questions that will lead our research process:

RSQ1: *Which context-independent DQ metrics exist in the literature for text data?*

As mentioned, DQ research on the text has just begun (Kiefer, 2019, p. 2; Cai & Zhu, 2015, p. 4), so it is important to take into consideration existing metrics and, based on this, derive additional metrics for context-independent measurement of TDQ.

RSQ2: *What is the best way to normalize metric results, and what are the implications?*

There are existent requirements for DQ metrics (Heinrich et al. 2017, p. 5), which will help us build high-quality metrics, and we need to analyze if it is possible to achieve each of them.

RSQ3: *What approaches are for integrating context-independent quality metrics into a measurement?*

Mylavarapu (2020, p. 27) assures that to evaluate DQ, dimensions of quality should be considered, and these dimensions are composed of the following indicators, which could be enriched by employing the requirement for DQ metrics: 'Sound Aggregation of the Metric Values' (Henrich et al. 2017, p. 3). It explains that our understanding should be built on smaller, detailed evaluations when we look at DQ on a larger scale. In other words, to get a complete picture of the DQ, we start by looking closely at each part and then combining these insights.

1.4 Thesis Structure

This paper presents a Literature Review in Chapter 2 about the main theoretical concepts such as DQ, context-independent metrics, requirements for DQ metrics, and quality assessment framework; the second part of this chapter includes a discussion of related work. Chapter 3 presents the methodology, including the data processing process and the definition of the context-independent metrics, dimensions, and measurement of TDQ. Chapter 4 shows the implementation of the Context-Independent Measurement of TDQ in a case study. The results will then be discussed and analyzed in Chapter 5, where we can find the results and their comparison with Sentiment Analysis Performance. Finally, Chapter 6 presents the conclusion, including the answers to the research question, the limitations, and the future directions of this research.

2. Theoretical Background

2.1 Data Quality

Data is defined as high quality if it meets specific quality criteria or is consistent with the expected goal (Mylavarapu, 2020, p. 24). These criteria are called Data Quality Dimensions (DQD) (Makhoul, 2022, p. 2).

DQD is calculated through one or more metrics that should have an interval between [0,1], where 0 indicates the lowest quality, and 1 indicates the highest quality (Kiefer, 2016, p.16). Normalization becomes essential to maintain uniformity and comparability among different metrics, facilitating the aggregation of diverse metrics. According to Suraj Juddoo (2015, p. 1), the components of DQ are: DQD, which indicates what we measure to know how to improve quality, then we have the metrics which demonstrate how to measure and quantify these DQD, and in the lowest level we have data profiling, data cleaning and discovery of DQ rules (Figure 1). Authors as Azeroual et al. (2018, p. 11) and Juddoo (2015, p. 5) suggest the normalization of DQD, such as accuracy, completeness, and consistency, can be calculated by reducing the result of the division: the number of unwanted results between several of all results, from the total, it means 1.

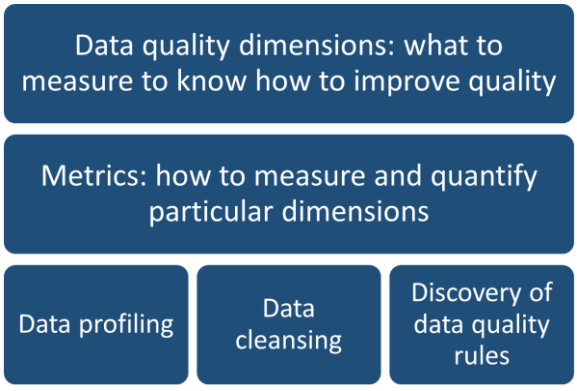


Figure 1. Data Quality Components (Juddoo, 2015, p. 1)

Wang & Strong (1996, p.15-18) suggest four categories for DQD: Intrinsic DQ, Contextual DQ, Representational DQ, and Accessibility DQ (Figure 2). Intrinsic DQ indicates that the data have quality in isolation; it includes Believability, Accuracy, Objectivity, and Reputation. Contextual DQ emphasizes the condition to consider context, including value-added, relevance, timelessness, completeness, and appropriate amount of data. Representational DQ highlights data format (Concise representation and Representational consistency) and data meaning (Interpretability and Ease of understanding). Accessibility DQ indicates that the system should be accessible and secure.

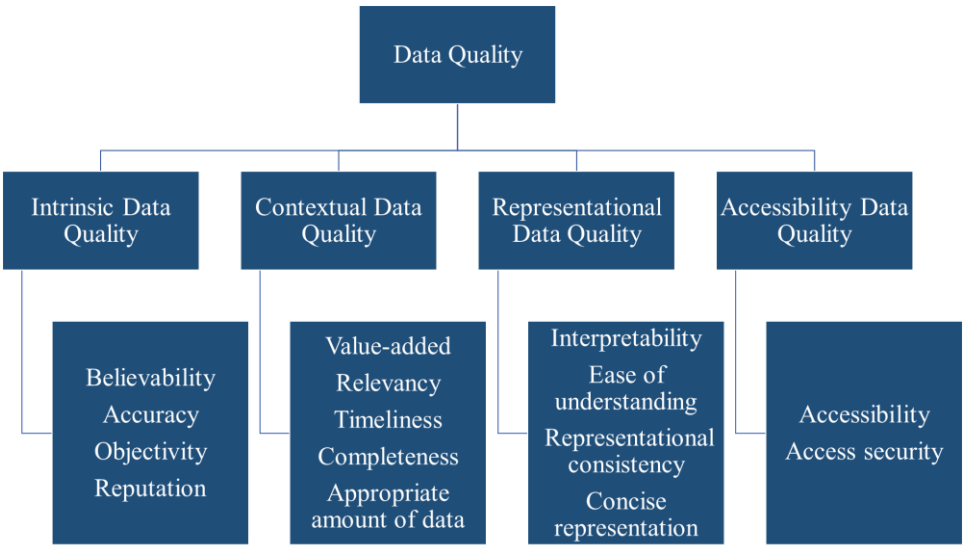


Figure 2. A Conceptual Framework of Data Quality (Wang & Strong, 1996, p. 17)

On the other hand, Taleb et al. (2015, p. 3) propose two categories for DQD based on Wang & Strong (1996), diving the DQD into Intrinsic (native data attributes) and Contextual (related to the content) (Figure 3). On the intrinsic aspect, completeness evaluates the absence of missing entries or values, consistency examines the format and structure, accuracy ensures the correctness of recorded values, and timeliness assesses the currency of the data.

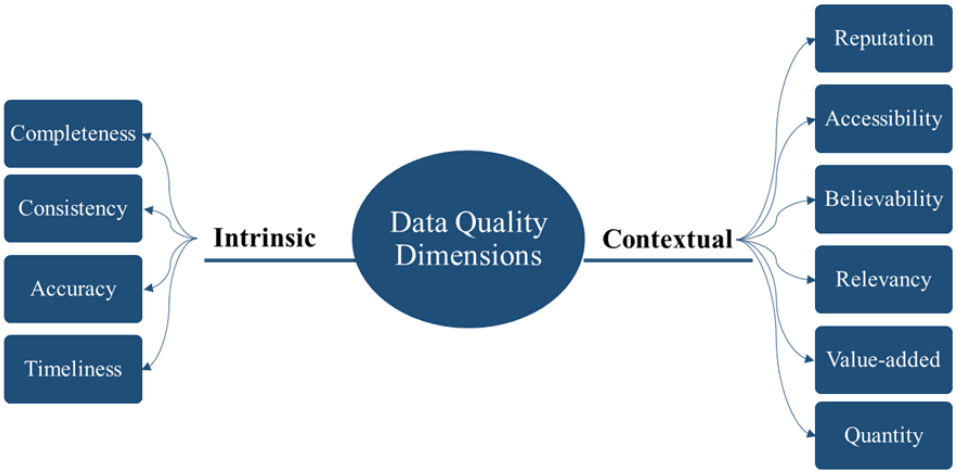


Figure 3. Data Quality Dimensions (Taleb et al. 2015, p. 3)

Most of these dimensions were designed for structured data without considering the analysis dimensions of interpretability, Relevancy, and Accuracy to assess the quality of unstructured data, and Mylavarapu (2020, p. 86-88) recommends measuring Accuracy and Consistency for TDQ.

DQ is usually linked with data preprocessing, profiling, and cleaning for follow-up tasks, such as data integration or data analysis (Ehrlinger & Wöß, 2022, p.1). Unstructured data, as text data, is a generic sequence of symbols, usually encoded in natural language (Sidi et al. 2013, p. 3). The absence of a defined structure in text data makes it sensitive to changes, so it can change dramatically with minor revisions or improvements (Mylavarapu, 2020, p. 93). These little improvements can be applied through NLP, which Azeroual et al. (2018, p. 9) defined as "an application of computational linguistics, which is responsible for the communication and interaction of humans and computers." NLP aims to understand human language, from single words to large complex text, and process its content or, depending on certain conditions, generate natural language as a result. The NLP methods are used to create a structure in text documents to understand its meaning (Azeroual, 2019, p. 4). The first steps in preprocessing data suggested by Nesca et al. (2022, p. 3) include removing stop words (common words in a language), removing punctuation, tagging part of speech (identity or labeling) (Figure 4), and

transforming abbreviation into words or phrases so that they can be understandable. Once data has been collected and cleaned, the next step is pattern recognition (Vezquezsoler & Yankelevich, 2001, p. 7).

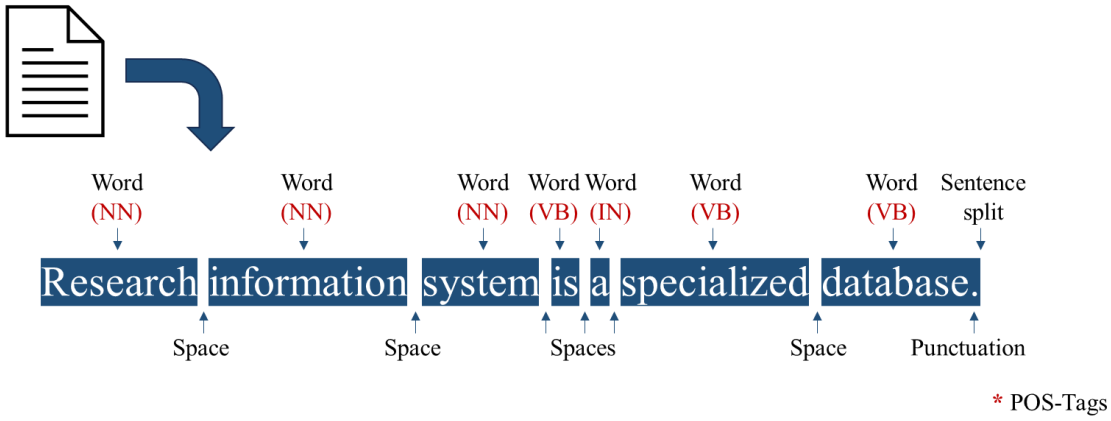


Figure 4. POS data annotation (Azeroual, 2019, p.4)

2.2 Context-independent Metrics of Text Data Quality

DQ metrics are a key element in the estimation of the importance of data-driven decisions (Ehrlinger & Wöß, 2022, p. 1) that provide unbiased information about data (Wang et al. 2003, p. 5).

Ge & Helfert (2007, p.5) summarize DQ problems for context-independent (Table 1) from two perspectives: 1) Data perspective, which shows problems that could be reduced or solved by data cleaning algorithms, data mining rules, statistical process control or dictionary matching routines and 2) User perspective that automated processes cannot solve. These authors consider it essential to define, measure, analyze, and improve DQ to deliver high-quality information to consumers (Ge & Helfert, 2007, p. 11).

Context Type	Data perspective	User perspective
Context-independent	• Spelling error • Missing data • Duplicate data • Incorrect value • Inconsistent data format • Outdated data	• The information is inaccessible • info information is insecure • The information is hardly retrievable

	• Incomplete data format • Syntax violation • Unique value violation • Violation of integrity constraints • Text formatting	• The information is difficult transformation • Errors in the information transformation
Context-dependent	• Violation of domain constraint • Violation of organization's business rules • Violation of company and government regulations • Violation of constraints provided by the database administrator	• The information is not based on fact • The information is doubtful credibility • The information presents an impartial view • The information is irrelevant to the work • The information consists of inconsistent meanings • The information is incomplete • The information is compactly represented • The information is hard to manipulate • The information is hard to understand

Table 1. Classification of DQ problems (Ge & Helfert, 2007, p.5)

Cornelia Kiefer (2019, p. 3-6) presents 9 metrics for TDQ (Table 2) divided into two components: text data and text analysis modules. Table 2 shows an additional column with this information.

Group	Indicator ID	Indicator Description
Data	1	Percentage of abbreviations
	2	Percentage of spelling mistakes
	3	Lexical diversity
	4	Percentage of uppercased words
	5	Percentage of ungrammatical sentences
	6	Average sentence length
Text Analysis Modules	7	Fit of (default) training data
	8	Confidence of standard processing modules
	9	Percentage of unknown words

Table 2. Text Data Quality Indicators (Kiefer, 2019, p. 3-6)

Indicator 1. The percentage of abbreviations is determined based on the Stanford Named Entity Recognizer in Python, which is a classifier to recognize named entities such as persons, cities, and company's names. This metric was adapted to take into consideration dimensions such as

word length, symbols, contains period, sentences dependencies, sequence of vowels and consonants and wordform (sequence of upper and lowercased characters).

Indicator 2. The percentage of spelling mistakes is calculated using Python implementation PyEnchant, but any other spell-checking module could be used.

Indicator 3, 4, 6 & 9. Lexical diversity, Percentage of uppercased words, Average sentence length and Percentage of unknown words are measured using standard methods of Natural Language Toolkit (NLTK) in Python for counting words depending on the objective. Example: Lexical diversity = unique words/ total amount of words.

Indicator 7. Fit of (default) training data assesses the text similarity using Cosine Similarity metric from between an operational dataset and the default training data.

Indicator 8. Confidence of standard processing modules can be evaluated for many classifiers such as part-of speech tagger. This development followed the OpenNLP documentation library, and the confidence value is expressed as a number in the interval [0,1].

Crossley (2020, p. 4) considers that lexical quality in text could be measured with lexical diversity (number of unique words), lexical density (number of content words: nouns, verbs, adjectives, and adverbs) and lexical sophistication (take the difference between counting the number of letters in a word versus calculate the actual word frequency).

2.3 Requirements for Data Quality Metrics

Heinrich et al. (2017, p. 4-5) categorized existing DQ requirements in six groups (Table 3), some requirements are within more than one group, that provide freedom for interpretation depending on the objective or application of the metrics. Group 1 covers requirements for DQ metrics to take values within a given range [0, 1], defining lower and high quality. Group 2 provides the requirements concerning the scale of measurement of the metric values. Group 3 comprises requirements that call for an interpretation of the metrics values, for example, simple

ratio where the metric could be interpreted as a percentage. Group 4 requirements establish that DQ metrics should be able to appropriately consider the context of application, for instance by means of weightings that decrease or increase the influence of contextual features. Group 5 refers to the consistent aggregation of the metric values on different data view levels. Group 6 deals with the enforcement of the DQ metric from a cost-benefit perspective.

Group	Keyword	Requirements
1	Range	Normalization, validity range, clarity of definition (range), simple ration (bounded in [0; 1]), representation consistency (range), measurability
2	Scale	Interval scale, definition of scale (scale)
3	Interpretation	Interpretability, clarity of definition (interpretation), simple ratio (interpretation), interpretation consistency (interpretation), comparability, comprehensibility, definition of scale (interpretation), representation consistency (interpretation)
4	Context	Weighted average (context), impartial-contextual consistency, adaptivity
5	Aggregation	Aggregation, consistency, min or max operation, weighted average (aggregation), interpretation consistency (aggregation),
6	Cost	Cost/benefit, feasibility, acceptability, business relevance

Table 3. Groups of Requirements (Heinrich et al. 2017, p. 5)

Heinrich et al. (2017, p. 8-15) propose the following five requirements for DQ metrics based on the mentioned groups to support decision-making under uncertainty and economically oriented DQ management.

Requirement 1. Existence of Minimum and Maximum Metric Values: The justification for this requirement is that exclusively one metric value must represent an excellent DQ (1) and another one the poor DQ (0) (Heinrich et al. 2017, p. 8-9).

Requirement 2. Interval-Scaled Metric Values: this requires that differences and intervals be meaningful. If the metric values are scaled, they will show a concise interpretation that helps comprehend the DQ level's real meaning. A good example cited is: “A metric value of 0.6 is twice as large as a metric value of 0.3” (Heinrich et al. 2017, p. 10-11).

Requirement 3. Quality of the Configuration Parameters and the Determination of the Metric Values: “It must be possible to determine the configuration parameters of DQ metric according to the quality criteria objectivity, reliability and validity”. Objectivity expresses the degree to which the parameters and values are independent of external factors. Reliability indicates the accuracy with a metric is measured, ensuring the reproducibility of the results and validity ensures the degree of reliability, so that the calculation accurately measures what it should measure (Heinrich et al. 2017, p. 11-13).

Requirement 4. Sound Aggregation of the Metric Values: this requirement guarantees that the DQ metric can be applicable to a single data value up to a dataset, such as a database, in a consistent way. Therefore, a DQ metric must be applicable to diverse levels of data visualization and must have a coherent aggregation of these metric values (Heinrich et al. 2017, p. 13).

Requirement 5. Economic Efficiency of the Metric: the expected additional benefit from the intended application of a metric must exceed the expected costs for determining the configuration parameters and metric values. (Heinrich et al. 2017, p. 14-15).

2.4 Quality Assessment Framework

Best practices in DQ Assessment are evolving, recognizing that more data and better models drive better results, ensuring the success of data-driven initiatives (Gudivada, et al., 2017, p. 2).

According to Gudivada, et al. (2017, p.12), DQ assessment is “the process of evaluating DQ to identify errors and discern their implications. The assessment is made in the context of an intended use and suitability of data for that use is evaluated”. Based on Maydanchik (2007, p.25) the goal of DQ assessment is to identify incorrect items of data and estimate the impact on data-driven business processes (Maydanchik, 2007, p. 25).

During DQ Assessment, the DQD are selected, they determine the data characteristics to be evaluated. DQ metrics are chosen based on the DQD. The aggregation of DQ metrics can be

defined as weighted averages of other metrics of more granularity (Serra et al., 2023, p. 6). This structure is represented in Figure 5.

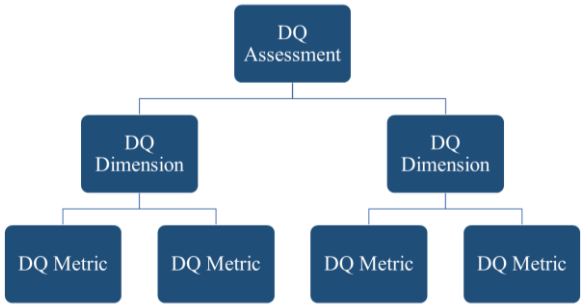


Figure 5. DQ Assessment Structure

2.5 Related Work

The most related research is by Kiefer (2019), where nine context-independent indicators are measured. Looking for similar approaches, Crossley, S. (2020) suggests some metrics that could be used to evaluate lexical sophistication (lexical diversity and density) and syntactic complexity (e.g., greater number of words before the main verb). Lexical diversity was mentioned on Kiefer (2019) paper.

In existing research, the DQ assessment for structured data uses the following dimensions: accuracy, completeness, consistency, and validity, where each of the metrics has a defined formula. For text data, most of the approaches evaluate accuracy and consistency in a similar way to structured data. One research (Sai & Sarma, 2020) shows contribution includes context-dependent extraction for both structured and text data. They also focused on identifying the dataset. Additionally, they conducted quality assessments for structured data, encompassing dimensions like completeness, validity, accuracy, and consistency. Moreover, they performed accuracy assessments and consistency evaluations for text data, all while considering the semantics of the text. In future directions, context-independent measurements are not considered. These measurements are crucial as they offer a more comprehensive and unbiased

assessment of Text Data Quality (TDQ). They remain unaffected by specific contextual or subjective factors, ensuring a standardized and objective evaluation.

The following section explains the methodology for the development of the Context-Independent Measurement for TDQ, including its metrics and dimensions.

3. Methodology

In this section, we explain the methodology for building the proposed Context-independent Measurement of TDQ. This measurement is based primarily on Context-independent metrics of TDQ. The metrics are mainly based on the indicators defined by Kiefer (2019). Each metric will be explained, represented by a formula, then assigned to a dimension (Accuracy, Completeness, or Consistency), and analyzed based on the requirements for DQ Metrics (Heinrich et al. 2017) to standardize and generate high quality metrics. Once the metrics have been defined, the integration of the metric will be shown in DQD and into the single DQ measurement.

3.1 Context-independent Metrics for TDQ

After the literature review, it was decided to propose the following ten metrics (Table 2), considering only quality metrics that are context-independent and can be applied to any text. Then, they were classified into Context-independent dimensions for TQD. After this, an analysis of the metrics was carried out according to the requirements for DQ metrics proposed by Henrich et al. (2017) to standardize them and create a high-quality metric.

The eight metrics were taken from Kiefer (2019, p. 4-6) and Crossley (2020, p. 4), and two more were added based on own criteria that we consider elemental during the data processing (remove punctuation and stop words). In the case of the metrics proposed by Kiefer, only six of them were focused on data. Her research paper mentions nine indicators but three of them are focused on Text Analysis Modules, which are not related to our goal. One of the metrics recommended by Crossley had already been included by Kiefer (Lexical diversity). And the other two are mentioned in the same article claiming to be part of the TQD, enriching lexical richness and sophistication of the text. Table 4 shows a summary of the metrics, including the dimension, the original name of the metric and source.

ID	Metric	Original metric	Source
1	Abbreviation metric	Percentage of abbreviations	Kiefer, 2019, p. 4
2	Spelling mistake metric	Percentage of spelling mistake	Kiefer, 2019, p. 5
3	Unknown words metric	Percentage of unknown words	Kiefer, 2019, p. 6
4	Grammatical sentence metric	Percentage of ungrammatical sentences	Kiefer, 2019, p. 4
5	Lexical density metric	Lexical density	Crossley, 2020, p. 4
6	Lexical diversity metric	Lexical diversity	Kiefer, 2019, p. 5 Crossley, 2020, p. 4
7	Stop word metric	New proposed metric	
8	Average sentence length metric	Average sentence length	Kiefer, 2019, p. 4
9	Uppercased word metric	Percentage of uppercased words	Kiefer, 2019, p. 4
10	Punctuation metric	New proposed metric	

Table 4. Assessment Framework Metrics

Based on papers that classify context-independent metrics into context-independent dimensions (Table 5) and on the characteristics of the selected metrics, it was determined to classify the metrics into the dimensions: Accuracy, Completeness and Consistency. When explaining each of the metrics, the characteristics by which the classification was decided will be mentioned.

	Interpretability	Relevancy	Accuracy	Consistency	Completeness
Kiefer, 2016, p.7	x	x	x		
Mylavarapu, 2020, p. 86-88			x	x	
Taleb et al. 2015, p. 3			x	x	x
Taleb et al. 2016, p. 4			x	x	x

Table 5. DQD for Context-independent metrics

The DQ metric requirements (Henrich et al. 2017, p. 8-15) that are planned to be analyzed are shown in Table 6, the column labeled "Considered" indicates the requirements that will be considered for our metrics.

ID	Requirement	Considered
1	Existence of Minimum and Maximum Metric Values	x
2	Interval-Scaled Metric Values	x
3	Quality of the Configuration Parameters and the Determination of the Metric Values	x
4	Sound Aggregation of the Metric Values	x
5	Economic Efficiency of the Metric	

Table 6. Considered Requirements for Data Quality Metrics

Metric 1: Abbreviation metric

The abbreviation metric was proposed by Kiefer (2019, p. 4), and it is based on the Stanford Named Entity Recognizer, which is a classifier that automatically recognizes named entities such as persons, cities and companies that uses NLP, as POS and syntax. This model had to be trained based on a new data set for the implementation in that paper. In our development, the rationale for the function to identify abbreviations is as follows: for uppercased text, the function looks for 2-3 letters followed by a single point and a space, after removing short, common, and stop words. When the text is not uppercased, two patterns were created. The first pattern identifies acronyms, searching for 2-3 capital letters after a non-uppercased word. The second pattern recognizes titles or date abbreviations (e.g., Mr., Dr., Mon., Aug.), looking for a first capital letter, 2-3 lowercase letters, a dot, and a space. Before applying this pattern identification, the shortest common words and stop words were removed to avoid misunderstandings, especially in uppercased texts. To normalize the metric, the abbreviation percentage is calculated and subtracted from 1. The formula is shown in Equation 1.

$$Abbreviation\ metric = 1 - \frac{Count\ of\ abbreviations}{Total\ count\ of\ words}$$

Equation 1. Abbreviation metric

Dimension: Accuracy. This metric analyzes the correct by ensuring precise usage of abbreviations, a critical factor to avoiding misunderstandings.

Requirement analysis. All four requirements are fulfilled. The minimum and maximum values are delimited in the range between 0 and 1, where 0 indicates low DQ and 1 high DQ, the metric

is on an interval scale, so the variances can be identified. It is a mathematical formula, so it is objective, reliable and valid complying with the quality configuration and determination requirement, and the results can be aggregated into an average for the complete dataset.

Metric 2: Spelling mistakes metric

The spelling mistakes metric is based on the proposed metric "The percentage spelling mistake" by Kiefer (2019, p. 5). Using percentage spelling mistakes, the minimum value will represent good quality and the maximum value poor quality. Our contribution is to normalize the result by subtracting the percentage of spelling mistakes from 1. For our development, we used SpellerCheck function from SpellerCheck library. This function works by comparing each word in each text against a dictionary of correctly spelled words. The spell checker identifies words that do not match any entry in the dictionary and suggests corrections based on the closest matches or possible alternatives. See Equation 2 for the formula of this metric.

$$Spelling\ mistake\ metric = 1 - \frac{Corrected\ words}{Total\ count\ of\ words}$$

Equation 2. Spelling mistakes metric

Dimension: Accuracy. This metric validates if the values were recorded correctly, the aim is to detect typos or misspelled words.

Requirement analysis. The four requirements are fulfilled after the normalization, bounded interval is between 0 and 1, the metric is interval-scaled, so the differences can be determined and meaningful, the metric is a mathematical formula, so it is objective, reliable and valid and the results could be aggregated into an average.

Metric 3: Unknown words metric

The unknown words metric was proposed by Kiefer (2019, p. 6), she applied the standard POS tagger implemented in NLTK to the texts, which has a class for unknow words (X, Appendix

A). An unknown word is typically defined as a word that is not recognized or present in each reference dictionary or language model. In our proposal, we use the Speller Checker, that includes a function called unknown, this function identifies words do not present in its dictionary, helping users recognize and it marks that word as "unknown" or potentially misspelled. Before applying the function, the text is preprocessed, deleting tags, links, punctuation, and numbers, because they were causing noise in the previous results. The formula applied for this metric is provided in Equation 3.

$$Unknown\ word\ metric = 1 - \frac{Count\ of\ unknown\ words}{Total\ count\ of\ words}$$

Equation 3. Unknown words metric

Dimension: Accuracy. This metric identifies unknown words, validating if the values were recorded correctly and can be understandable.

Requirement analysis. All four requirements have been met. To normalize the percentage of unknown words, we subtracted this percentage from 1. This approach ensures that the lowest DQ value is 0, and the maximum is 1. The metric is interval-scaled, it is a percentage so it is a mathematical formula, that is objective, reliable and valid and the results also could be aggregated into an average.

Metric 4: Grammatical sentence metric

The grammatical sentence metric is based on the proposed metric "Percentage of ungrammatical sentences" from Kiefer (2019, p. 4), in this case the article does not provide an explanation about this metric. This metric was one of the most complex metrics to develop in, because first we need to define and understand what a grammatical sentence is and how it can be built. There are four types of sentences: simple sentences, compound sentences, complex sentences, and compound-complex sentences (Das, et al. 2018, p. 3). The grammatical sentences for this metric are simple sentences and complex sentences. Simple sentences are

formed by a noun phrase and a verb phrase, that could include a prepositional phrase. Compound sentences basically are two simple sentences connected with a coordinating or sub coordinating conjugation. In Figure 6, we can find the words that could be included in each of the phrases.

Simple sentence											Compound sentence	
Noun phrase			Verb phrase					Prepositional Phrase			(Sub-) Coordinating conjugation	+ Simple sentence
			Verb phrase		Noun phrase							
PRON			AUX	VERB				ADP	DET	NOUN	CCONJ	
PROPN			AUX	VERB	PRON			ADP	PRON		SCONJ	
DET	NOUN		AUX	VERB	PROPN							
DET	ADJ	PROPN	AUX	VERB	NOUN							
DET	ADJ	NOUN	AUX	VERB	DET	ADJ	PROPN					
			AUX	VERB	DET	ADJ	NOUN					

Must be there
Could be one or more times
Optional

Figure 6. Grammatical sentence's structure, based on Agarwal et al. (2020, p. 6)

To calculate the grammatical sentence metric, it is necessary to count grammatical sentences and divide them into the total sentences (Equation 4).

$$Grammatical\ sentence\ metric = \frac{Count\ of\ grammatical\ sentences}{Total\ count\ of\ sentences}$$

Equation 4. Grammatical sentence metric

Dimension: Accuracy. This metric reflects the precision in the use of proper grammatical structure,

Requirement analysis. The four requirements are fulfilled. The metric has a bounded interval from below an above, where 0 means poor DQ and 1 means perfect DQ, the metric is interval-scaled, because we are talking about a percentage, the metric is a mathematical formula, so it is objective, reliable and valid and the results could be aggregated into an average from individual content in one dataset.

Metric 5: Lexical density metric

The Lexical density metric is defined by Crossley (2020, p. 4) as the number of content words in the text. The content words refer to nouns, adjectives, verbs, and adverbs. These words indicate that a text is a more concentrated and relevant set of words, which contributes to clearer content. The identification of these words will be made using POS and the formula can be found in Equation 5.

$$Lexical\ density\ metric = \frac{Count\ of\ content\ words}{Total\ count\ of\ words}$$

Equation 5. Lexical density metric

Dimension: Completeness. Completeness is associated with lexical density as a higher lexical density implies a greater proportion of meaningful words, contributing to a more comprehensive and complete representation of the text.

Requirement analysis. The four requirements are fulfilled. The minimum and maximum values are bounded in the interval (0,1) where 0 means poor DQ and 1 perfectly good DQ, the metric is interval-scaled, so the differences can be determined and meaningful because it is a percentage and so it is a mathematical formula, that is objective, reliable and valid, and the results could be aggregated into an average.

Metric 6: Lexical diversity metric

The lexical diversity metric was proposed by Kiefer (2019, p.5) and Crossley (2020, p.4). Crossley defines lexical diversity as the number of unique words in text. Lexical diversity is relevant for the quality of data because it provides insights into the richness and variety of vocabulary used in a text. See Equation 6 for metric formula.

$$Lexical\ diversity\ metric = \frac{Unique\ words}{Total\ count\ of\ words}$$

Equation 6. Lexical diversity metric

Dimension: Completeness. Lexical diversity contributes to completeness by suggesting a broader range of unique words, potentially enhancing the overall richness of the text.

Requirement analysis. The four requirements are fulfilled. The existence of minimum and maximum values is already bounded in the right way. The metric is interval-scaled because it is a percentage where mean that is a mathematical formula, that is objective, reliable and valid and the results also could be aggregated into an average in a dataset.

Metric 7: Stop word metric

The stop word metric is the first newly proposed metric. The reason is that during the cleaning process, removing stop words is a common step. A text with many stop words is considered low quality because stop words, typically contribute little to the substantive meaning or information of the text. The formula applied is presented in Equation 7.

$$Stop\ word\ metric = \frac{Total\ count\ \ of\ words\ after\ remove\ stop\ words}{Total\ count\ \ of\ words}$$

Equation 7. Stop word metric

Dimension: Completeness. This metric could be considered in Completeness dimension because the metric is analyzing the percentage of words that will be considered as meaningful.

Requirement analysis. The four requirements are fulfilled. The interval is bounded between 0 and 1, being 1 the most perfectly DQ. The metric is interval-scaled that as a consistent meaning, it is a percentage so it is a mathematical formula, that is objective, reliable and valid and the results also could be aggregated into an average.

Metric 8: Average sentence length metric

The average sentence length metric was inspired by the indicator average length proposed by Kiefer (2019, p.4). This metric was modified to get a result bounded between 0 and 1. Taking into consideration that more than one sentence adds complexity to a text and that short sentences

are better; it was decided to measure the variability between the average length of the sentence and the total length of the text, this variability will be divided by 100, because the values obtained are always greater than 0 and in some cases may even be greater than 100. For more detail, read section 6.3. The applied formula for the average sentence length metric is shown in Equation 8.

$$Average\ sentence\ length\ metric = 1 - \frac{Variability}{100},$$

$$where\ Variability = \frac{Text\ length - Length\ average}{Number\ of\ sentences},$$

$$\text{where } Length\ average = \frac{Text\ length}{Number\ of\ sentences}$$

Equation 8. Average sentence length metric

This example shows a lower Data Quality (DQ) when comparing texts with the same average length but a higher number of sentences. Conversely, another example exhibited the same percentage of reduction, where the metric performed better for shorter sentences, as illustrated in Figure 7.

Length	Sentences	Length average	Receded %	100-%
20	1	20	0	100
20	2	10	5	95
20	3	6.67	4.44	95.56
20	4	5	3.75	96.25
30	1	30	0	100
30	2	15	7.5	92.5
30	3	10	6.67	93.33
30	4	7.5	5.62	94.37
40	1	40	0	100
40	2	20	10	90
40	3	13.33	8.89	91.11
40	4	10	7.5	92.5
50	1	50	0	100
50	2	25	12.5	87.5
50	3	16.67	11.11	88.89
50	4	12.5	9.37	90.62

Same avg length
Same % reduced

Figure 7. Example of Average sentence length metric

Dimension: Consistency. For this metric, we analyze the variability between length of the sentences and try to create a more uniform pattern.

Requirement analysis. The four requirements are not fulfilled, this time the metric is not interval-scaled, because the reduction will change not in a regular or standard way. The existence of minimum and maximum values, the quality configuration parameters and the determination of the metric values and the aggregation requirement are fulfilled.

Metric 9: Uppercased word metric

The uppercase word metric was proposed by Kiefer (2019, p. 4). For this metric, the text is tokenized and then an uppercase word list is created, the percentage of uppercased word is calculated and then the result is subtracted from 1. Formula can be visualized in Equation 9.

$$Uppercased\ word\ metric = 1 - \frac{Uppercased\ words}{Total\ count\ of\ words}$$

Equation 9. Uppercased word metric

Dimension: Consistency. This metric identifies and maintains matching format for words that should be consistently uppercase throughout the text.

Requirement analysis. The four requirements are fulfilled. The existence of minimum and maximum values range is between 0 and 1, where 0 represents the lowest DQ and 1 the highest DQ. The metric is interval-scaled and easily compared, it is applied a mathematical formula, so it is objective, reliable and valid and the results can be aggregated into an average.

Metric 10: Punctuation metric

In this paper, we propose the inclusion of a Punctuation metric. This decision emerged from observations during the development of previous metrics, where punctuation was initially removed as it introduced noise into the results, particularly affecting the spelling mistake and unknown words metrics. Recognizing the significance of punctuation in the overall TDQ, we now consider it crucial to incorporate this aspect into our evaluation. The formula for this metric is shown in Equation 10.

$$Punctuation\ metric = \frac{Length\ of\ the\ text\ without\ punctuation}{Total\ count\ of\ characters}$$

Equation 10. Punctuation metric

Dimension: Consistency. This metric preserves the syntactic and grammatical coherence of text, promoting uniformity across the text.

Requirement analysis. The four requirements are fulfilled. The existence of minimum and maximum values is bounded (0,1). The metric is interval-scaled with a concise interpretation, the quality of configuration parameters and the determination of the metric values is objective, reliable and valid and the results can be aggregated into an average for a complete dataset.

After having defined the DQ metrics that will be used in our proposed Context-independent Measurement of TDQ, the composition of the integrated dimensions and quality measurement will be explained in the next section.

3.2 Context-Independent Measurement of TDQ

The Context-Independent Measurement proposed for TDQ is calculated based on three DQD: Accuracy, Completeness and Consistency. The assignation of the dimension was defined during the specification of each metric. To summarize DQD, the calculations used in each of the dimensions are presented below.

Accuracy dimension consists of calculating the average of the following metrics: Abbreviation, Spelling mistakes and Unknown words (Equation 11).

$$Accuracy = \frac{1}{4}(Abbreviation\ metric\ +\ Spelling\ mistake\ metric + Unknown\ words\ metric\ + Grammatical\ sentence\ metric)$$

Equation 11. Accuracy

Completeness dimension is evaluated with the metric mean of: Grammatical sentences, Lexical diversity, Lexical density and Stop words. According to the calculation shown in Equation 12.

$$Completeness = \frac{1}{4}(Lexical\ diversity\ metric + Lexical\ density\ metric + Stop\ word\ metric + Average\ sentence\ length\ metric)$$

Equation 12. Completeness

Consistency dimension is determined by calculating the average of the following metrics: Average sentence length, Punctuation and Uppercased words. See Equation 13.

$$Consisteny = \frac{1}{2}(Puntuation\ metric + Uppercased\ word\ metric)$$

Equation 13. Consistency

The proposed Context-independent DQ Assessment (Figure 8) is obtained by the weighted average of the dimensions (Equation 14) or the average of the ten defined metrics (Equation 15).

$$Context - independent\ measurement\ of\ TDQ = 0.4(Accuracy) + 0.4(Completeness) + 0.2(Consistency)$$

Equation 14. Context-independent of Text Data Quality by Dimensions

$$Context - independent\ measurement\ of\ TDQ = \frac{1}{10}(Abbreviation\ metric + Spelling\ mistake\ metric + Unknown\ words\ metric + Grammatical\ sentence\ metric + Lexical\ diversity\ metric + Lexical\ density\ metric + Stop\ word\ metric + Average\ sentence\ length\ metric + Puntuation\ metric + Uppercased\ word\ metric)$$

Equation 15. Context-independent of Text Data Quality by Metrics

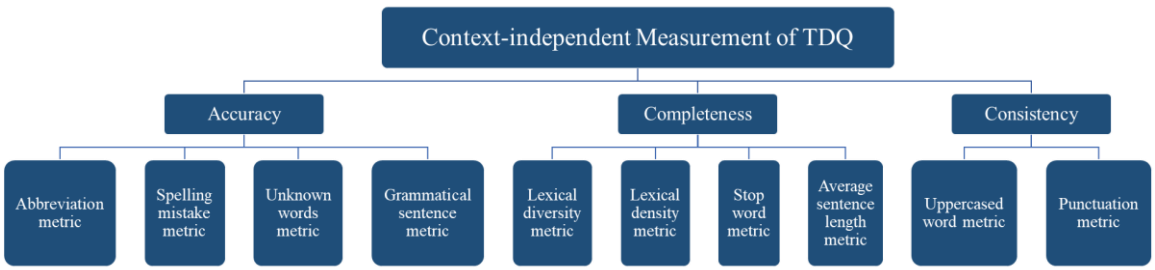


Figure 8. Proposed Context-independent measurement for TDQ: structure

4. Implementation

This section shows the implementation of the context-independent measurement of TDQ on a sample dataset for original text and processed text, then a sentiment analysis model is applied to both columns, with the aim of analyzing Context-independent measurement of TDQ and its effect on the performance of the sentiment analysis model. The development of the implementation is coded in Python version 3.9.12. And the development could be found in the following GitHub link: https://github.com/mlozanog96/Context-independent-Measurement-of-Text-Data-Quality. The structure of the folder including related section is shown in Table 7 where "…/" indicates the root GitHub link. The detailed content will be explained in the mentioned section.

#	Folder	File	Section
1	…/	README.md	4.
2	…/	Dataset.ipynb	4.1
3	…/	utils.py	4.2, 4.3
4	…/	Implementation.ipynb	4.2, 4.3, 4.4
5	…/	Context-Independent Measurement of TDQ Analysis.pbix	5
6	…/Data/Raw Data/Originals	Twitter.csv	4.1
7	…/Data/Raw Data/Originals	Reviews_0-250.csv	4.1
8	…/Data/Raw Data/Originals	News_COVID.csv	4.1
9	…/Data/Raw Data	Twitter_1000.xlsx	4.1
10	…/Data/Raw Data	Reviews_1000.xlsx	4.1
11	…/Data/Raw Data	News_COVID_1000.xlsx	4.1
12	…/Data/Processed Data/	dataset.pkl	4.1, 4.2
13	…/Data/Processed Data/	processed_dataset.pkl	4.2, 4.3
14	…/Data/Results	dataset_TDQ.pkl	4.3
15	…/Data/Results	dataset_to_analyze.xlsx	4.4, 5

Table 7. Context-independent Measurement of TDQ Git Hub Structure

4.1 Data

The only requirement necessary to evaluate a dataset using the proposed Context-independent measurement of TDQ is a table with a column including text on which it is possible to measure TDQ. It is recommended to clean the dataset by removing unnecessary columns, which will help us reduce the size of the dataset allowing us to manage more data efficiently.

The dataset for the implementation consists of a table based on three datasets downloaded (CSV files) from Kaggle. The selection of these datasets was done by looking for a range of different lengths and text types that could be used as classifications during the results. Table 8 shows the dataset name, link, text type (added during the creation of the new dataset) and related files from Table 7 (original file, file with human sentiment column). The human sentiment column was added in the second file to compare it against the result obtained by applying a sentiment analysis model on texts of different qualities.

#	Dataset Name	Link	Text Type	Related files
1	Sentiment140 dataset with 1.6 million tweets	https://www.kaggle.com/datasets/kazanova/sentiment140/	Twitter	6, 9
2	Sephora Products and Skincare Reviews	https://www.kaggle.com/datasets/nadyinky/sephora-products-and-skincare-reviews	Review, review title	7, 10
3	Covid-19 News Sentiment Analysis	https://www.kaggle.com/code/sameer1502/covid-19-news-sentiment-analysis/input	News, news title	8, 11

Table 8. Kaggle datasets

The selected column to analyze from dataset 1 was text (Column F, original dataset without columns name), classified as Twitter in text type. From dataset 2, the chosen columns were "review_text" and "review_title", classified as review and review title. In dataset 3, "Description" and "Headline" columns were elected, classified as news and news title. This classification will be used to compare results.

For each dataset, the first 1000 rows were taken, taking into consideration different aspects:

- Twitter: Rows with English content and enough information (no empty rows after processing function, for example rows only including links and mentions).
- Reviews: Rows with English content in which the review title is not empty.

- News: Rows with English content where the Covid column is 1, indicating it talks about it (when the content was not about Covid it was difficult to classify its sentiment because it was from a region of which I did not have enough knowledge).

The Jupyter notebook file called Dataset (file 2, Table 7) contains the code for the consolidation of analyzed dataset. This code creates a pickle file (file 12, Table 7) with 5 columns: text, human_sentiment, text_type, length, and length_classification. The dataset consists of 5000 rows where min length is 3 characters and maximum length 1354, distributed by 5 text types (equally) and length classification. Length classification has three categories: short (0-50 characters), medium (50-250 characters) and long (more than 250 characters), in a distribution of 26%-46%-27% respectively. Mean length (156) was taken as "medium" category, including a range of +/- 100 characters. For more details about the distribution of the dataset see Table 9. Jupyter notebook file includes some histograms by text type and length classification.

Text Type	Length Overview				Length Classification			
	Mean	Media	Min	Max	Short	Medium	Long	Total
Twitter	75	71	8	150	309	691	0	1000
Review	254	211	50	1354	0	626	374	1000
Review title	20	18	3	52	988	12	0	1000
News	358	360	287	404	0	0	1000	1000
News title	71	71	41	90	11	989	0	1000
Total	156	79	3	1354	1308	2318	1374	5000

Table 9. Dataset Overview: length analysis by text type

4.2 Data Processing

The aim of the data processing is to have two texts with different Context-independent measurement of TDQ. The data processing should help us to improve TDQ.

The python file "utils" (file 3, Table 7), includes a function called processed_text which generates a new column "processed_text", basically the function improves the TDQ in the text by removing usernames, links, punctuation, numbers and stop words, converts contractions into words and text into lower case, as well as corrects spelling mistake in the text. Figure 14 shows

the process of this function. The relevant libraries used in the function are re, nltk, gensim.parsing.preprocessing and spellchecker, find more details in utils.py.

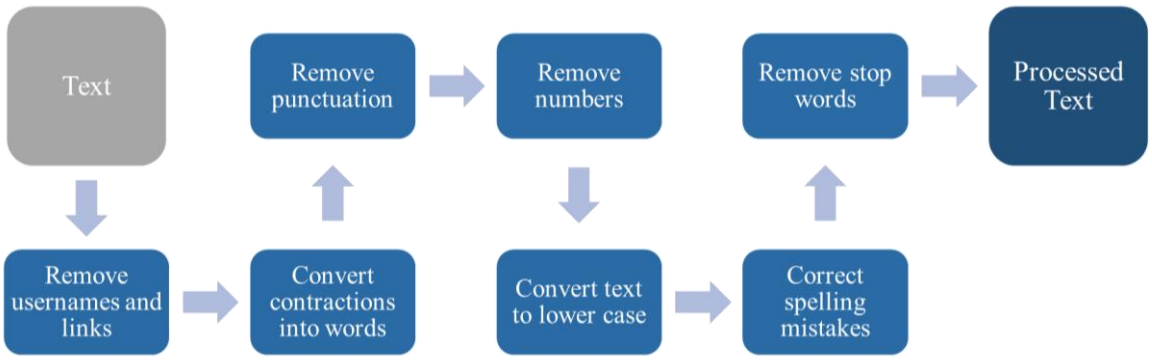


Figure 9. Data Processing Function

The function is loaded into the Jupyter Notebook called Implementation (file 4, Table 7), where the pickle file dataset is loaded, then processed_text function is applied and as a result the processed_dataset pickle file (file 13, Table 7) is generated.

4.3 Context-independent Measurement of TDQ

The first step to get the Context-independent measurement of TDQ is to calculate their metrics. Table 10 lists the metrics and their functions from python file "utlis" (file 3, Table 7). The description of each function is below the table. All metric functions include the same arguments: df (dataset), column (text to be measured), new_column (name of new column). Each metric is applied to the columns text and processed_text in the Jupyter notebook file: Implementation (file 4, Table 7).

#	Metric	Function
1	Abbreviation	abbreviation_metric()
2	Spelling mistake	spelling_mistake_metric()
3	Unknown words	unknown_word_metric()
4	Grammatical sentence	grammatical_sentence_metric()
5	Lexical density	lexical_density_metric()
6	Lexical diversity	lexical_diversity_metric()
7	Stop words	stop_word_metric():
8	Average sentence length	average_sentence_length_metric()

9	Uppercased word	uppercased_word_metric()
10	Punctuation	symbol_punctuation_metric()

Table 10. Context-independent metrics functions

Metric 1. The abbreviation metric function first identifies if the text is uppercased or not, if it is, the code identifies an abbreviation pattern after removing uppercased shortest common and stop words using "remove_stopwords" function from "gensim. parsing.preprocessing" library, and if it is not, the pattern identifies acronyms, titles and dates abbreviations. In both cases a list with abbreviations is created to then measure the abbreviation percentage and subtracted from 1.

Metric 2. The Spelling mistake metric function creates an instance of SpellChecker into a variable to evaluate the spelling mistakes. The data is processed by removing tags, links, punctuation, numbers, and contractions. After processing, the content is divided into words, the "correction" spell check function is applied, and then a difference set is created. The length of this list will be divided by two (because we have the list with the correct and incorrect word) and then divided by the length of the content, obtaining the percentage of misspellings, finally this percentage is subtracted from one.

Metric 3. The Unknown words metric function creates an instance of SpellChecker into a variable, then the text is preprocessed, we tokenize the text splitting it into words, and then the function creates a list of unknown words using unknown function from loaded instance. Then we get the percentage of unknown words, and to normalize the text the percentage is subtracted from 1.

Metric 4. The Grammatical sentence metric function loads the pre-trained English model from spaCy library to preprocess our content, after this, the content per row is divided into a list of sentences, for each sentence we use Part Of Speech (POS) tagger function to transform the sentences into the sentences structure (see Appendix A, for the meaning of the transformation), then a list with the grammatical structure sentences is created, where we can identify the

grammatical sentences through patterns using function “findall” from library “re”. Finally, we divide the count of grammatical sentences by total count of sentences.

Metric 5. The Lexical density metric function loads the pre-trained English model from spaCy library to preprocess the text, then we create a list of content words, where the word is added if the POS taggers are ‘NOUN’, ‘ADJ’, ‘VERB’ or ‘ADV’. After this, the percentage of content words is calculated.

Metric 6. The lexical diversity metric function is calculated using the same method applied by Kiefer (2019, p. 5) through the calculation of unique words using the pandas library, measuring the length of the value set (unique words) divided by the number of words in the text.

Metric 7. For the stop words metric function, the stop words are removed from the text, it is not necessary to load the list again because they were previously loaded (during the cleaning process of abbreviation metric). After using the "remove_stopwords" function from the "gensim.parsing.preprocessing", the function measures the percentage of stop words.

Metric 8. The metric function Average sentence length first loads the pretrained English model from the spaCy library to preprocess the text, then it creates a list by dividing the text into sentences, then it calculates the length of the text minus the average length divided by the number of sentences and the result is validated to check that it is not greater than 100, if so it will be converted to 100, this validation is needed because we can get negative numbers from the formula applied, find the detailed explanation in section 6.3 Limitations, then the resulting value is subtracted from 1.

Metric 9. The Uppercased word metric function is programmed to be used by indicating a data set and a text column, then it splits the text into words, to create a list including uppercased words, then the percentage of uppercased is measured and subtracted from 1.

Metric 10. The Punctuation metric function cleans the text, using the "strip_punctuation" function from "gensim.parsing.preprocessing" library, then another symbols not included in this function are removed. And finally, the percentage of clean text characters is calculated.

Figure 10 illustrates that our second step will be to calculate the Context-independent dimensions of TDQ, and the third step will be to calculate the Context-independent measurement of TDQ. In both steps we will use the columns_average function found in the python utils file (file 3, Table 7) where we must provide as arguments: df (dataset), columns (list of columns to average) and new_column (name of the new column).

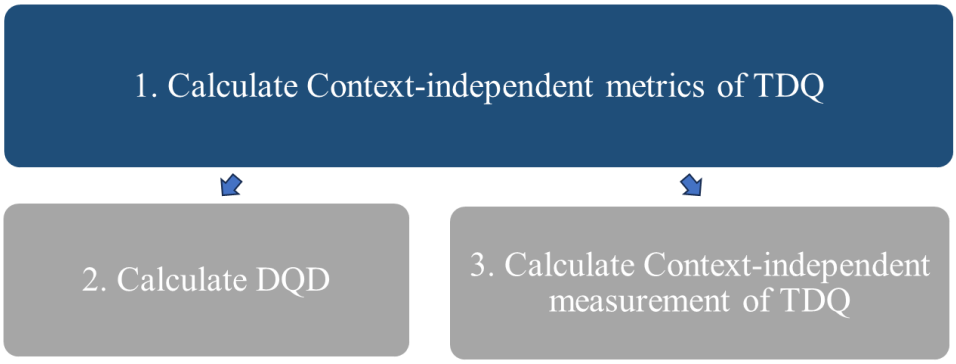


Figure 10.Context-independent measurement of TDQ steps for implementation

Table 11 shows the columns used to calculate each the Context-independent dimensions for TDQ. This is applied to the metrics of the text and the processed text (new columns).

#	Dimension	Columns to average	New columns
1	Accuracy	Abbreviation metric, Spelling mistake metric, Unknown words metric, Grammatical sentence metric	dqd1_accuracy_txt dqd1_accuracy_ptxt
2	Completeness	Lexical density metric, Lexical diversity metric, Stop word metric	dqd2_completeness_txt dqd2_completeness_ptxt
3	Consistency	Average sentence length metric, Uppercased word metric, Punctuation metric	dqd3_consistency_txt dqd3_consistency_ptxt

Table 11. DQD settings for implementation

Finally, to determine the Context-independent measurement of TDQ, the function "columns_average" is used using as column argument all the metrics evaluated for text and

processed text respectively. The new columns are called: “TDQ_txt” and “TDQ_ptxt”. The resultant file from this step is the pickle file “dataset_TDQ” (file 14, Table 7).

4.4 Sentiment Analysis Model

The applied sentiment analysis model belongs to the NLTK library, and is loaded with the SentimentIntensityAnalyzer() function, once the model is loaded, each row of the data set is evaluated by assigning it a polarity score value, the 'compound' column is taken and classified according to Table 12 into three values: positive, negative, and neutral.

Sentiment Classification	Polarity Score Range
Positive	Polarity Score >= 0.05
Neutral	-0.05 < Polarity Score < 0.05
Negative	Polarity Score <= 0.05

Table 12. Polarity Score Ranges for Sentiment Analysis Classification

After applying the sentiment analysis model to the columns: text and processed text in the Jupyter Notebook file “Implementation” (file 4, Table 7), it creates two new columns comparing obtained results with human_sentiment column, and the result is transformed to integer to facilitate the subsequent analysis.

The dataset including all metrics evaluated during the implementation results in an excel called "dataset_to_analyze" (file 15, Table 7), which is analyzed in a Power BI report (file 5, Table 7) in the following section.

5. Results and Discussion

This section contains the results obtained during the implementation, which include the results of the Context-independent metrics, dimensions and measurement of TDQ applied to text and processed text in a dataset, analyzed by length classification (short, medium and long) and text type (twitter, review, review title, news, news title), so as their comparison with the performance of the sentiment analysis model applied to both texts, including the discussion of the findings. The report including all charts, can be found in the Power BI file "Context-Independent Measurement of TDQ Analysis" (file 5, Table 7).

5.1 Context-independent Metrics

The results of the context-independent metrics are presented in Figure 11, where each metric is shown in a different color. The first column lists the name of the metric indicating the text on which it has been performed, followed by a bar displaying the result of the evaluation on each text, and finally in the third column, the differential between the processed text and the original text is reported, this differential representing the increase or decrease per metric.

Figure 11. Context-independent metrics results

Metrics differentials 1 and 6 display values close to zero, denotating an insignificant change after data processing. Differentials of metrics 2, 3, 4, 9 and 10 represent a low-moderate increase

in the metric, going from values greater than 0.01 and lower than 0.15. Metric differentials 5, 7 and 8 show values around 0.25, indicating a moderate improvement of the metric result.

Context-independent metrics by Length Classification

The Context-independent metrics results by length classification (Figure 12) provides the results of each metric evaluated in the text and the processed text according to its length classification: long, medium, and short (assignation conditions in section 4.1) Each metric is represented with a different color.

	long	medium	short	Total
M1: Abbreviation metric (text)	1.00	0.99	1.00	1.00
M1: Abbreviation metric (processed text)	1.00	1.00	1.00	1.00
M2: Spelling mistake metric (text)	0.94	0.91	0.96	0.93
M2: Spelling mistake metric (processed text)	1.00	1.00	1.00	1.00
M3: Unknown word metric (text)	0.93	0.90	0.95	0.92
M3: Unknown word metric (processed text)	1.00	1.00	1.00	1.00
M4: Grammatical sentence metric (text)	0.20	0.30	0.12	0.22
M4: Grammatical sentence metric (processed text)	0.82	0.23	0.03	0.34
M5: Lexical diversity metric (text)	0.44	0.51	0.58	0.51
M5: Lexical diversity metric (processed text)	0.72	0.78	0.76	0.76
M6: Lexical density metric (text)	0.81	0.95	0.99	0.93
M6: Lexical density metric (processed text)	0.81	0.97	0.99	0.93
M7: Stop word metric (text)	0.64	0.71	0.80	0.71
M7: Stop word metric (processed text)	1.00	1.00	0.87	0.96
M8: Average sentence length metric (text)	0.26	0.87	0.99	0.73
M8: Average sentence length metric (processed text)	0.98	1.00	1.00	0.99
M9: Uppercased word metric (text)	0.97	0.95	0.92	0.95
M9: Uppercased word metric (processed text)	1.00	1.00	1.00	1.00
M10: Punctuation metric (text)	0.97	0.96	0.96	0.96
M10: Punctuation metric (processed text)	1.00	1.00	1.00	1.00

Figure 12. Context-independent metrics results by Length Classification

Metric 1 denotes the reduction of abbreviations occurred in the medium-length texts of the processed text; the rest of the texts did not show any change.

For metric 2, no definite tendency is noted. The medium texts are the ones that present more spelling mistakes (0.91), and the short texts have a lower number of spelling mistakes (0.95), while the processed text does not have these mistakes.

In metric 3, it is possible to find that the medium texts are the ones with the highest number of unknown words (0.90) and the short text shows a lower number of them (0.95), while the processed text does not have any unknown words.

It can be appreciated results of metric 4 are low values in comparison with other metrics, in the medium text we can see the highest number of grammatical sentences (0.3) while in the short text a lower value is detected (0.12) if you only compare original text. When analyzing the processed text, we can observe that the short-processed text has a much lower result (0.03), reducing by 75% the value of the short original text, and the most relevant in this metric is to detect the longer processed texts had a higher value (0.82) with an increase of 400% when compared to the original long text.

The metric 5 reveals in the original text: as the length increases, the lexical diversity decreases. The processed text does not have the same result, in fact we can notice a higher lexical diversity in the medium processed texts (0.78). The processing of the data in all the categories increased but it was not in the same proportion.

In metric 6, there is no difference between long and short in both texts. The only remark is an increase of 0.02 in the lexical diversity of the medium processed text.

Metric 7 shows a lower result for the long text and a higher value for the short text, indicating that the long text has a higher number of stop words in the original text. While in the processed text, we can see that the only value that is not 1 is in the short texts (0.87).

We can remark in metric 8 that the value of the original long text is considerably lower (0.26), compared to the short (0.99) or medium text (0.87). Once the text is processed the 3 categories show a higher result (0.98-1.0).

For the metric 9, it is detected a better result in the long texts (0.97), evidencing that they have a lower content of capital letters than the medium (0.95) or short (0.92) texts.

Metric 10 shows a lower use of punctuation and symbols in the long text (0.97) and the medium and short text show a slightly lower value (0.96).

Context-independent metrics by Text Type

The Context-independent metrics results by Text Type for text and processed text are illustrated in Figure 13. The text type was assigned in section 4.1: news, news title, review, review title and twitter. Each metric is represented with a different color.

	news	news title	review	review title	twitter	Total
M1: Abbreviation metric (text)	1.00	0.99	1.00	1.00	1.00	**1.00**
M1: Abbreviation metric (processed text)	1.00	1.00	1.00	1.00	1.00	**1.00**
M2: Spelling mistake metric (text)	0.93	0.84	0.97	0.97	0.96	**0.93**
M2: Spelling mistake metric (processed text)	1.00	1.00	1.00	1.00	1.00	**1.00**
M3: Unknown word metric (text)	0.92	0.84	0.96	0.96	0.92	**0.92**
M3: Unknown word metric (processed text)	1.00	1.00	1.00	1.00	1.00	**1.00**
M4: Grammatical sentence metric (text)	0.23	0.36	0.19	0.08	0.27	**0.22**
M4: Grammatical sentence metric (processed text)	0.91	0.18	0.41	0.01	0.19	**0.34**
M5: Lexical diversity metric (text)	0.42	0.51	0.51	0.60	0.50	**0.51**
M5: Lexical diversity metric (processed text)	0.70	0.81	0.78	0.77	0.73	**0.76**
M6: Lexical density metric (text)	0.82	0.98	0.86	0.99	0.96	**0.93**
M6: Lexical density metric (processed text)	0.81	0.99	0.89	0.99	0.97	**0.93**
M7: Stop word metric (text)	0.66	0.79	0.62	0.82	0.67	**0.71**
M7: Stop word metric (processed text)	1.00	1.00	1.00	0.84	0.98	**0.96**
M8: Average sentence length metric (text)	0.21	0.99	0.58	1.00	0.88	**0.73**
M8: Average sentence length metric (processed text)	0.99	1.00	0.98	1.00	1.00	**0.99**
M9: Uppercased word metric (text)	0.98	0.94	0.95	0.91	0.95	**0.95**
M9: Uppercased word metric (processed text)	1.00	1.00	1.00	1.00	1.00	**1.00**
M10: Punctuation metric (text)	0.97	0.97	0.97	0.96	0.94	**0.96**
M10: Punctuation metric (processed text)	1.00	1.00	1.00	1.00	1.00	**1.00**

Figure 13. Context-independent metrics results by Text Type

It can be observed in metric 1 that abbreviations are found in new titles (0.99).

In metric 2, we can find the highest amount of spelling mistakes also in news titles (0.84), followed by news (0.93), and the rest of the classifications can be found with a higher result (0.96-0.97).

Metric 3 reveals that news titles also had the highest number of unknown words (0.84), followed by twitter and news with 0.92 and finally reviews and reviews title with 0.96.

In the previous metrics, no comments were made about the processed text because in all metrics the result was 1.

For metric 4, several results were found. First, within the text and processed text the lowest result of grammatical sentences was for review titles (0.08, 0. 01). Then, in three text types of the result of this metric was reduced after applying the data processing: news title (from 0.36 to 0.18), review (from 0.08 to 0.01) and twitter (from 0.27 to 0.19). And two types of processed text improved: news (from 0.23 to 0.91) and reviews (0.19 to 0.41).

At metric 5, we found an increase of lexical diversity between text and processed text of 0.25, showing in all text types of the same tendency to increase (between 0.17 and 0.3). News is the text type with the lowest lexical diversity (text: 0.42, processed text: 0.7), despite its increase of 0.28. News title is the processed text with the highest increase (0.3), as well as the highest result (0.81). The original text with the best lexical diversity was review title (0.60).

In metric 6, we can see that lexical density improves after data processing except for news where it decreases 0.1 (from 0.82 to 0.81) and for review title where it stays the same (0.99). News is the worst lexical density result (0.82, 0.81). The best results in processed data were for news title (0.99) and review title (0.99).

Metric 7 revealed an improvement when comparing the processed text versus the original text for all categories. The reviews title obtained the best result (0.82) in the original text, but the worst after data processing (0.84). The rest of the data types improved to 0.98 or 1. The biggest increase was in reviews (0.62 to 1).

It can be seen in metric 8, the worst result in the original texts is news (0.21), followed by reviews (0.58) and twitter (0.88). After data processing, all five data types have scores between 0.98 and 1.

Metric 9 showed a low number of capital letters denoting scores above 0.9 before data processing. The best result corresponded to news (0.98) and the worst to review titles (0.91).

In metric 10 twitter is the lowest result (0.94) due to its greater use of punctuation and symbols. The best result was 0.97 for news, news titles and reviews.

5.2 Context-independent Dimensions of TDQ

The Context-independent dimensions of TDQ results are illustrated (Figure 14), where each dimension is plotted with different colors. The first column indicates the name of the dimension with the text on which it has been performed, the second column displays a bar with the evaluation results and the third column reflects the differential between the processed text and the text.

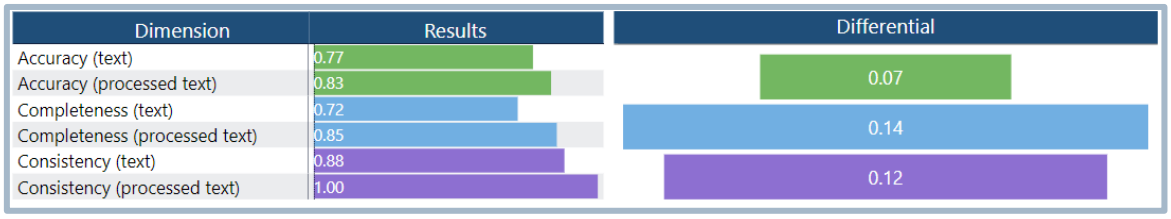


Figure 14. Context-independent dimensions results

Accuracy scores 0.07, Completeness 0.14 and Consistency 0.12. In the original text the best rated dimension was Consistency (0.88) and the worst was Completeness (0.72). After data processing, Consistency remained the best rated dimension with 1.0, but Accuracy became the lowest (0.83).

Context-independent dimensions by Length Classification

The results of the context-independent dimensions are given by length: long, medium, and short (Figure 15). Each dimension is highlighted in a distinct color.

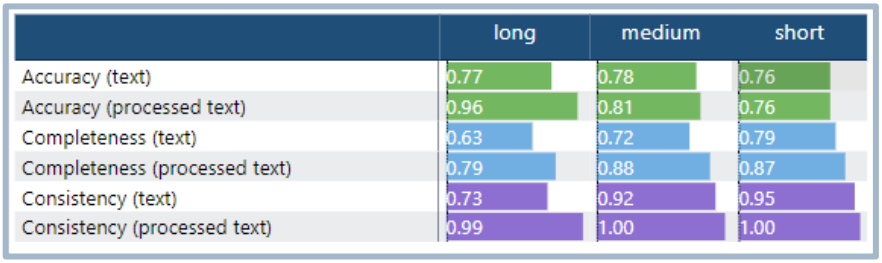


Figure 15. Context-independent dimensions results by Length Classification

Accuracy did not change for short texts before and after data processing representing the lower value in both texts, medium texts increased their value by 0.3 and long texts showed the biggest growth (0.19) from 0.77 to 0.96 becoming the best accuracy value in processed texts.

The completeness values increase if the original text is shorter. The long text obtained a value of 0.63, while the short text obtained 0.79. Once the text was processed, the three classifications were improved in various proportions, with the medium text obtaining the best completeness value (0.88).

Consistency results in the original text seem to follow the same trend as the completeness values, showing an increase if the text is shorter, where the lowest value was represented by long texts (0.73) and the highest by short texts (0.95). Meanwhile, in the processed texts, the short and medium texts represent the maximum value (1) and the long text improved considerably, reaching 0.99.

Context-independent dimensions by Text Type

The scores for the Context-independent dimensions are indicated by text type: news, news tittle, review, review title and twitter (Figure 16). The dimensions are emphasized with a distinct color.

	news	news title	review	review title	twitter
Accuracy (text)	0.77	0.76	0.78	0.75	0.79
Accuracy (processed text)	0.98	0.80	0.85	0.75	0.80
Completeness (text)	0.63	0.76	0.66	0.81	0.71
Completeness (processed text)	0.79	0.92	0.84	0.87	0.86
Consistency (text)	0.72	0.97	0.84	0.96	0.92
Consistency (processed text)	0.99	1.00	0.99	1.00	1.00

Figure 16. Context-independent dimensions results by Text Type

Accuracy lowest value for text and processed text is review title with 0.75 in both scores. The best accuracy was obtained in original text by twitter (0.79), followed by reviews (0.78) and news (0.77), and for processed text, news is the best result (0.98), followed by reviews (0.85),

then news title and twitter scored with 0.80. All text types show a growth after data processing, except for review title.

The results on the completeness of the original text give the best score to review title (0.81), slightly ahead of news title (0.76) and twitter (0.71). In the processed texts, the best scores are for the news title (0.92), closely followed by the review title (0.87) and twitter (0.86). An increase is also evident after data processing.

Consistency in the original text showed high scores for news title (0.97), review title (0.96) and twitter (0.92). The lowest value is for news (0.72). On the processed text, the results for news and reviews reached 0.99 and the rest achieved the maximum score (1). This may be related to the fact that news and reviews tend to be longer.

5.3 Context-independent Measurement of TDQ

The Context-independent measurement of TDQ results for Text and Processed Text are shown in Figure 17. Where we can appreciate the result obtained from both texts, and the differential between Processed Text and Text.

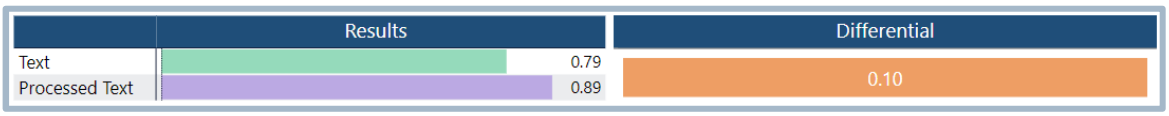


Figure 17. Context-independent Measurement results

The Context-independent measurement of Text is 0.79, and the Context-independent measurement of Processed Text is 0.89, therefore the difference of 0.1 between them can be seen.

Context-independent measurement of TDQ by Length Classification

Figure 18 presents the Context-independent measurement of both texts by length classification in a bar chart, where the text is presented in mint green and the processed text in lilac. Below the table includes the differential between processed text and text.

	long	medium	short
Differential	0.20	0.08	0.04

Figure 18. Context-independent Measurement results by Length Classification

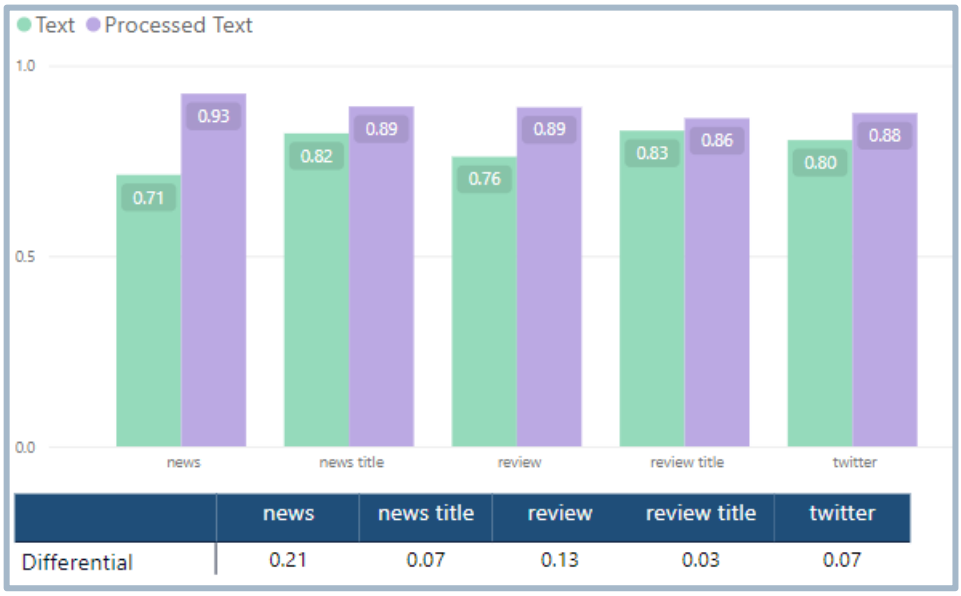
The results of Context-independent text measurement analyzed by length classification indicate the highest DQ for short texts (0.83), followed by medium texts (0.80) and finally long texts (0.72).

This is totally opposite to the results of the processed text, where short text presents the lowest value (0.86), followed closely by medium text (0.89), and where long text presents the best quality (0.92).

In the differential, the improvement between short and medium text is doubled, from 0.04 to 0.08. Meanwhile, in the long text, the differential of the short text increases 5 times over the short text (from 0.04 to 0.20). The increment represents the effect provided by the processing of the data.

Context-independent measurement of TDQ by Text Type

The results of the Context-independent measurements by text type are reported in figure 19, where we find the same color coding for text and processed text. In addition, a table with the differential between processed text and text is included.



	news	news title	review	review title	twitter
Differential	0.21	0.07	0.13	0.03	0.07

Figure 19. Context-independent Measurement results by Text Type

The results of the Context-independent text measurement analyzed by text type indicate that the lowest text DQ was in news (0.71), followed by reviews (0.76), review twitter (0.80), news title (0.82) and with the highest quality review title (0.83).

Once the text is processed the results shift significantly, news has the highest DQ value (0.93), next news title and reviews (0.89 both), then twitter (0.88), and review title is the text type with the lowest DQ (0.86).

Looking at the differentials reveals that news improved its DQ by 0.21, followed by review with 0.13, then news title and twitter with 0.07 and finally review title with a small change of 0.03.

5.4 Sentiment Analysis Performance and Context-independent Measurement of TDQ

The Sentiment Analysis Performance and its comparison with the Context-independent measurement is shown in Figure 20, where we see the performance is represented by the bars and DQ are the diamonds. On the right side of the figure, we can also find the most relevant resulting values, including the performance and DQ differentials.

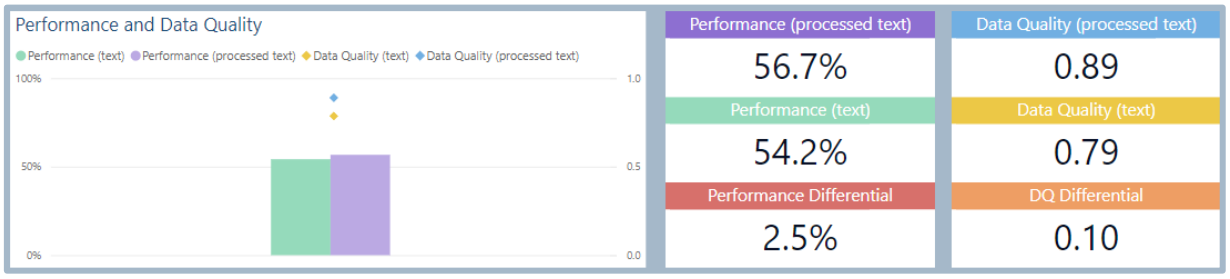
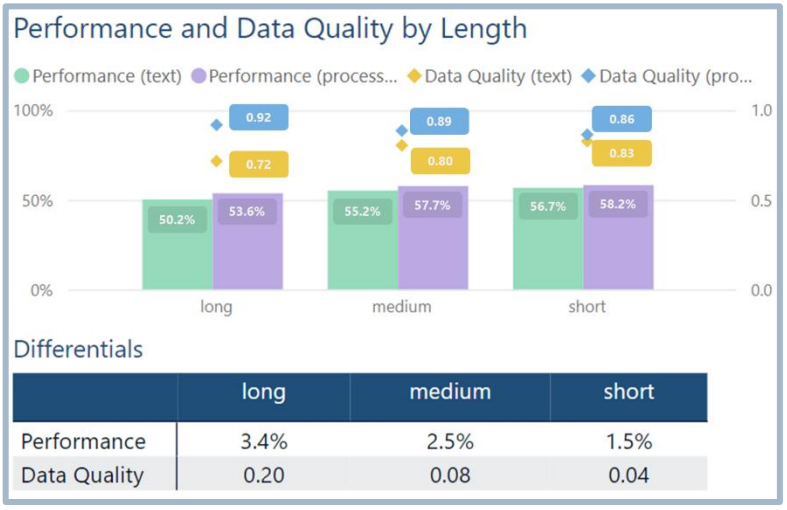


Figure 20. Sentiment Analysis Performance and Context-independent Measurement of TDQ results

The results suggest that both Sentiment Analysis Performance and Context-independent Measurement of TDQ were improved in the processed data. The Sentiment Analysis Performance increased by 2.5% from 54.2% to 56.7%, while Context-independent Measurement of TDQ increased from 0.79 to 0.89 with a 0.10 increase.

Sentiment Analysis Performance and Context-independent measurement of TDQ by Length Classification

The results of Sentiment Analysis Performance and Context-independent measurement of TDQ by length classification (long, medium, and short) are shown in Figure 21. The green bars represent the Sentiment Analysis Performance in text and the purple bars represent one in the processed text. The diamonds indicate Context-independent measurement of TDQ, yellow means from the text and light blue means from the processed text. The differential of each of them can be found in the table below the graph in the same figure.



	long	medium	short
Performance	3.4%	2.5%	1.5%
Data Quality	0.20	0.08	0.04

Figure 21. Sentiment Analysis Performance and Context-independent Measurement of TDQ results by Length Classification

Sentiment Analysis Performance is highest in both texts when the texts are short (text: 56.7%, processed text: 58.2%), followed by medium texts (text: 55.2%, processed text: 57.7%) and finally long texts (text: 50.2%, processed text: 53.6%). One factor that is notable is that the differential between these texts was greatest in the long texts (3.4%), followed by the medium texts (2.5%) and finally the short texts (1.5%).

Context-independent Measurement of TDQ by length classification was explained detailed in Section 5.3, where the same trend is clear as that shown in the differential of Sentiment Analysis Performance.

Sentiment Analysis Performance and Context-independent measurement of TDQ by Text Type

Figure 22 displays the Sentiment Analysis Performance and Context-independent measurement of TDQ results by text type (news, news title, review, review title and twitter). The color coding is the same as that used in Figure 21 and it can be read in the same figure. Below the graph it is possible to find a matrix with the Sentiment Analysis Performance and Context-independent measurement of TDQ differentials by text type.

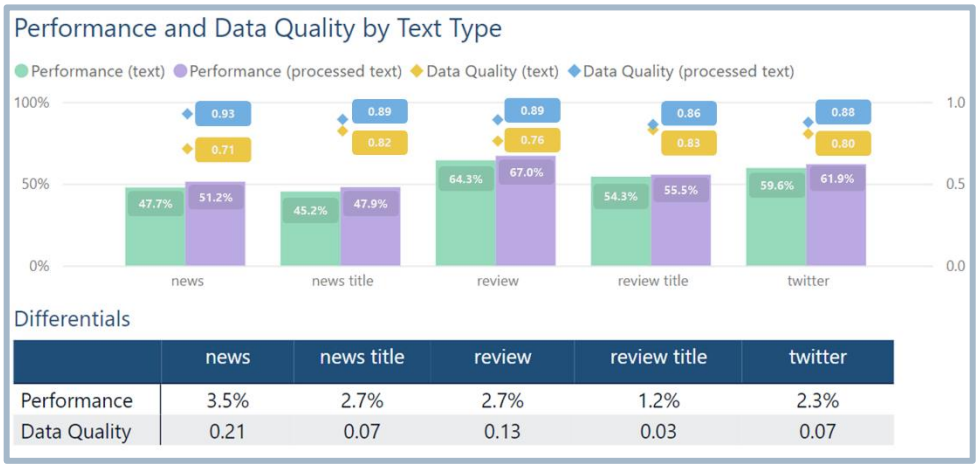


Figure 22. Sentiment Analysis Performance and Context-independent Measurement of TDQ results by Text Type

The best Sentiment Analysis Performance in text was found in reviews (64.3), followed by twitter (59.6%), review title (54.3%), news (47.7%) and news title (45.2%). After processing

the data, we still found the same order in the text type, but Sentiment Analysis differential varies, the highest differential was found in news with 3.5%, followed by news title and review with 2.7%, closely shadowed by twitter with 2.3%, and the lowest differential was found in review title with 1.2%.

The context-independent measurement of TDQ by text type was discussed in detail in Section 5.3. When comparing Sentiment Analysis Performance versus Context-independent Measurement of TDQ, we can notice that for example we see a Context-independent Measurement of TDQ differential of 0.07 in news title and twitter, but their changes in Sentiment Analysis Performance differential are not exactly proportional, in this case we see that the Sentiment Analysis Performance in news title improved 2.7% and 2.3% in twitter. Another case is that Sentiment Analysis Performance differential of news title and review are equal (2.7%), and the Context-independent Measurement of TDQ differential in the review is almost double (0.13 compared to 0.7).

After showing the results of our implementation, the next section will share the conclusions reached in this project.

6. Conclusion

In this section, the findings of this thesis are discussed, including an overall summary, the analysis of the research questions, the limitations identified throughout the project and their possible future directions.

6.1 Summary

The goal of this master thesis is to measure and evaluate Context-Independent Measurement of TDQ including a study case where we implement the proposed measurement and as result, analyze the effects of TDQ improvement. The implementation resulted in the conclusions found in Table 13.

Section	Conclusions
5.1	Metric 5 (Lexical diversity), Metric 7 (Stop word metric) and Metric 8 (Average sentence length metric) reflect the largest increase in individual metric evaluation. These metrics are related to the reduction of text length, which is usually reduced after data processing. When analyzing the metrics by length, Metric 4 (Grammatical sentence metric) is the only metric that decreases its quality in medium and short texts but caused a positive effect in long texts. Looking at the results by text type, it is possible to observe the same decreasing effect for News title, review title and twitter, which tend to be short or medium texts, and a considerable increase in news and reviews, which tend to be long texts.
5.2	Accuracy remained stable for short texts but increased for medium and long texts. Completeness and consistency improved across all length classifications after processing, with the most significant growth observed in long texts.

5.3	The Context-independent measurement for the Text improved 0.1 after data processing. Suggesting that data processing has a positive impact on the overall TDQ. In the original text, short texts show the highest TDQ and long texts the lowest. However, after data processing, the pattern changes, with long texts showing the highest TDQ and short texts the lowest TDQ. While these differences represent an overall improvement, they highlight the substantial improvement in data processing, especially evident in the long texts.
5.4	Both Sentiment Analysis Performance and Context-independent Measurement of TDQ showed improvements in processed data. When comparing Sentiment Analysis Performance with Context-independent Measurement of TDQ, variations in the differentials highlight non-proportional changes across text types.

Table 13. Conclusion Table by Section

In conclusion, the detailed analysis of metrics, length classification and text types emphasize the impact of data processing on TDQ. The improvements observed in the metrics and dimensions across different texts demonstrate the robustness of the processing techniques employed. The changing patterns in TDQ after processing, especially the reversal of trends between short and long texts, highlight the dynamic nature of data improvements. These findings provide valuable information for refining data processing strategies adapted to specific textual characteristics (for example, length classification), contributing to the optimization of TDQ. Given that both the performance of the sentiment analysis and the context-independent measurement of TDQ show positive trends after data processing, the study encourages further

exploration of the interaction between metrics, dimensions, and context-independent measurement of TDQ.

6.2 Research Questions

This study explores the critical issues surrounding the Context-independent metrics of TDQ to measure and evaluate the context-independent measurement of TDQ. The general research question aims to propose a methodology for assessing context-independent quality in textual data. For this, two specific research sub-questions (RSQ1 and RSQ2) explore the existing landscape of Context-independent metrics of TDQ metrics in the literature and explore optimal requirements for metric results and their implications. Moreover, a third sub-question (RSQ3) examines the available approaches to quantifying and seamlessly integrating context-independent quality metrics.

Responding to RSQ1: " *Which context-independent DQ metrics exist in the literature for text data?*" we found some metrics that can be consulted in Section 2, then these metrics were transformed based on the response of RSQ2: " *What is the best way to normalize metric results and what are the implications?*", where through the Literature Review we discovered that it is not only important to normalize the metrics, as this only covers one of five requirements for quality metrics, the proposal tries to cover at least four of the five proposed requirements: Existence of Minimum and Maximum Metric Values, Interval-Scaled Metric Values, Quality of the Configuration Parameters and the Determination of the Metric Values and Sound Aggregation of the Metric Values (Henrich et al. 2017). This last-mentioned requirement is of great contribution to RSQ3: "*What are the approaches for integrating context-independent quality metrics into a measurement?*". In addition to this, it was decided to integrate the metrics into dimensions and a single measurement.

Therefore, the conclusion to answer the research question "*How can we measure context-independent quality in text data?*" is that Context-independent Measurement of TDQ was measured through 10 metrics (Abbreviation, Spelling mistake, Unknow words, Grammatical sentence, Lexical diversity, Lexical density, Stop word, Average sentence length, Uppercased words and Punctuation) that can be integrated into dimensions (Accuracy, Completeness and Consistency) or a single measurement. These metrics tried to meet at least 4 of the 5 requirements for DQ metrics (Henrich et al. 2017).

6.3 Limitation

Although the research goal has been reached measuring and evaluating Context-Independent Measurement of TDQ in a study case, one limitation found in Metric 8: Average sentence length metric, where we can find negative values applying the formula if we do not validate that the residual is greater than 100, in this case the validation was applied in the code. This will happen if the sentence number is more than 1 and it exceeds a maximum length represented by Equation 16, where $\llcorner\lrcorner$ represent a rounding down.

$$Maximum\ Length\ Limit\ per\ sentence = \left\llcorner \frac{Number\ of\ Sentences * 100}{1 - \frac{1}{Sentences\ of\ Sentences}} \right\lrcorner$$

Equation 16. Maximum Length Limit per Sentence.

Figure 23 shows average length values for 1 to 25 sentences using the maximum length. The maximum average length limit will always be higher than 100, and the higher the number of sentences, the closer it is to 100. The result of the metric for using this maximum value will be 0 or a number close to 0.

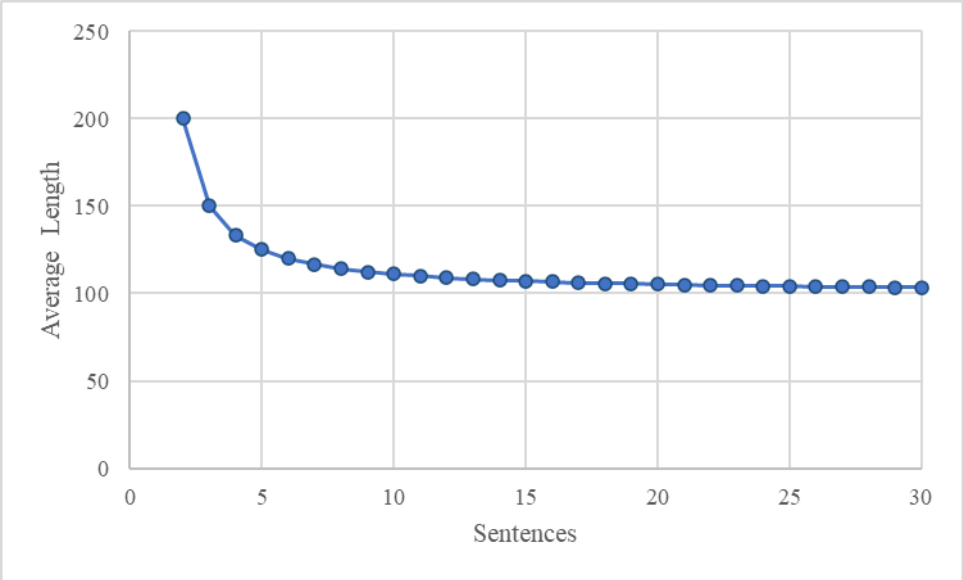


Figure 23. Average Length per Sentence using max length value

This leads us to consider that this metric could be improved, due to time constraints it was not continued with a better solution.

6.4 Future Directions

In future directions of the Context-independent measurement of TDQ, I would recommend initially validating the proposed metrics, and creating a more accurate measurement by doing an integration based on how important the metrics are and not just taking the overall average of the metrics, or weighted average of the dimensions. I would also suggest refining Metric 8, so that it can be more representative in long texts and in a single sentence as well. And finally apply this measurement on a larger dataset and compare it with other Machine Learning models.

7. References

[1] Agarwal, N., **WaniMudasir**, & Bours, P. (2020). Lex-Pos Feature-Based Grammar Error Detection System for the English Language. *Electronics*, *9*. https://doi.org/10.3390/electronics9101686

[2] Azeroual, O. (2019). A Text and Data Analytics Approach to Enrich the Quality of Unstructured Research Information. *Computer and Information Science*, *12*, 84. https://doi.org/10.5539/cis.v12n4p84

[3] Azeroual, O., Saake, G., Abuosba, M., & Schöpfel, J. (n.d.). Text data mining and data quality management for research information systems in the context of open data and open science.

[4] Azeroual, O., Saake, G., & Wastl, J. (2018). Data measurement in research information systems: Metrics for the evaluation of data quality. *Scientometrics*, *115*. https://doi.org/10.1007/s11192-018-2735-5

[5] Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*. https://doi.org/10.5334/dsj-2015-002

[6] Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), Article 3. https://doi.org/10.17239/jowr-2020.11.03.01

[7] Das, B., Majumder, M., & Phadikar, S. (2018). A Novel System for Generating Simple Sentences from Complex and Compound Sentences. *International Journal of Modern Education and Computer Science*, *1*, 57–64. https://doi.org/10.5815/ijmecs.2018.01.06

[8] Ehrlinger, L., & Wöß, W. (2022). A Survey of Data Quality Measurement and Monitoring Tools. *Frontiers in Big Data*, *5*. https://www.frontiersin.org/articles/10.3389/fdata.2022.850611

[9] Ge, M., & Helfert, M. (2007). A Review of Information Quality Research—Develop a Research Agenda. In *Proceedings of the 2007 International Conference on Information Quality, ICIQ 2007* (p. 91).

[10] Gudivada, V. N., Apon, A., & Ding, J. (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations.

[11] Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. (2017). Requirements for Data Quality Metrics. *Journal of Data and Information Quality*, *9*(2), 1–32. https://doi.org/10.1145/3148238

[12] Juddoo, S. (2015). *Overview of data quality challenges in the context of Big Data*. https://doi.org/10.1109/CCCS.2015.7374131

[13] Kiefer, C. (2016). Assessing the Quality of Unstructured Data: An Initial Overview. https://ceur-ws.org/Vol-1670/paper-25.pdf

[14] Kiefer, C. (2019). *Quality Indicators for Text Data*. Gesellschaft für Informatik, Bonn. https://dl.gi.de/handle/20.500.12116/21801

[15] Makhoul, N. (2022). *Review of data quality indicators and metrics, and suggestions for indicators and metrics for structural health monitoring*. *3*. https://doi.org/10.1186/s43251-022-00068-9

[16] Maydanchik, A. (2007). *Data Quality Assessment*. Technics Publications.

[17] Mylavarapu, S. S. G. S. (2020). *Context-aware quality assessment of structured and unstructured data*. https://shareok.org/handle/11244/328620

[18] Nastachowski, B. (n.d.). *Academic Guides: Grammar: Sentence Structure and Types of Sentences*. Retrieved November 13, 2023, from https://academicguides.waldenu.edu/writingcenter/grammar/sentencestructure

[19] Nesca, M., Katz, A., Leung, C., & Lix, L. (2022). A scoping review of preprocessing methods for unstructured text data to assess data quality. *International Journal of Population Data Science*, *7*, 1757. https://doi.org/10.23889/ijpds.v6i1.1757

[20] *NLTK :: Natural Language Toolkit*. (n.d.). Retrieved November13, 2023, from https://www.nltk.org/

[21] Oliveira, P., Rodrigues, F., & Rangel Henriques, P. (2005). A Formal Definition of Data Quality Problems. In *Proceedings of the 2005 International Conference on Information Quality, ICIQ 2005*.

[22] Sebastian-Coleman, L. (2012). Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework. Newnes.

[22] Serra, F., Peralta, V., Marotta, A., & Marcel, P. (2023). *Context-Aware Data Quality Management Methodology* (pp. 245–255). https://doi.org/10.1007/978-3-031-42941-5_22

[23] Sidi, F., Hassany Shariat Panahy, P., Affendey, L., A. Jabar, M., Ibrahim, H., & Mustapha, A. (2013). *Data quality: A survey of data quality dimensions*. https://doi.org/10.1109/InfRKM.2012.6204995

[24] Taleb, I., Dssouli, R., & Serhani, M. (2015). *Big Data Pre-Processing: A Quality Framework*. https://doi.org/10.1109/BigDataCongress.2015.35

[25] Taleb, I., El Kassabi, H., Serhani, M., Dssouli, R., & Bouhaddioui, C. (2016). *Big Data Quality: A Quality Dimensions Evaluation*. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122

[26] Taleb, I., Serhani, M., & Dssouli, R. (2018). *Big Data Quality Assessment Model for Unstructured Data*. https://doi.org/10.1109/INNOVATIONS.2018.8605945

[27] *Universal POS tags*. (n.d.). Retrieved November 13, 2023, from https://universaldependencies.org/u/pos/

[28] Vazquezsoler, S., & Yankelevich, D. (2001). Quality Mining: A Data Mining Based Method for Data Quality Evaluation. (p. 172).

[29] Vikholm, O. (n.d.). Dealing with unstructured data.

[30] Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

[31] Wang, Y., Kon, H., & Madnick, S. (2003). Data quality requirements analysis and modeling. *Massachusetts Institute of Technology (MIT), Sloan School of Management, Working Papers*.

8. Appendices

Appendix A. Universal POS Tags

POS tag	Meaning	Examples
ADJ	adjective	Big, old, green, African, incomprehensible, first, second, third
ADP	Ad position	In, to, during
ADV	Adverb	Very, well, exactly, tomorrow, up, down Interrogative/relative adverbs: where, when, how, why, whenever, wherever (including when used to mark a clause that is circumstantial, not interrogative, or relative) Demonstrative adverbs: here, there, now, then Indefinite adverbs: somewhere, sometime, anywhere, anytime Totality adverbs: everywhere, always Negative adverbs: nowhere, never
AUX	Auxiliary	Tense auxiliaries: has *(done)*, is *(doing)*, will *(do)* Passive auxiliaries: was *(done)*, got *(done)* Modal auxiliaries: should *(do)*, must *(do)* Verbal copulas: He is a teacher. Agreement auxiliaries: [quc] la (2nd person singular formal), alaq (2nd person plural formal)
CCONJ	Coordinating conjunction	And, or, but
DET	Determiner	Articles: a, an, the Demonstrative determiners: this as in *I saw* this *car yesterday.* Interrogative determiners: which as in *"*Which *car do you like?"* Relative determiners: which as in *"I wonder* which *car you like."* Quantity determiners (quantifiers): indefinite any, universal: all, and negative no as in *"We have* no *cars available."*
INTJ	Interjection	Psst, ouch, bravo, hello
NOUN	Noun	Girl, tree, etc., beauty, decision
NUM	Numeral	0, 1, 2, 3, 4, 5, 2014, 1000000, 3.14159265359 11/11/1918, 11:00 one, two, three, seventy-seven k (abbreviation for thousand), m (abbreviation for million), etc. I, II, III, IV, V, MMXIV
PART	Particle	Possessive marker: 's Negation particle: not

PRON	Pronoun	Personal pronouns: I, you, he, she, it, we, they Reflexive pronouns: myself, yourself, himself, herself, itself, ourselves, yourselves, themselves Interrogative pronouns: who, what as in What *do you think?* Relative pronouns: who, that, which as in *a cat* who *eats fish*; who, what as in *I wonder* what *you think*. Indefinite pronouns: somebody, something, anybody, anything Total pronouns: everybody, everything Negative pronouns: nobody, nothing Possessive pronouns (which usually stand alone as a nominal): mine, yours, his, hers, its, ours, theirs Attributive possessive pronouns (in some languages; others use DET for similar words): my, your
PROPN	Proper noun	Mary, John, London, NATO, HBO
PUNCT	Punctuation	Period: . Comma: , Parentheses: ()
SCONJ	Subordinating conjunction	that as in *I believe* that *he will come.* If, while
SYM	Symbol	$, %, §, © +, −, ×, ÷, =, <, > :), ♥‿♥, 😝 john.doe@universal.org, http://universaldependencies.org/, 1-800-COMPANY
VERB	Verb	run, eat runs, ate running, eating
X	Other	xfgh pdl jklw