CONTEXT-AWARE QUALITY ASSESSMENT OF STRUCTURED AND UNSTRUCTURED DATA

By

SESHA SAI GOUTAM SARMA MYLAVARAPU

Bachelor of Technology in Computer Science and Engineering
Jawaharlal Nehru Technological University
Hyderabad, India
2012

Master of Science in Computer Science
Oklahoma State University
Stillwater, Oklahoma
2015

Submitted to the Faculty of the
Graduate College of
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
JULY, 2020

CONTEXT-AWARE QUALITY ASSESSMENT OF STRUCTURED AND UNSTRUCTURED DATA

Dissertation Approved:

Dr. Johnson Thomas

Dissertation Adviser

Dr. K. M. George

Dr. Christopher Crick

Dr. Weihua Sheng

Name: SESHA SAI GOUTAM SARMA MYLAVARAPU

Date of Degree: JULY, 2020

Title of Study: CONTEXT-AWARE QUALITY ASSESSMENT OF STRUCTURED AND UNSTRUCTURED DATA

Major Field: COMPUTER SCIENCE

Abstract: Data analysis is a crucial process in the field of data science that extracts useful information from any form of data. The ease of access and maintenance makes structured data the most popular choice among many organizations even today. On the other hand, with the rapid growth of technology, more and more unstructured data, such as text and image, are being produced in large amounts. Apart from the techniques used, the quality of the data plays a prominent role in the accurate analysis. Data quality becomes inferior to poor maintenance and mediocre data generation strategies employed by amateur users. This problem escalates with the advent of big data. Data cleaning is one possible solution to this problem. However, it requires a great deal of domain knowledge and expert inference to verify and repair the data. Data Quality Assessment (DQA) is an effective alternative that differentiates between good and bad quality data. Although DQA requires domain knowledge, since it does not repair or change the inherent data, it is more viable to automate the process. In this dissertation, we propose two quality assessment models for structured data and textual form of unstructured data. The context of data plays an important role in determining the quality of the data. Therefore, we automate the process of context extraction in structured data using machine learning techniques. For textual data, we use natural language processing to identify data errors and assess quality. However, an accurate source of information is necessary to identify data errors. Therefore, we propose an automated mechanism to identify the closest dataset using deep neural networks with minimal user intervention. In addition, we also look into multiple dimensions of data quality such as completeness, accuracy, and consistency, to create a comprehensive quality assessment model. Our experimental results show the importance of the data context and multiple dimensions in quality assessment.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

CHAPTER I

INTRODUCTION

1.1 Motivation

Data is available in various formats. Data analysis will yield new information. This makes data a valuable resource. Data science is a collection of fundamental principles that promote extracting information and knowledge from data [106]. Data analysis is an important branch of data science that helps to understand data. In recent times, data analysis has become an integral part of business irrespective of domain. Businesses can make crucial decisions for growth and future investment based on the information extracted from data. MicroStrategy Enterprise Analytics reported that 90% of organizations worldwide state that data and analytics are important for their digital transformation initiatives [78]. Therefore, data analysis has evolved as an important research topic.

Data analysis and its complexity vary according to the type of data. The complexity of analysis is associated with several aspects such as data resources, the accuracy of analysis, and domain dependence. Structured data has a pre-defined format, making it a less complicated type to analyze. On the other hand, unstructured data has no pre-defined format which therefore makes it more complex to analyze. In the pre-internet era, the majority of data were generated by machines and transactions in industry. With the rapid growth of the internet, data formats took diverse forms such as social media, online transactions, etc. According to Gartner Inc. [1], 80% of the data produced over the next five years will be unstructured. They consist of images, videos, documents, and other types of rich media that consumers and businesses are producing every second.

International Data Corporation (IDC) predicts that the total amount of digital data worldwide will grow from 33 Zettabytes (ZB) in 2018 to 175 ZB by 2025 [36]. This irrepressible data generation brings numerous challenges including data management, intricate analysis, privacy and security issues. For this reason, more and more companies are adopting big data technologies to handle their issues. Furthermore, with the advent of the Internet of Things (IoT), an unprecedented amount of real-time data is being produced by individuals, sensors, and enterprises. IDC also predicts that, by 2025, real-time data will comprise more than a quarter of all data created, with 95% of that data generated by IoT [36]. This massive amount of data, if utilized properly, can open doors for advances in science, technology, business and our personal lives.

Data analysis provides reliable results if the volume and variety of data are high. Hence data analysis provides the most benefits when it is applied to big data. However, data analysis can also produce undesirable outcomes, resulting in substantial losses for companies [64]. Although there are many reasons for inaccurate analysis, quality of data is one of the most important factors. Bad quality data makes analysis dubious and leads to error-prone conclusions which results in unreliable decision-making thereby producing adverse effects on business. The problem is compounded with data decay which is a very common and important problem where the quality of data diminishes over time. Hence there is a need to examine the data to assess the quality of data.

1.2 Problem Statement

Inferior quality data is becoming a major concern to organizations as they have a negative impact on the company's growth. According to a survey conducted by Gartner Inc., organizations estimate that poor-quality data is costing them on average $14.2 million annually [47]. Raw data can never be used to perform any analysis, as it is susceptible to inconsistent formats, missing data, human errors, and false interpretations. Data provenance is directly proportional to the quality and integrity of the data. Although many data-driven companies have ample resources to handle data, they are still prone to inaccurate auto-generation such as sensor data. Companies rely more on consumer-generated data as they are more influential and highly impact business decisions [101]. However, data generated by individuals contribute more to the existing data quality problem as they can have a diverse range of formats and inputs.

Data can be inherently defective or the quality of the data may depreciate in the process of transformation performed for a specific use. There are multiple reasons for the existence of erroneous and poor-quality data. A few common reasons include:

- excessive manual entry of data;
- multiple variants of the same data or data duplication;
- loss of data fragments due to migration from one format to the other;
- data obfuscation to protect sensitive information;
- integration of data from multiple sources;
- inaccurate data generated by individuals in social media.

Traditional data management tools are designed to handle structured data. With the rapid advancement of technology, semi-structured and unstructured data is occupying a major share in today's data generation. Therefore, we need new methods to deal with these distinct formats of data. As a result, big data technologies have evolved to manage large datasets and perform resource-intensive data analysis. Despite having these efficient techniques to govern big data and execute various operations, the most important element is the quality of data. As mentioned earlier, the high volume of data affects the analysis positively, but it is only beneficial if the data is of good quality. Thus, regulating the quality of data takes center stage to achieve worthwhile analysis. Moreover, with the advent of the Internet of Things, the increased growth of real-time data leads to the need to instantly resolve quality issues.

From the perspective of data quality, two methods to improve analysis are data cleaning and quality assessment. Data cleaning (also called data cleansing) refers to the process of identifying and correcting errors within the data to get more value from the data [114]. On the other hand, Data Quality Assessment (DQA) is a scientific and statistical evaluation of data to determine if the quality of data is suitable to perform a specific operation [40]. Both approaches require domain expert knowledge and need to be tailored for each specific domain. However, data cleaning is much more domain-dependent than quality assessment, as it requires the creation of new values and the replacement of existing values.

1.3 Data Quality: Key Aspects

With ever-growing data in multiple domains, data has become the biggest asset to businesses. Data helps enterprises to make critical decisions such as workforce management, risk minimization, cost reduction, investment decisions etc. Decision-makers in companies can maximize their profits with the proper use of data. This makes DQA an essential component to incorporate in companies. This section reviews the key aspects which are needed during quality assessment.

Similar to data analysis, the assessment of data quality also changes based on the type of data. Since structured data has a more definite and precise format, the assessment rules and requirements differ when compared to unstructured data. Besides, quality issues also change according to the type of data. For example, missing values are common in structured data, but not very prevalent in unstructured data. Therefore, the data format is a fundamental classification that needs to be considered for DQA.

Although data quality is a single entity, it comprises multiple dimensions. Quality cannot and should not be determined by a single measure. Different dimensions serve distinct purposes and are intended to measure multiple aspects of data quality [84][110]. The most commonly used quality dimensions are completeness, validity, accuracy, and consistency. Completeness and validity verify the structure of the data, whereas accuracy and consistency deal with the actual value. Even though each dimension is measured individually, all the dimensions are interlinked to form a single entity called quality. These quality dimensions are discussed in detail in Chapter II.

The quality of data varies depending on its usage. Since a single dataset can serve various purposes, the information within the data is utilized in multiple ways. As a result, the data which is appropriate for a specific purpose, may not meet the needs of a different objective. This scenario is generally considered as the context of the data. Accordingly, the quality of data is measured differently for the different context of the data. This is the most essential aspect of data quality.

1.4 Plausible Illustrations

Example 1:

Table 1.1 shows a hypothetical example of medical record data of patients suffering from hypertension.

First Name	Last Name	Age	Systolic Blood Pressure	Diastolic Blood Pressure
John	Smith	54	118	80
Kelly		49	132	84
Veola	See	61	142	945

Table 1.1: Example of Structured Data

There are two quality issues in this medical record. The ‘Last Name’ column of the second patient is missing a value. This issue refers to the completeness dimension of data quality. The absence of information drastically reduces the quality of the dataset. Even though the name is not a unique field, a person’s identification can be determined to some extent using his/her full name. However, among some ethnicities, there are many common first names. As a result, the last name carries more importance in a person’s identification. In this example, the quality of data is low as we cannot be certain about the identity of the patient.

The other quality issue in this example is validity. The diastolic blood pressure of the patient named Veola See is unusually high. Although the value follows the required data type i.e. an integer, the value ‘945’ is not valid as a measure for diastolic blood pressure.

These problems can be detected and possibly resolved by quality assessment and cleaning. We can identify the completeness and validity issues using some pre-defined rules. However, not all issues can be resolved using these techniques. For example, the name of the patient can be recorded if there is another instance of the same patient. On the other hand, the blood pressure value of a patient is a measurable unit and cannot be replaced or corrected with a new accurate value.

Example 2:

Consider the following example biography of a fictitious tennis player.

"John Smith was born and raised in Los Angeles, CE. He is a 38-year old passionate tennis player representing the state of California. Besides tennis, Mr. Smit also likes to play baseball."

The above textual data has no grammatical errors and clearly describes a person. However, there are two quality issues in this passage. The last name of the person is different in the first and second lines. This makes the data inconsistent. Although the last name 'Smith' in the first line is more probable and natural, we cannot be sure which name is correct, as a name can be anything. In addition to consistency, this text also carries an accuracy issue. No state in the U.S. is represented by a code 'CE'. Such accuracy issues are detected and possibly corrected using a source of correct information.

1.5 Contributions

Data quality is a multi-dimensional measure that changes according to the context. Since DQA is an extensive task, there are multiple phases involved in developing such a model. Our research focuses on context extraction and quality evaluation of both structured and unstructured data. In this work unstructured data is also referred to as "textual data", whereas structured data is called as "structured." We show the importance of context in quality assessment and how multiple dimensions contribute to the overall quality of data. We propose to automate this process to minimize domain dependency and expedite the quality assessment process. We achieve this objective by employing machine learning methodologies. In this thesis we look at DQA models for both structured data and unstructured data in the form of text. Our contributions in building a comprehensive DQA model are as follows.

- Context Extraction in Structured Data

  Without examining the context, quality estimation is imprecise. In this dissertation, a novel context extraction framework has been proposed for structured big data. Since context is heavily domain-dependent, we apply a machine learning approach to extract the context with negligible domain dependence.

- Identification of Closest Dataset

  Some quality dimensions such as accuracy and consistency, require an accurate source of information to compare with. Following the context extraction, we propose a mechanism to automatically link the records and identify a dataset to acquire an accurate closest dataset to compare with. We also propose a similar model for textual data.

- Quality Assessment of Structured Data

  Since we aim to measure quality in multiple dimensions, our main contribution is to design a mechanism to estimate quality in the accuracy and consistency dimensions. These dimensions require a closest dataset identified in the above step. We accomplish this goal by implementing deep learning and statistical techniques. This module considers the extracted context and utilizes the closest dataset obtained from our previous works.

- Context Extraction in Textual Data

  The interpretation and the way the data is organized creates the fundamental difference between structured and unstructured data. As a result, the models devised for structured data will not suffice for unstructured data. In this dissertation, we propose a context extraction model for textual data.

- Accuracy Assessment of Textual Data

  Unlike structured data, accuracy assessment in textual data cannot be performed by comparing the words in two texts. The information within the textual data can be expressed in multiple ways. Therefore, we need to assess the accuracy of the information in textual data. We carry out this task by performing sentiment analysis on text data by identifying the subjects and objects in each sentence.

- Consistency Evaluation of Textual Data

Similar to the accuracy assessment, we employ natural language processing to extract the dependency rules between the multiple elements in each sentence to assess the consistency of textual data.

To summarize, DQA in our proposed approach is composed of context extraction followed by accuracy and consistency assessment for both structured and textual (unstructured) data. Context extraction yields the relevance of the data to the problem at hand and accuracy and consistency assessment measures the quality of the context relevant data.

1.6 Dissertation Organization

The rest of this dissertation is organized as follows. Chapter II reviews the background and literature in this area. In Chapter III, the framework to extract the context of structured data is presented. In Chapter IV, the mechanism to identify the closest dataset is described in detail. Chapter V explains the methodologies to assess various dimensions of structured data quality. The framework to extract the context from textual data is explained in Chapter VI. In Chapter VII, the framework for the assessment of textual data quality is proposed. Finally, Chapter VIII concludes this dissertation and provides directions for future work.

CHAPTER II

BACKGROUND AND LITERATURE REVIEW

Data Quality Assessment (DQA) is a multi-step process that identifies the real essence of data. The value of data is key in the proper utilization of data, which can be used for several purposes. To validate the quality assessment of data, the evaluation process must include some crucial aspects such as quality dimensions, data context, etc. The fundamental steps involved in DQA are error identification, context evaluation, and quality estimation. Numerous approaches have been implemented in the literature to accomplish these tasks. However, since each of these operations is performed individually for various purposes, not all approaches are suitable for quality assessment. This chapter discusses the earlier work conducted in these areas.

2.1 Types of Data Errors

Understanding the source of data is certainly essential in estimating quality. DQA changes with the change in the type of error residing in the data. A good DQA model must be capable of identifying the error type and consider it while estimating the quality. However, an error can be defined in numerous ways depending on how and where it exists. This section reviews the literature about the types of data errors and their classification.

Many studies classified data errors from different perspectives such as violation of rules, inducing errors, etc. The different viewpoints in error classification are due to the difference in motives and approaches to solve the same data quality problem. Therefore, it is not feasible to have an extensive classification of data errors that satisfies all conditions.

2.1.1 Violations

The most common errors in data are defined based on whether or not the data satisfies certain rules within a domain. Ilyas et. al. [54] and Chu et. al. [29] consider data errors as violations of patterns and rules. Although there can be many variants in data error classification, a basic taxonomy of error types is defined in [7]. Figure 2.1 shows the error type taxonomy proposed in [7].

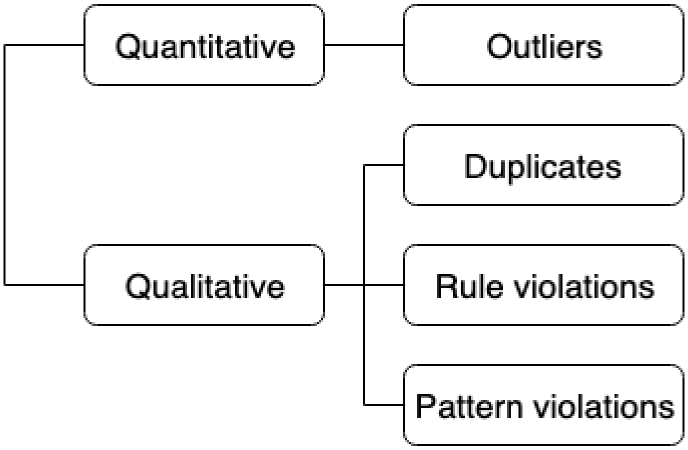

Figure 2.1: Error Type Taxonomy [7]

For the classification of errors shown in the figure, the authors defined an error as a deviation from ground truth value. The errors are mainly divided into quantitative and qualitative errors, meaning measurable and non-measurable errors respectively. The errors are further classified as Outliers, Duplicates, Rule violations, and Pattern violations.

- Outliers are measurable errors which include data values that differ from a range of possible values for a particular attribute.

- The existence of redundant copies of the same record reduces the quality of data.

These multiple references to the same data are called duplicates.

- Rule violation errors refer to the data that does not follow certain pre-defined rules that add meaning to the data. These rules usually define the nature of the attribute or data item such as uniqueness, null values, etc.,
- Data items that do not satisfy the syntactic and semantic constraints are categorized as pattern violation errors.

2.1.2 Data Entry Errors

Another prevalent reason for low-quality data is due to the introduction of errors into data, either deliberately or accidentally. There are many ways for errors to creep into the data [52][62]. The impact of each error on data quality varies depending on how the errors are introduced into the data. Thus, a classification of data entry errors is necessary to understand the cause and difference between them. A classification of data quality problems with different levels was proposed by Rahm et. al. in [86]. It also consists of a classification for data entry errors, which is a part of the data quality problem (see figure 2.2).

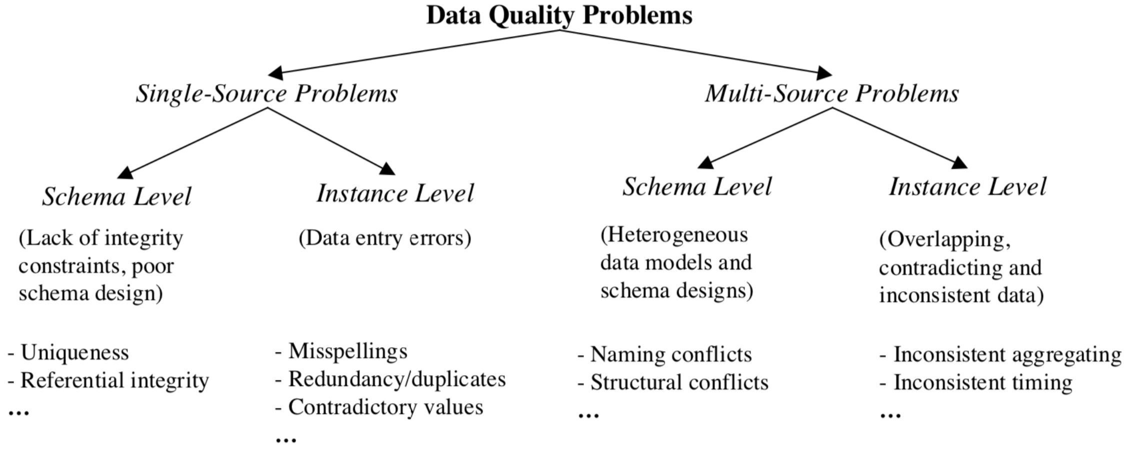

Figure 2.2: Classification of data errors in data sources [86]

Data quality and its problems change based on the source of the data. Rahm et. al.'s classification (figure 2.2) also presents the analysis of data problems based on the number of sources that generated the data. Rule violations are the primary cause for data errors that happens at the schema level, irrespective of the number of sources. However, data entry errors are highly observed at the instance level. Data duplication, data obfuscation, spelling mistakes, and inconsistent values are a few examples of data entry errors. These problems are aggravated further when data is integrated from multiple sources. However, multi-source data have more problems such as inconsistent aggregation, inconsistent timing, and missing data due to migration.

In structured data, rule violations can be avoided by enforcing strict guidelines during data generation. On the other hand, data entry errors are difficult to control as it involves distinct human interventions, especially in multi-source data. However, in semi-structured data, even rule violations are hard to overcome, since defining a smaller number of rules is the primary difference between structured and semi-structured data.

2.1.3 Errors in Textual Data

The absence of a definite arrangement of information makes unstructured data a viable choice for individuals to adopt. However, in terms of data quality, unstructured data is disparate and highly complex to assess. Moreover, since most of the unstructured data providers are not experts, the quality of data further diminishes. There are numerous formats in unstructured data such as video, audio, text, etc., This dissertation focuses on the textual form of unstructured data.

The errors that reside in textual data are dissimilar to the errors in structured data. However, the nature of errors in text data is similar to that of structured data. Many studies in the literature focused on identifying and categorizing the error types in text data [55][100]. Table 2.1 shows the classification of common data error types in text data. Similar to structured data, errors are mainly categorized into two types, namely: rule violations and data entry errors. Nonetheless, unlike the constraints in structured data, the rules enforced in textual data are corresponding language grammar rules such as punctuation, determiner, etc., for English. The introduction of common errors either intentionally or unintentionally are misspelled words, unwanted words, etc. This insight of the error types and classification will be worthwhile in the efficient identification of errors in the data.

Error Category	**Error type**
Rule Violations	Punctuation errors
	Determiner errors
	Prepositional errors
	Subject-Verb agreement
Manual or Automated Data Entry Errors	Misspelled words
	Missing words
	Unwanted words
	Duplicate words

Table 2.1: Classification of Errors in Textual Data

2.2 Data Quality Dimensions

Data quality is commonly described as the condition of the information residing in the data. Although quality also considers the structural aspects of data, the information has more impact on the overall quality. There are many definitions of data quality, but data is mainly considered as high quality if it satisfies certain quality criteria or if it fits for its intended purpose [87][41]. The quality criteria (termed as dimensions) usually differ from one domain to the other. However, there exist some definite dimensions that are common in almost every domain. This section mainly examines the most commonly used quality dimensions.

2.2.1 Survey of Quality Dimensions

Data Quality dimensions represent the views, criteria, or measurement attributes for data quality problems that can be assessed, interpreted, and possibly improved individually [91]. Data quality comprise of numerous dimensions, each contributing to the overall quality of data. A method to construct data quality dimensions was proposed by Wang et. al. in [109]. The authors performed two surveys to extract the data quality attributes and dimensions perceived by data consumers. The first survey extracted 179 quality attributes performed using factor analysis. All the attributes are further categorized into 20 quality dimensions in the second survey. Table 2.2 shows the data quality dimensions and a description for each dimension.

Quality Dimension	**Description**
Access Security	Access to data can be restricted
Accessibility	Data must be retrievable and easily available
Accuracy	Data must be reliable and error-free
Appropriate	Amount of Data The amount of data must be appropriate
Believability	Data must be believable or credible
Completeness	The scope of the information contained in the data
Concise	Data must be compact and well-organized
Consistency	Data are represented in a consistent format and compatible with previous data
Cost-effectiveness	Cost of data accuracy and data collection
Ease of Operation	Data must be easily customized and integrated
Ease of Understanding	Data must be clear and easily understood
Flexibility	Data must be adaptable and flexible
Interpretability	Data must be useful to extract information
Objectivity	Data must be unbiased
Relevancy	Data must be applicable or relevant to the task
Reputation	Source of the data is trusted
Timeliness	Age of data
Traceability	Data is well-documented and verifiable
Value-added	Data add value to the task
Variety	Variety of data and its sources

Table 2.2: Data Quality Dimensions [109]

2.2.2 Common Data Quality Dimensions

Since data quality is a necessary aspect, quality dimensions must be considered. The total number and dimensions considered could be different. However, certain quality dimensions are popular and necessary to evaluate the overall quality of data. Depending on the domain and usage, these dimensions can have distinct definitions. In this dissertation, we focus on four important data quality dimensions namely Completeness, Validity, Accuracy, and Consistency.

Completeness

Data should not include extra information, nor should it be missing relevant data values. The lack of completeness represents the data entry error type of data errors. We present the common definitions of data completeness extracted from [24][46][59] respectively.

Definitions:

Data completeness is defined as a state where users have access to all data they deem important to the information-based service in which they are involved.

Completeness is defined as the degree to which data collection provides the values or all attributes of entities that are supposed to have values.

Completeness is the extent to which data are not missing and are of sufficient breadth and depth for the task at hand.

Validity

Data is invalid if it does not meet the required constraints of data. It belongs to the rule violation error type of data errors.

Definitions:

Validity refers to the proportion of data with given attributes that truly has the required characteristics. [23]

Property of validity means that data is valid if it guarantees to satisfy the constraints set on the data. [38]

Accuracy

Data is considered accurate if the data values are true and close to real-world data. It belongs to the data entry error type of data errors.

Definitions:

Data accuracy refers to whether the data values stored for an object are the correct values. To be correct, the data value must be the right value and must be represented in a consistent and unambiguous form [82].

Accuracy is defined as the closeness between a value v and a value v', considered as a correct representation of the real-life phenomenon that v aims to represent [92].

Addressing the accuracy dimension is straightforward. If the recorded value is not what it should be, the data unit is labeled as defective [16].

Consistency

Consistency is maintaining the identical data values at multiple instances. It represents the data entry error type of data errors.

Definitions:

The consistency dimension can be viewed from many perspectives, one being the consistency of the same (redundant) data values across tables [84].

Consistency is defined as the degree to which data managed in a system satisfies specified integrity constraints or business rules [91].

2.2.3 Relationships of Quality Dimensions

Although each quality dimension has its purpose and contribution in defining the overall quality of data, there exist inter-relationships between the quality dimensions. The relationships between the dimensions can vary depending upon how and where the data is used. At times, one quality dimension can either positively or negatively impact another dimension. Sometimes, the same two dimensions may not have any relation with each other. In some domains, some quality dimensions overlap with other dimensions, resulting in a single measurement for multiple such dimensions. A description of the importance of data quality and its dimensions are discussed in [13]. Table 2.3 shows the common quality dimensions and their related dimensions to access the overall data quality.

Data Quality Dimension	**Related Dimensions**
Completeness	Validity and Accuracy
Validity	Accuracy, Completeness, Consistency and Uniqueness
Accuracy	Validity
Consistency	Validity, Accuracy and Uniqueness

Table 2.3: Related Data Quality Dimensions

Although there exist multiple related dimensions for each quality dimension, the degree of relatedness varies from dimension to the other. One important observation from table 2.3 is that validity and accuracy are quite essential dimensions that are commonly related to every other quality dimension. Validity refers to the correct representation of data value, whereas, accuracy refers to the true real-world value. Therefore, if the data is corrupt syntactically and semantically, the measurement of other quality dimensions becomes meaningless.

2.3 Data Quality Assessment Methods

There are many ways to assess the quality of data. Since data quality is a multi-dimensional concept, each approach has its limitations in terms of assessing all the dimensions in data quality. Therefore, all the dimensions of data quality cannot be measured using one single technique. Moreover, the primary objective and problem at hand vary with every dimension. The functions of the most common data quality dimensions are listed below.

- Comparison of data values
- Identifying missing values
- Analysis of data context
- Discovering data constraints

Although DQA is extensively studied in the literature, many approaches have been proposed to stabilize with data-related problems and improvise the process. This section examines the overview of various approaches used in the literature to measure the quality of data. A more detailed review of related work for corresponding quality dimensions are discussed in further chapters.

2.3.1 Statistical Approaches

A comparison of data values appears to be a simple task, but it gets complicated as the format of the data changes. Depending on the type of data, many comparison methods such as exact matching, similarity identification, and distance calculation are used. Exact matching is a simple approach that verifies whether two data items have an identical value or not [30]. If the comparison is between two tuples with numerical values or strings represented as numerical values, a mathematical method called cosine similarity can be used as shown in [49]. Cosine similarity measures the cosine angle between two vectors projected in a multi-dimensional space. An alternate string-matching technique based on the distance between two characters or words was proposed in [31].

Identifying missing values is a critical aspect of DQA. Null values refer to the missing data values, whether or not they are necessary. Nevertheless, the quality of the data only depends on the necessary data values. Therefore, identification of necessary missing values is the primary focus in DQA. Statistical methods based on the linearity and non-linearity of missing values was implemented in [73][96].

Context defines the importance and usage scenario of the data. This is very essential in the evaluation of the quality of data, as the quality changes with the need and use of the data. A statistical approach called TF-IDF (Term Frequency-Inverse Document Frequency) is used in the literature to extract the information from textual data [53]. TF-IDF calculates the product of the frequency of every word and its inverse document frequency to identify its importance in a document. Many variants of TF-IDF are also proposed in the literature to extract the important topic of the document [25] [79], which is further considered as the context.

Verification of data constraints is necessary to assess the validity and consistency dimensions of data quality. Since the constraints vary widely with multiple data domains, a domain expert knowledge is necessary to attain them and examine the quality of data. A constraint-based quality assessment and data repairing methods have been proposed in [42][21].

2.3.2 Machine Learning

With the rapid growth of data in every industry, the importance of data quality increases exponentially to make use of the valuable information inherent in the data. Human interaction and domain knowledge dependency are the biggest barriers to implement efficient statistical methods to evaluate the quality of data on a larger scale. This makes machine learning a worthwhile choice to perform a quality assessment.

Although few simple operations in DQA such as missing value identification, basic data value comparison, etc., do not require machine learning, complex procedures such as context analysis, constraint extraction, require machine learning to perform DQA on high-dimensional big data. However, identifying missing values and data comparison becomes complicated in unstructured data and therefore requires an automated mechanism. Richman et. al. proposed a machine learning mechanism to identify and predict missing data values using Support Vector Machines (SVM) and Artificial Neural Network (ANN) [89].

A comparison of data values in textual data is more complex compared to structured data. Albeit structured data also contains strings or words, there exists a relation between attributes. However, there is no relationship between individual data items of different attributes. In other words, there are only syntactic constraints within structured data. Textual data, on the other hand, contains both syntactic rules (grammar) and semantic relationships between words in a sentence and between sentences in a document. Therefore, techniques such as exact matching, distance calculation, etc., used for structured data are not adequate for unstructured textual data. As a result, data comparison in textual data is widely studied in the literature. Automated mechanisms for record linkage in structured data were proposed by the authors in [28][111][107]. String and sentence matching mechanisms based on semantic rules were proposed in [71][9].

Automation of more intricate operations such as context extraction and data constraint discovery is a challenging task even in structured data. This problem was addressed in [63]. However, the application of Natural Language Processing (NLP) is necessary to analyze and extract the information residing in textual data. Many NLP techniques such as sentiment analysis [22] and topic modeling [108] are necessary to extract information and determine the quality of data. A comprehensive analysis of automation mechanisms for data quality assessment is discussed in the following chapters.

2.4 Big Data Technologies

In addition to the problems associated with the quality of data, the accelerated growth of data demands the design of more advanced technologies to handle large amounts of data. Aforementioned, the analysis of unstructured data is a tedious task even with the use of machine learning techniques. Inevitably, unstructured data covers a larger proportion of today's big data. In this section, we discuss the most prominent big data frameworks used in this dissertation.

2.4.1 Apache Hadoop

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models [2]. Hadoop comprises two key components namely Hadoop Distributed File System (HDFS) and MapReduce.

Hadoop Distributed File System (HDFS)

A typical file in HDFS is gigabytes to terabytes in size [3]. The architecture of a conventional HDFS cluster is shown in Figure 2.3. A large dataset is divided into a pre-defined size of smaller components called data blocks. The segregation of data into blocks provide diverse advantages such as efficient use of disk storage, block-level abstraction, reliability, and fault-tolerance. However, once the administrator sets the block size of the dataset, a user cannot access the individual blocks of data, rather the dataset can be accessed as a whole.

The ability to access the data from a different node during a node failure makes Hadoop a fault-tolerant system. As shown in figure 2.3, a Hadoop cluster usually consists of a single name node and multiple data nodes. A master-slave relationship exists between the name node and the data nodes. The name node contains only the metadata of the dataset, whereas, the entire dataset, which is divided into smaller blocks are stored only in the data nodes. The name node always receives a heartbeat message from the data nodes to ensure that they are active.

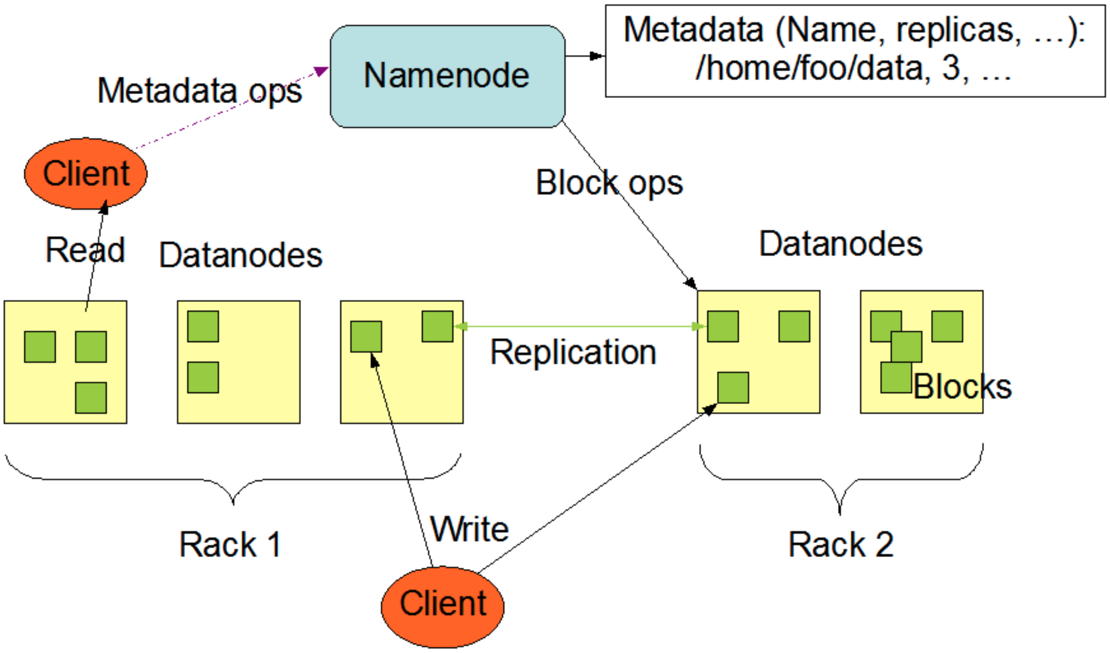

Figure 2.3: HDFS Architecture [2]

MapReduce

MapReduce is a programming model in HDFS where the jobs are logically divided into a batch of subtasks. MapReduce consists of four components namely: HDFS, client, job tracker, and task tracker. All the resources required by a MapReduce job are stored in the HDFS. The client the user who submit jobs to the MapReduce model. The job tracker is responsible for keeping track of all the jobs submitted to the Hadoop cluster. The batch of subtasks divided in the MapReduce model is defined as a map and reduce tasks, which are monitored by the task tracker. The input and output format in MapReduce are key-value pairs. The parallelism in a MapReduce model is performed using the map tasks to manage big datasets.

2.4.2 Apache Spark

Similar to the MapReduce model, Apache Spark is a distributed cluster computing framework used to divide and process the data in parallel [67]. The data in Spark is distributed using a read-only architecture called the Resilient Distributed Dataset (RDD). Apache Spark processes the data in-memory, in contrast to Hadoop, which transfers the data in and out of the disk frequently. Unlike Hadoop, Spark includes powerful libraries to support intensive tasks such as real-time streaming, machine learning, etc. Figure 2.4 shows the modules in Apache Spark. In addition to the libraries, Spark also supports integration with Hadoop [60]. Data processing in Spark can be standalone, or it can utilize the efficient processing of HDFS.

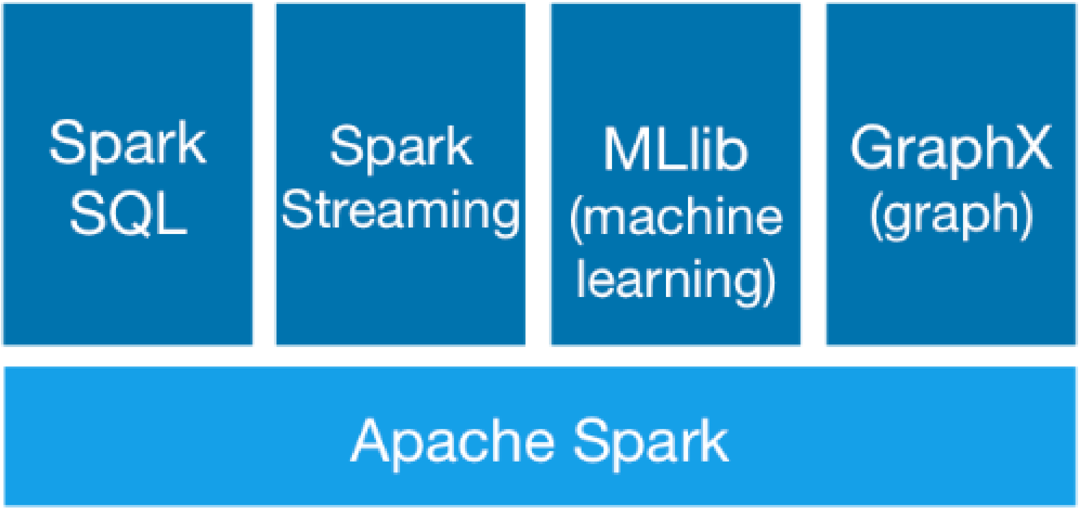

Figure 2.4: Apache Spark Modules [67]

Spark SQL

Spark SQL module supports the structured data processing in Spark [4]. Besides the regular processing of data, this module allows the users to perform standard SQL operations and queries on the data residing in Spark. It also supports the more optimized form of relational databases called DataFrames.

Spark Streaming

Spark Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams [5]. The stream processing in Spark is supported by discretized stream or Dstream. Dstream is a continuous stream of data represented by a sequence of RDDs [5]. The streaming data in Spark can be obtained from many sources, which then can be processed by other modules.

MLlib

MLlib in Spark is a machine learning library that consists of the most commonly used machine learning algorithms. The algorithms in MLlib support some essential machine learning operations such as classification, regression, recommendation, topic modeling, and clustering. Although there exist various machine learning libraries for many programming languages, the main objective of Spark's MLlib is to make machine learning scalable.

2.5 Summary

In this chapter, we discussed the different types of data errors and their classification. Although the complexity of errors differs based on the nature of data, the classification of data errors is similar for both structured and unstructured data. Data errors are categorized as rule violations and data entry errors.

We examined a survey of numerous data quality dimensions consisting of 179 quality attributes. These attributes are further condensed into 20 quality dimensions. We discussed the most commonly used data quality dimensions, namely: completeness, validity, accuracy, and consistency. These quality dimensions not only contribute to the overall quality of data, but they are also interconnected to each other.

Depending on the quality dimension and the type of the data, the procedure to identify the data error changes. The objectives to evaluate the quality of data include a comparison of data values, identification of missing values, analysis of context, and the discovery of data constraints. We presented a broad literature review of both statistical and machine learning approaches to identify data errors and assess the quality of data.

We explained the need for and importance of using big data technologies in DQA. We presented an overview of the two big data frameworks used in our research. The architecture and working of Apache Hadoop are explained. We also discussed the importance of Apache Spark, along with the essential libraries available in Spark.

In most cases, only one quality dimension cannot determine the complete quality of the data. Although few studies in the literature focused on DQA in multiple dimensions, they are heavily domain-dependent. Moreover, existing techniques do not consider the context of the data in the assessment of data quality. This thesis presents a context-aware comprehensive data quality assessment with minimal domain dependence.

CHAPTER III

CONTEXT EXTRACTION IN STRUCTURED DATA (CES)

3.1 Introduction

As discussed earlier, context plays a prominent role in data quality assessment. There are numerous definitions of context. In the field of data science, the most commonly used meaning of context is the situation or the purpose of data usage[12]. Context-aware data quality assessment is defined as determining the quality of data based on the situation it is used. In other words, checking whether the data meets the quality standards for the purpose of the data being used, which is frequently termed as *"fitness for use"*.

Although context has a coherent definition, the meaning is not apparent in relation to structured data. In order to associate a context with structured data, we can utilize the fact that the importance of variables or features of structured data changes according to the context [6]. Therefore, we consider the most important subset of features in structured data as the context of the data. However, obtaining the context of structured data from a domain expert can become a challenging task, depending on the size of data. With the growth of big data, high-dimensional datasets have become the norm. This situation demands extended support from a domain expert and hence escalates the processing time. The delay in quality assessment can negatively impact the quality of data [17]. To minimize the processing time and to marginalize domain dependency, an automated context extraction mechanism is essential. This chapter focuses on developing an automated Context Extraction module for Structured data (CES).

3.2 Related Work

The quality of the data is determined based on where and how the data is used. In other words, the context of the data plays a major role in determining the quality of the data. Although this seems apparent, in effect, the context of the data is not readily obtainable. Manually defining context is an exhausting job as the same data can be used in various contexts resulting in endless human intervention. Therefore, the context of the data is usually ignored in the quality assessment of the data. However, the reliability of data analysis becomes inferior. This problem is largely discussed in the literature [95][18]. A classification of numerous problems associated with contextual and non-contextual data is presented in [112].

A few studies in the literature focused on contextual data quality assessment. Nevertheless, context is not automatically extracted from the data, instead, a domain expert provides the context during quality assessment. Malaki et. al. proposed a framework for multi-dimensional contexts of data quality assessment [75]. The authors introduced contextual hierarchies as components of contexts for data quality assessment.

A quality assessment model for in-use big data was proposed by Jorge et. al. in [77]. This model supports the three characteristics of data quality namely: contextual adequacy, operational adequacy, and temporal adequacy. Although this model supports data from any domain, the contextual assessment is heavily domain-dependent. A similar domain-dependent framework for contextual information quality assessment was proposed by Stvilia et. al. in [103]. Moreover, these models do not consider the necessary dimensions of data quality during the assessment.

Apart from comprehensive DQA models that consider the context of data, a few models were proposed which consider data from specific domains. The quality assessment models in [85][88] are devised only for specific contexts in the healthcare domain. Most of the models in the literature have multiple problems including heavy domain dependency for contextual DQA, limited DQA with a fewer number of quality dimensions, and deficient quality assessment of big data. To the best of our knowledge, our context extraction module is the first to support automation with negligible domain expert support.

3.3 Feature Selection

"In machine learning and statistics, feature selection, also known as variable selection, is the process of selecting a subset of relevant features for use in model construction." [56]

Extracting the subset of features from a dataset gives the important features for a given situation. In other words, feature selection provides the context of the dataset. However, not all datasets that need to be assessed for quality are context-dependent. Context independent datasets are those that have no definite purpose to serve apart from a general analysis. Context is neither necessary nor does it play a crucial role in the quality assessment of such datasets. On the other hand, context-dependent datasets have a clear reason for existence. Therefore, the analysis varies depending on the change of context or the features in a dataset.

Accordingly, there are many ways feature selection can be done depending on whether the dataset is context-dependent or not.

Feature selection methods are mainly divided into three types [51]

Filter methods:

Filter methods calculate a statistical measure to weigh the importance of each feature by assigning a score. In most of the cases, the interdependency between any two features is ignored by just considering the importance of each feature. Some examples of filter methods are correlation coefficient scores, chi-squared tests, etc.

Wrapper methods:

Like filter methods, wrapper methods also rank the features based on their importance. But instead of using a statistical measure, wrapper methods use a predictive model to score the features. Wrapper methods use different search strategies to find the possible subsets of features that will predict the output. Based on the accuracy of the prediction model, the appropriate feature subset is selected. The selected feature subset is directly considered as the important feature set or furthermore used to assign scores to individual features.

Embedded methods:

As the name suggests, embedded methods are a combination of feature selection and predictive models. The primary purpose of these methods is a prediction. It is similar to wrapper methods with the only exception being that the feature selection is a part of the model construction process for prediction. In other words, the predictive model itself has a built-in feature selection mechanism to choose the best features to produce better accuracy. Some examples include LASSO, ridge regression, etc.

Based on the type of data, the CES module chooses the appropriate feature selection method to extract the context of the data. To extract the context from a context-dependent dataset, the model used in this dissertation is a predictive model that requires a target or output variable. As opposed to the domain expert defining a context for a structured dataset, this context extraction module automatically defines the context when provided with a target variable. Since feature selection is an implicit process in embedded methods, there are comparatively fewer possibilities to choose from and there is a chance of data transformation in some embedded methods [11]. This is not necessarily a problem, but the primary purpose of using feature selection here is to extract the context and assess the quality of original data rather than prediction. So, the CES module ignores embedded methods to extract the context as embedded methods include prediction and may also change the actual data.

Whenever the context of the data is not required or the target variable is unknown for prediction, the filter methods are used for quality assessment. Though computationally expensive, the wrapper methods are the most efficient and comprehensive methods for context extraction where the optimal feature subset can be found with a choice of numerous strategies for both subset selection and prediction model while preserving the original data.

3.4 Framework

Figure 3.1 shows the framework of context extraction for structured data (CES). Since the CES module requires the context of the data, the first component of the framework is context extraction. We use machine learning algorithms to extract the context of the data. The extracted context is further used to measure the quality of data. If the data is not context-dependent, filter methods are used to remove the redundant or highly correlated variables and the resultant variables are used to measure the data quality. Whereas if the data is context-dependent, wrapper methods are applied to extract the context by sending the output to the subset selector to select the most significant variables in the data. Note that this is domain-dependent as the wrapper methods require a target or output variable to use the prediction models.

3.5 Context Extraction

As discussed in section 3.3, the CES module extracts the context of structured data using feature selection techniques. However, there are many ways feature selection can be performed, and each has its advantages and disadvantages, depending on the data. Since CES is intended to be used for any domain, this module consists of a mix of several feature

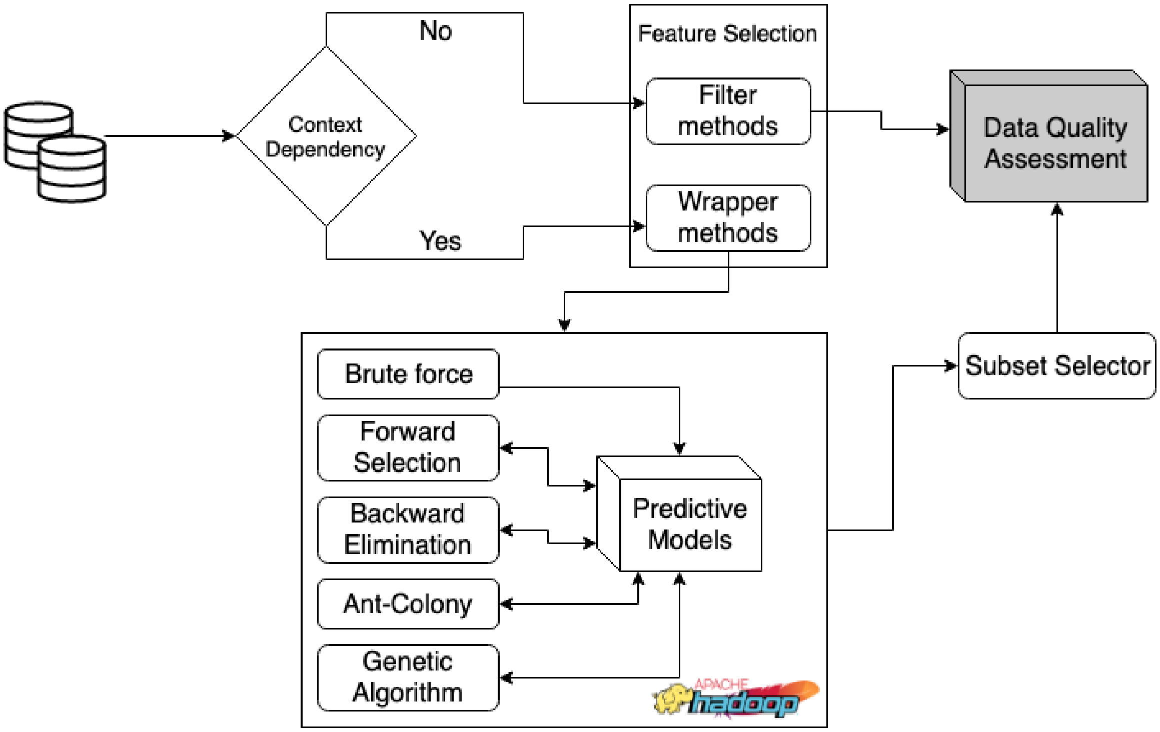

Figure 3.1: Context Extraction Framework for Structured Data

selection algorithms. As mentioned earlier, embedded methods of feature selection are ignored for context extraction, while filter and wrapper methods are used according to the context-dependency. The overview of context extraction process is shown in algorithm 3.1.

3.5.1 Context Dependency

This is the first phase of the CES module, where the process of context extraction is determined, i.e., whether or not the data is context-dependent, which is determined by the presence or absence of a target variable. If the dataset is context-independent or the target variable for prediction is unknown, the context extraction is determined by a filter-based feature selection mechanism. This is not a context per se, but it produces a reduced feature size. However, if the input dataset is context-dependent, CES uses the wrapper-based feature selection mechanism. Since wrapper methods use numerous ways to obtain a feature subset, this CES module uses a variety of search strategies described later to determine the

Algorithm 3.1 Context Extraction Algorithm for Structured Data

```
Determine whether the dataset is context dependent or independent based on the target variable
if dataset is context independent then
    Apply filter method
    Obtain feature subset
else
    Apply multiple wrapper methods
    Choose the (best) smallest subset from all wrapper methods
    Consider the best subset as the context of the dataset
end if
Deliver the context for Closest Dataset Search (CDS)
```

best feature subset.

3.5.2 Context Independent

As the target variable for prediction is not considered for the extraction of the feature subset, this is a straightforward process that carries the basic feature filtering mechanism like removal of redundant variables, highly correlated variables and null variables. The CES module uses Pearson's Chi-squared test to determine the importance of each feature in the dataset.

3.5.3 Context-Dependent

Wrapper methods require a target variable to use a prediction model and determine the best feature subset. The choice of prediction models, selection of feature subsets, and their combinations produce different results and indeed defines context differently. There is no single machine learning algorithm that best suits all kinds of data. With this notion, a set of search strategies and prediction models are implemented in the CES framework to choose the best feature subset for data quality assessment. The following two sections describe the search strategies and prediction models used in the CES module to generate possible feature subsets. These feature subsets are further assessed to obtain an optimal feature subset which finally defines the context of a particular dataset.

3.5.4 Subset generation: Search strategies

In this section, the approaches to generate multiple subsets of features to obtain the best subset are discussed. Different search strategies can be implemented in wrapper methods to generate the feature subsets ranging from a methodical search to a heuristic search. We used brute-force, forward selection, and backward elimination as a part of a heuristic search and ant-colony optimization and genetic algorithm as a part of stochastic subset selection. Each subset is tested for accuracy by using the subset for prediction. The subset that gives the best accurate prediction is the best set of features.

Brute-force:

Brute force subset selection is a straightforward method where it considers all possibilities of the subsets to generate the feature subset. If ‘n’ is the number of features of a dataset, 2n-1 gives the number of possible subsets of a dataset. Though this method might give the optimal subset, it is computationally expensive as the time complexity is linear. This method does not consult the prediction model continuously, but rather it consults after producing all the subsets and the prediction model chooses the best subset based on model accuracy.

Forward Selection:

Unlike brute-force, forward selection does not produce multiple feature subsets to choose from. It is an iterative process of feature selection where initially a null set is considered and each feature is added to the existing subset based on the selection criteria. The selection criteria are the predictive model accuracy, where the current feature is added to the feature subset only if the accuracy is improved compared to the previous subset and omitted otherwise.

Backward Elimination:

Backward elimination is like the forward selection but in reverse order. Each feature is selected for removal sequentially to check its importance in deriving the output. Initially, the complete set of features is considered as the subset, and if the current model accuracy without a particular feature is improved when compared to the existing accuracy, the removal of the feature from the existing subset is confirmed and is no longer added to the subset. Eventually, the feature is added back to the subset, if the model accuracy decreases as it indicates the importance of that particular feature in determining the output.

Ant-Colony Optimization:

Ant-colony optimization [113] is a probabilistic algorithm to find the optimal path using graphs based on the behavior of artificial ants. This algorithm is inspired by real ants seeking a path to the source of food in their colony. The shortest path is discovered using pheromone trails deposited by the ants moving in a random fashion initially. The probability of a path increases with an increase in pheromone level. Based on the probability, the ants follow the shortest path that leads to the destination. We used this optimization technique to derive the feature subset that gives the best accuracy with a prediction model. The probability is calculated using equation 3.1.

$$P_{i,j} = \frac{\left(\tau_{i,j}^{\alpha}\right)\left(\eta_{i,j}^{\beta}\right)}{\Sigma\left(\tau_{i,j}^{\alpha}\right)\left(\eta_{i,j}^{\beta}\right)} \tag{3.1}$$

where, τ = amount of pheromone on edge i,j

α and β are the parameters to control the influence $\big($set to one$\big)$

η = desirability of edge i,j $\left(\frac{1}{W_k}\right)$ with w being weight of feature obtained from Pearson's correlation matrix

Each feature is treated as an artificial ant, where each ant traverses through the existing set of features and calculates the prediction accuracy. Based on the accuracy, the pheromone matrix is updated, which defines the probabilities of each path. Equation 3.2 gives the formula for the pheromone update.

$$\tau_{i,j} = \big(1-\rho\big)\tau_{i,j} + \Delta\tau_{i,j} \tag{3.2}$$

where, τ = amount of pheromone on edge i,j

ρ = rate of pheromone evaporation $\big($set to 0.01$\big)$

Δ = amount of pheromone deposited, given by

$\frac{1}{W_k}$ if ant k travels on edge i,j

0, otherwise

Genetic Algorithm:

Genetic algorithms [44] are inspired by Darwin's theory of survival of the fittest and mimic the biological reproduction process. Initially, the individuals are randomly distributed into populations representing the chromosomes and the fitness of each individual is calculated based on the behavior of the population. Two individuals are selected based on their fitness value to serve as parents for crossover intending to produce better offsprings. In the feature selection problem, each feature represents a gene and the collection of genes is a chromosome (the subset of features). Each chromosome is represented by a string of 0s and 1s. Where 1 represents the presence and 0 represents the absence of the feature.

3.5.5 Prediction Models

As the selection of feature subsets is diversified between several algorithms and approaches, different algorithms are used to develop a better predictive model as the outcomes vary hugely depending on the combination of a predictive model and a feature subset. The machine learning algorithms used for prediction as a part of context extraction are Decision trees (DT) and Logistic Regression (LR). A better model is chosen for context extraction when all the wrapper methods (subset generation) are implemented with each of these machine learning algorithms.

3.6 Subset Selector

Each wrapper method produces different feature subsets with a combination of predictive models. It is important to choose the best feature subset as it derives the context of the data. Subset Selector module chooses the feature subset that best satisfies the factors of selection criteria. Prediction accuracy, feature subset size, and execution time are the important factors of subset selection. Each factor is given a weight to derive a uniform selection score. Subset selection is performed using equation 3.3. The subset with the highest score is considered as the best subset.

$$S_i = \frac{\big(0.5 * a_i\big) + \big(0.3 * n_i\big)}{0.2 * t_i} \tag{3.3}$$

where, S_i = Selection score for feature subset i

a_i = prediction accuracy of subset i

n_i = number of features in subset i

t_i = execution time (in seconds) of algorithm that produced subset i

3.7 MapReduce Environment

As the data grows, assessing the quality of data becomes a difficult task. On the other hand, context extraction requires more sophisticated implementation as the number of features increase. With the advent of big data, the number of features or variables of data is increasing rapidly. The average number of features ranges from a few 100s to even 1000s in current big datasets [93]. It is computationally an expensive task to perform machine learning on large datasets, especially if it involves an iterative process. In order to overcome this problem, the power of parallelism provided by Apache Hadoop's MapReduce framework [2] is used in the CES module.

3.8 Experimental Setup

Though brute-force, forward selection, and backward elimination algorithms are simple to implement in Hadoop, ant-colony and genetic algorithms require a different approach. For the ant-colony algorithm, the CES module splits the data and calculates the feature weights using Pearson's correlation coefficient in the mapper section. Model prediction and subset generation based on prediction accuracy including pheromone updates are implemented in the reducer. For the genetic algorithm, the mapper evaluates the fitness function and keeps track of the best individual. On the other hand, the reducer performs the selection and crossover steps of the algorithm. All the MapReduce jobs are implemented in 24-nodes Hadoop cluster. Python and Java are used to implement all the wrapper methods in Hadoop.

Dataset:

Medical Information Mart for Intensive Care (MIMIC-III) data [58] is used as a case study for the CES module. MIMIC-III data consists of de-identified health-related data associated with over 40,000 patients collected between 2001 and 2012 who stayed in critical care units. It is a large dataset with many details of patients including demographics, vital sign measurements, medications, caregiver notes, etc. It is a high-dimensional data with more than 750 features. The intent is to extract the context from the dataset using feature selection and assess the quality of the dataset.

3.9 Results and Discussion

The primary purpose of context-based data quality assessment is to observe how useful the data is, in a particular context. It is not appropriate to label the data as “bad quality” if it totally serves the purpose even though it is “visibly” bad. This indeed means that if a small subset of features defines the data, it is appropriate to consider only those features for quality assessment, provided it outperforms the bigger subset and defines the data better than any other subset. The target variable for context extraction in MIMIC-III dataset is the disease of each patient. Therefore, the context (best feature subset) is extracted based on the disease prediction. We first extract the feature subsets using our wrapper methods. Their performance or significance is then determined using prediction models. As discussed earlier, since there is no single prediction model that best suits for all datasets, we use two different prediction models (decision tress and logistic regression) to observe the context of the dataset. Therefore, the wrapper methods (including genetic algorithm) helps to produce the feature subsets, whereas, the prediction model determines the efficiencies of each feature subset. Figure 3.2 shows the sizes of the best subsets produced by each wrapper method for each prediction model.

It is observed that the genetic algorithm produces a smaller subset with MIMIC-III dataset, whereas backward elimination has the largest subset. However, the feature subset size cannot be the only criteria. The ultimate goal is to find the best subset that better defines the data, so this may be possible with other wrapper methods with slightly more features. Considering the prediction accuracy is another factor for subset selection. The respective accuracies for the above methods are shown in figure 3.3.

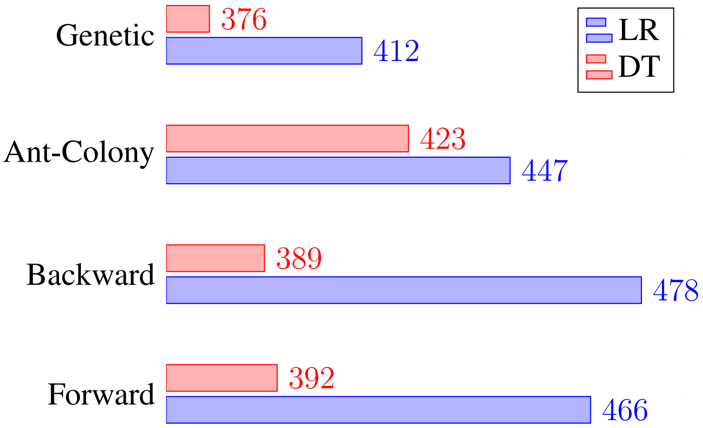

Figure 3.2: Feature subset size for each wrapper method

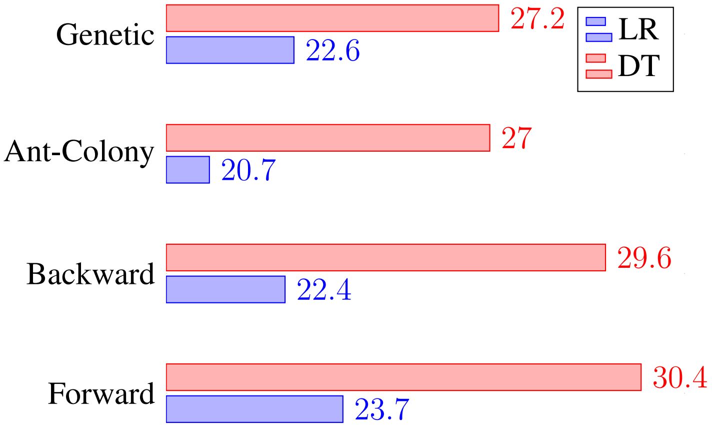

Figure 3.3: (%) Prediction accuracies for each wrapper method

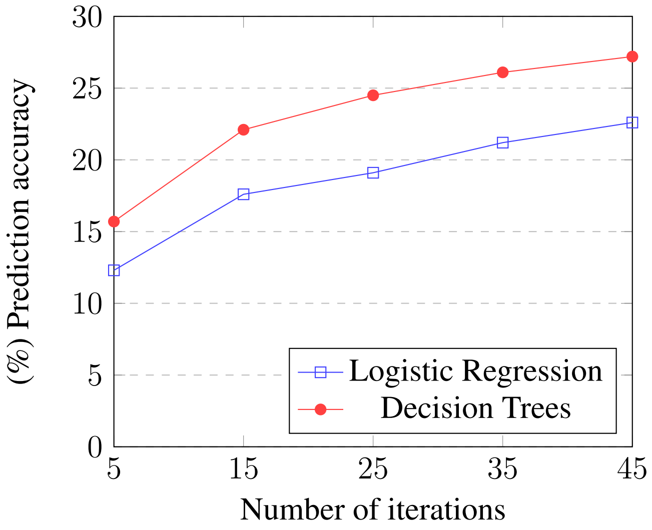

Figure 3.4: Genetic Algorithm: Prediction accuracy vs number of iterations

The subset selector component takes all these aspects into consideration and suggests a better subset that can optimally assess the quality of the dataset. Although some algorithms are fairly straightforward, the genetic algorithm apart from being resource-intensive is also highly arbitrary in design and the stopping criteria are difficult to meet. The aim is to observe its learning from iteration to iteration. Figure 3.4 gives the prediction accuracies of genetic algorithm per iterations.

The important factors that influence the best subset selection are observed in this section. From the results, it is understood that each wrapper method has a different strategy and outcome for the same data. Though some algorithms suggest a smaller feature subset, their prediction accuracy is relatively high. This is an important observation to extract the context of data. Prediction accuracy and feature subset size being the key factors, the subset selector considers these results to choose the best subset which is considered as the context of data.

The context of a dataset varies with the change in the purpose of its usage. In other words, if the target variable changes, the context of the dataset can vary depending on the variable. This makes context extraction very subjective not only between different datasets but within the same dataset for different target variables. Therefore, we intend to observe the results produced by our CES module for MIMIC-III dataset with disease prediction as the purpose or target variable.

Related Variables	**Unrelated Variables**
Age	Religion
Gender	Joint Crystals
Heart Rate	Immunofixation
Admission Type	CD Body Fluids
RBC Count	Iron

Table 3.1: All Algorithms: Common Related and Unrelated Variables

Table 3.1 shows the five related and five unrelated variables that were commonly chosen by all the algorithms that we used in the CES module. The most related variables are considered to be part of the context of MIMIC-III. On the other hand, the unrelated variables are eliminated by all the algorithms since they do not contribute to defining the context of the dataset.

Related Variables	**Unrelated Variables**
Expiry Flag	Language
Ethnicity	Breast Milk
Hemoglobin	Lipase
O2 Flow	Total Protein
LDL Cholesterol	Macrophage

Table 3.2: Forward Selection: Most Related and Unrelated Variables

However, the context of the MIMIC-III dataset for the purpose of disease prediction is not determined by all the algorithms. Since the subset selector determines the context by choosing the best algorithm, we identified that the Forward Selection mechanism produced the best subset of variables for this experiment. Table 3.2 shows the five most related and five most unrelated variables produced by the Forward Selection algorithm, in addition to the variables shown in 3.1, which are not produced by all the other remaining algorithms. Similarly, other algorithms selected other variables in addition to the ones listed in Table 3.1.

3.10 Summary

In this chapter, we discussed the importance of the context in data quality assessment. Although the implicit quality of structured data remains the same, the explicit quality of the data changes based on the situation where the data is used (i.e. the context). To build a comprehensive DQA model, it is necessary to extract the context of the data. We use the feature selection mechanism of machine learning to achieve this goal.

We discussed the advantages and disadvantages of using different types of feature selection algorithms for context extraction. In our research, we use the filter and wrapper-based mechanisms to perform feature selection and define the context of the data. We further explain our framework and its modules for context extraction in structured data.

At first, the context extraction model verifies whether the given dataset is context-dependent or not. If the dataset is context-independent, we simply implement filter methods to get a reduced set of features. On the other hand, if the dataset is context-dependent, we implement wrapper-based algorithms to extract the context of the dataset. Wrapper methods, however, can be implemented in many variations using different search strategies.

In addition to different search strategies, the performance and outcome of wrapper methods change according to the prediction models. In this dissertation, we use decision trees and logistic regression algorithms to finally define the context. We make use of the parallelism inherent in Apache Hadoop to perform resource-intensive tasks of wrapper methods.

Our experimental results show the importance of context extraction for assessing the quality of the data. We also demonstrate the significance of using multiple variants of wrapper methods for the process of context extraction.

CHAPTER IV

RECORD LINKAGE (RL) AND CLOSEST DATASET SEARCH (CDS)

4.1 Introduction

Data quality is typically studied as a single-measure entity. However, there are many dimensions of data quality such as accuracy, validity, consistency, etc. This dissertation focuses on data quality assessment from different dimensions. Structured data quality assessment (SDQA) is discussed extensively in chapter V. Some quality dimensions such as accuracy, consistency, etc. require domain knowledge to measure the quality. One of these requirements is a need for an accurate dataset or a correct source of information. This requirement gets complicated as the input dataset (dataset that is being assessed for quality) changes in schema and size. This chapter describes the Record Linkage (RL) and Closest Dataset Search (CDS) modules, which automate the process of finding a closest accurate dataset from a collection of datasets, to compare it with the input dataset. This process still requires a collection of datasets for each different domain, but heavily reduces the domain dependency as each input dataset needs little or no domain knowledge expert to determine accuracy.

4.2 Framework

As mentioned in the previous section, automating the process of finding the closest accurate dataset is essential to reduce the processing time and domain dependency, as the data varies widely with domain and time. In order to perform this action, two modules Record Linkage (RL) and Closest Dataset Search (CDS) are developed as shown in figure 4.1. The CDS framework consists of three phases: training, Record Linkage (RL), and Closest Dataset search (CDS). The following sections describe these phases.

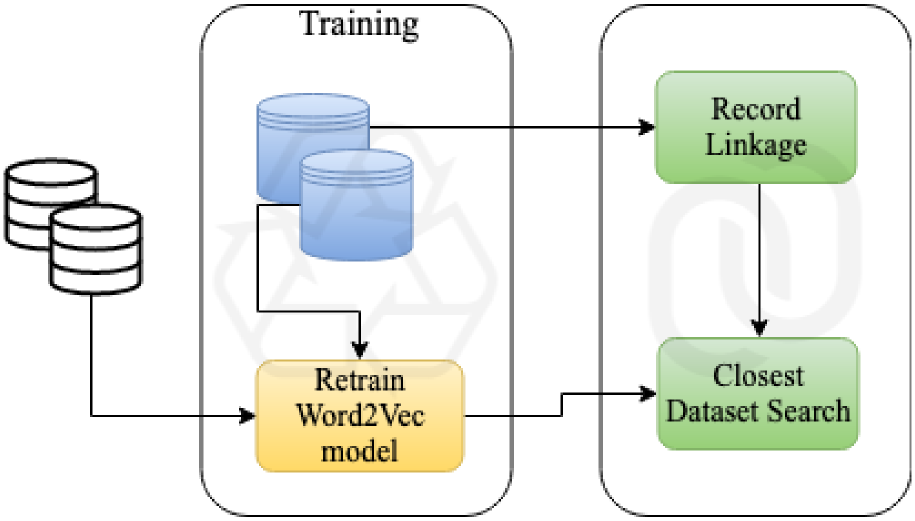

Figure 4.1: Framework for Closest Dataset Search

4.3 Training

Quality assessment of structured data is a challenging task, especially when the data can have diverse formats, as it is difficult to convert non-numeric variables into numeric variables. The conversion is essential for comparison of two different datasets or data items, as this is the primary goal in quality assessment.

Usually, all non-numeric variables are simply converted to categorical variables. A categorical variable is a variable that can take on one of a limited, and usually fixed number of possible values, assigning each individual or other units of observation to a particular group or nominal category based on some qualitative property [102]. This approach is followed in many accuracy assessment models [33][35]. Categorization is beneficial in the field of machine learning only to some extent. However, representing non-numeric variables as categories is not the best approach for quality assessment, especially in some cases such as incorrect spellings, and extremely low or extremely high number of categories.

To overcome these problems, we use word embeddings in our accuracy assessment model. Word embeddings are neural-network models that are trained to extract and compare the words based on relations and backgrounds of the words rather than a simple character matching. Since the relation between words can be obtained using word embeddings, the words can then be represented in a vector space where the distance between any two words gives the relation or similarity between them.

Regardless of the technique used, quality assessment for some dimensions is a straightforward process if the correct dataset or standard values are provided. However, in most domains, having a source of correct information or dataset is highly impossible [74]. Moreover, this requires identifying a correct or accurate dataset for each application domain. The goal of this module is not to provide a correct dataset for each domain as this will require excessive time overheads, particularly in today's realm of big data. Therefore, multiple datasets are used (both relevant and irrelevant) to train the word embedding model. This removes the need to have an accurate dataset for each domain. In order to build an efficient word embedding model, Google's Word2Vec word embeddings [48] model is used as the base model. Even though Word2Vec contains about 100 billion words, there is always a possibility of having new words, especially since the SDQA model can be used for any domain. Hence, as a primary step to the CDS module, the existing Word2Vec model is loaded and re-trained with new words from both the existing collection of datasets and the incoming input datasets (dataset for which the quality will be assessed).

4.4 Record Linkage (RL)

Another important problem that we address in this chapter is choosing an optimal dataset for quality assessment. As mentioned in section 4.3, having a single source of correct information is likely impossible in many domains. Thus, choosing the best dataset for comparison would be the crucial step before assessing the quality of the input dataset. This best dataset must be selected automatically without expert input. Since multiple datasets are considered, which includes relevant and irrelevant datasets of a particular domain, it is necessary to filter the datasets that are irrelevant to the input dataset. There are many efficient techniques to perform this action [19][90]. In this dissertation, the Python Record Linkage Toolkit [37] is used to find the subset of relevant datasets to the input dataset. There are multiple algorithms available in the toolkit for dataset indexing. The best algorithm is chosen by the toolkit depending on the requirement, such as the 'Blocking' algorithm for computationally intensive large datasets, 'SortedNeighborhood' algorithm for datasets with the possibility of a large number of spelling mistakes, etc. After performing the record linkage operation, the resultant subset comes as a collection of datasets relevant to the input dataset.

4.5 Closest Dataset Search (CDS)

The next step is to choose the optimal/closest dataset in the subset obtained from the previous RL step. To obtain the closest dataset, the word embeddings model obtained from the training step is used. After comparing the input dataset with the subset of relevant datasets, the dataset with the lowest average distance is considered as the closest dataset. In other words, this closest dataset is considered as the source of correct information available. The average distance between any two datasets is calculated using the equation 4.1.

$$dist(N, X) = \frac{\sqrt{\Sigma_{i=1}^{m}\Sigma_{j=1}^{n}\left|n_{ij} - x_{ij}\right|^{2}}}{m * n} \tag{4.1}$$

where, m = number of records in the dataset

n = number of variables in the dataset

N = new dataset

X = a dataset from the subset of relevant datasets

n_{ij} and x_{ij} represents the data item of i^{th} record and j^{th} variable of N and X datasets respectively

The process of closest dataset search is shown in algorithm 4.1.

Algorithm 4.1 Closest Dataset Search Algorithm

```
Create a pool of sample datasets
Train the Word Embedding model with the sample datasets and the input dataset
for each sample dataset do
    Perform Record Linkage with input dataset
    if the records and columns match then
        Add the sample dataset to the shortlisted datasets
    else
        Discard the dataset
    end if
end for
for each shortlisted dataset do
    Perform similarity measure for non-numeric values using Word Embedding
    Perform direct comparison for numeric values
end for
Identify the closest dataset to the input dataset
```

4.6 Summary

Data quality is a multi-dimensional entity. The requirements and measurement techniques vary depending on the dimension. The dimensions which deal with the semantic quality of the data require a source of correction information, in other words, an accurate dataset to compare with. Manual processing of numerous accurate datasets is a tedious task and requires high domain knowledge. To overcome this problem, an automated mechanism to identify a closest accurate dataset is designed in this chapter.

There are three stages to obtain a closest dataset namely: training, record linkage, and closest dataset search. In order to perform the comparison of data items, we used a neural network model called word embeddings to support character datatypes and acquire closest data items within a particular attribute. During the training stage, the neural network model is trained to process the new data items each time.

The record linkage module compares the input dataset with a collection of multiple relevant and irrelevant datasets. This module after comparison provides a subset of relevant datasets to the input dataset. At the final stage, the closest dataset is decided based on the lowest average distance between the input dataset and the subset of relevant datasets. Finally, the closest dataset is considered as the accurate dataset or the source of correct information.

Since the closest accurate dataset plays an important role in determining the quality of the input dataset, it is essential to observe its efficiency. However, the efficiency of the CDS module can only be determined by how well the module helps in evaluating the quality of the input dataset, and not by the accurate dataset itself. Therefore, we conduct experiments on the CDS module by combining it with the quality assessment (SDQA) in the next chapter.

CHAPTER V

STRUCTURED DATA QUALITY ASSESSMENT FRAMEWORK (SDQA)

5.1 Introduction

Data quality measures the value of the data in multiple aspects. Although there are numerous dimensions of data quality as discussed in chapter II, every dataset is different and may not possess all quality dimensions. Many data-driven companies have a definite set of quality dimensions to assess the overall quality of their data. However, certain quality dimensions are common to almost every domain. Therefore, in this dissertation, we design a DQA model that supports four important quality dimensions. Nonetheless, the proposed DQA model can adopt other quality dimensions with ease.

Assessment of data quality not only varies by domain but also changes based on the type of data. Since some quality dimensions require an accurate dataset or a correct source of information for measurement, we developed an algorithm to automatically identify the closest accurate dataset in the previous chapter. In this chapter, we propose a methodology to evaluate the overall quality of structured data. We accomplish this goal by measuring the individual scores of four quality dimensions namely: completeness, validity, accuracy, and consistency. Depending on the quality dimension, we either use the context or the closest accurate dataset or both obtained from the previous steps.

5.2 Related Work

Data quality dimensions are mainly classified into two types: intrinsic and contextual. Intrinsic quality dimensions refer to measuring the data values themselves outside of any association with a data element or a record. Contextual quality dimensions refer to the data elements concerning other data elements or from one record to other records. Before assessing the quality dimensions of structured data, it is essential to understand the nature of quality dimensions. David Loshin proposed a classification of quality dimensions for structured data in [74]. Figure 5.1 shows the categorization of quality dimensions into intrinsic and contextual.

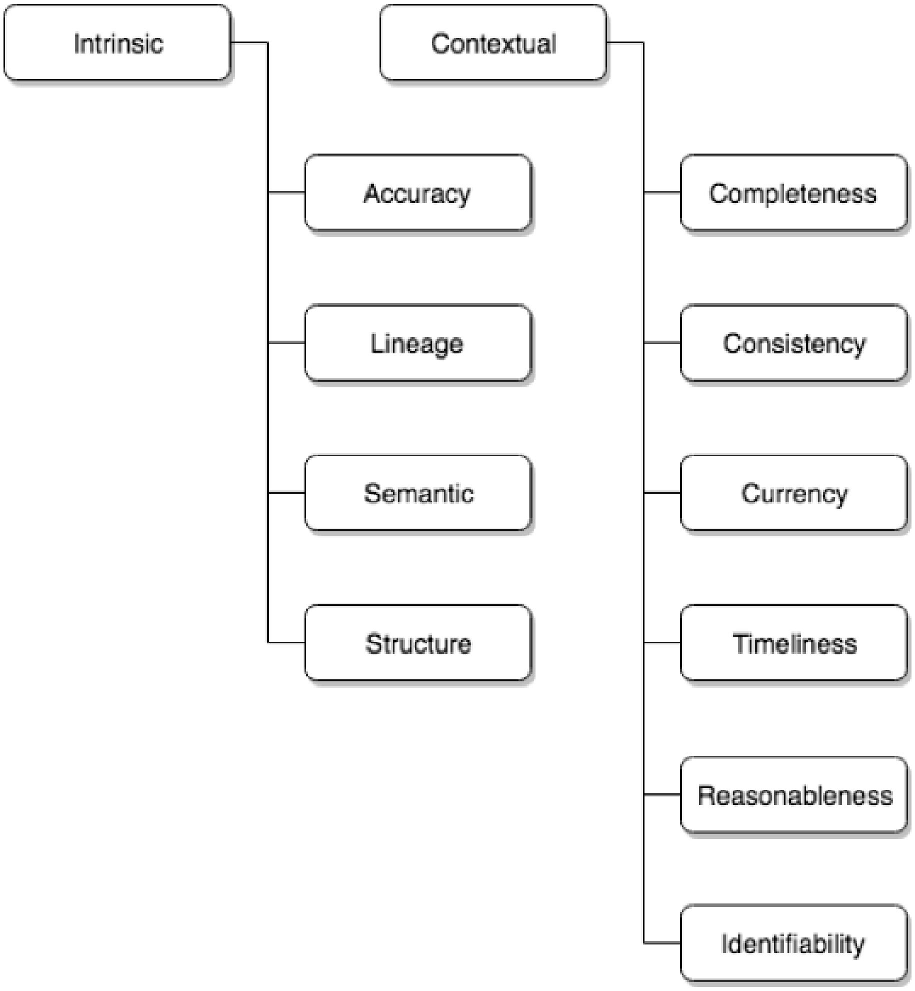

Figure 5.1: Categories of Data Quality Dimensions

Among the four quality dimensions that we consider in our research, validity (also called structure), and accuracy come under the intrinsic type of dimensions, meaning, the evaluation of these dimensions does not change with context. On the other hand, completeness and consistency belong to contextual quality dimensions which change according to the context of the data. However, evaluating completeness and validity is a straightforward process, as they do not require an accurate dataset to compare with.

Data accuracy is one of the most important and challenging dimensions of data quality. An Accuracy Assessment Algorithm (AAA) based on probability theory was proposed by V. Sessions et. al. in [94]. With no prior knowledge, this algorithm estimates the accuracy levels of a dataset based on a few predefined significance levels, learned using the PC algorithm proposed in [99]. Although this algorithm is capable of assessing the accuracy of the dataset without prior knowledge using Bayesian Networks, it is not suitable for large datasets.

Jingling Zhou et. al. proposed a search and score-based data accuracy assessment tool in [118]. This model considers only a small subset of the dataset that needs to be assessed and manually calculates the accuracy. The subset and the remaining dataset are then learned using Bayesian Networks using a score-based mechanism. Finally, the Euclidean and Jaccard distances between those datasets are calculated to determine the overall accuracy of the dataset. In [35], Robert Crone proposed a Veracity Assessment framework for big data. Like any other accuracy assessment model, this model requires an accurate dataset to compare with. However, it depends on a domain expert to provide this accurate dataset. The accuracy of a new dataset is calculated based on its usefulness when combined with an existing dataset. This leads to feature selection and creates a completely new dataset.

In the literature, data consistency is typically combined with data cleaning. Unlike DQA, data cleaning alters the data by repairing the inconsistent data values. These inconsistent data values are recognized with the help of similarity rules. Similarity rules are identical to functional dependencies in relational databases, which are usually described as the relationship between attributes within a dataset. Data is inconsistent if it does not follow the dependency rules and vice versa.

Samir Al-janabi et al in [10] proposed a data cleaning model that focuses on repairing inconsistent values and discover the accurate values in data. The authors utilize embedded density information in data and functional dependencies to fix errors. The density of data is determined based on assigned confidence scores. Though similar cleaning models were proposed in [32][39], all of them rely on predefined rules. Unlike small datasets, it is virtually impossible to define dependency rules for large datasets.

Wenfei Fan et. al. proposed a bread-first search technique to discover conditional functional dependencies in a dataset [43]. Similarly, a depth-first search approach to identify functional dependencies within a dataset was proposed by Catharine Wyss et al in [115]. The primary goal of these studies is to discover dependencies without the aim of data cleaning. Therefore, these methods are not suitable for data consistency.

A more efficient dependency discovery method was proposed by Loredana Caruccio et al in [27]. This model discovers Approximate Functional Dependencies (AFDs) using genetic algorithms. A similar AFD discovery algorithm was proposed by Sebastian Kruse et al [65]. In contrast to FDs and CFDs, AFDs are more practical for discovering dependency rules for large datasets. However, discovering AFDs for consistency assessment is computationally expensive and cannot handle datasets with more than 30 variables and 250,000 records [20].

For both accuracy and consistency assessments, there exist several drawbacks in the literature including:

- Failure to perform contextual assessments.

- Lack of support for large datasets.
- Heavy domain dependence for data comparison.

To overcome the problems in the literature, we propose a multi-domain context-aware quality assessment model that supports big data using machine learning techniques.

5.3 Framework

The aforementioned, Structured Data Quality Assessment (SDQA) has certain prerequisites to satisfy before measuring the individual scores of quality dimensions. We utilize the Context Extraction (CES), Record Linkage (RL), and Closest Dataset Search (CDS) modules proposed in chapters III and IV. However, all the quality dimensions do not require these modules for quality estimation. Our SDQA framework is shown in figure 5.2.

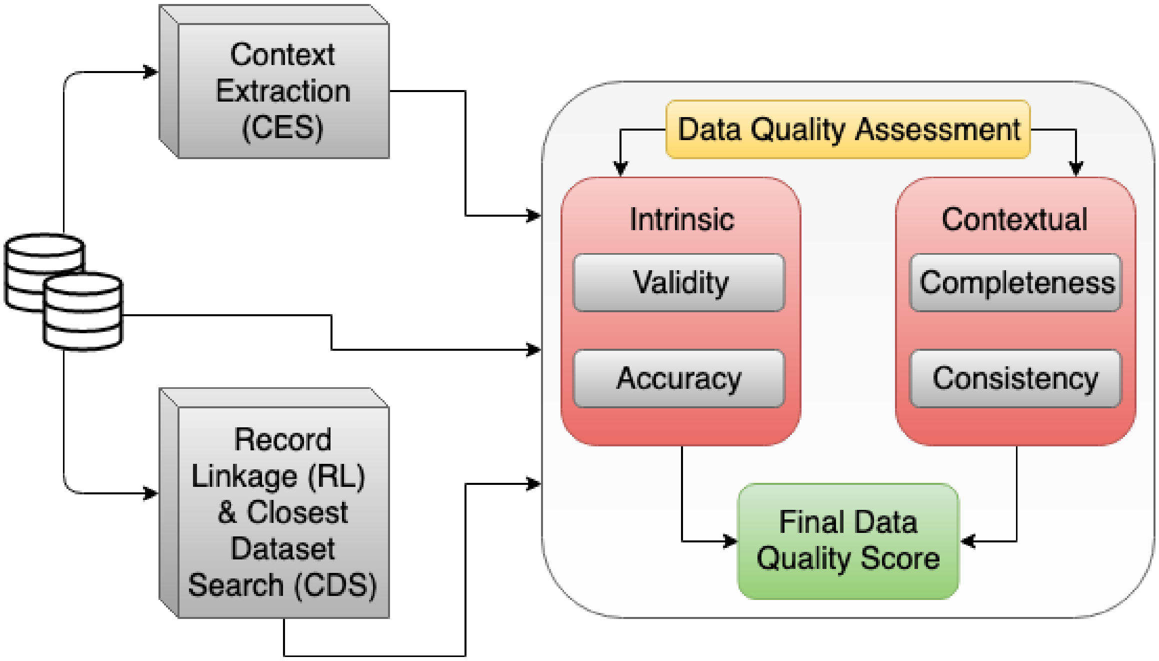

Figure 5.2: Structured Data Quality Assessment Framework

Initially, the input dataset is sent to the CES, RL, and CDS modules to extract the context of the dataset and obtain a closest accurate dataset to compare with. This information along with the input dataset is further transmitted to the quality assessment module. Since validity is an intrinsic quality dimension, it does not require any new information, except for the input dataset.

Although accuracy is an intrinsic quality dimension, it requires an accurate dataset to compare with the input dataset. Therefore, both the input dataset and the closest dataset obtained from the CDS are used for accuracy assessment. Because completeness is a contextual quality dimension, the input dataset along with its context is transmitted to calculate the completeness score.

A quality dimension that requires both context and an accurate dataset is consistency. Based on the context of the input dataset, the dependency rules possessed by the accurate dataset are verified to estimate the consistency score. Finally, the individual scores of these dimensions are combined to calculate a final data quality score.

5.4 Data Quality Assessment

For contextual quality assessment, we use only those features that define the context of the dataset obtained from the CES module. On the other hand, we consider all the features in the dataset for intrinsic quality assessment. In this section, we explain the methodologies to evaluate the individual scores of the four quality dimensions used in our research.

5.4.1 Data Completeness

Completeness is a measure of the availability of required information within a dataset. Apart from missing values, a dataset should not contain extra information. Certain features are expected to have complete values under any circumstance, whereas some features have an optional assignment of values in the data set. Either the lack of required information or the abundance of unnecessary information reduces the quality of data. This makes completeness a contextual quality dimension. We use equation 5.1 to calculate the completeness of an input dataset.

$$d_{completeness} = \frac{number\ of\ complete\ data\ items}{f_{contextual} * n} \tag{5.1}$$

where, $f_{contextual}$ = number of context based features that cannot have null values
n = number of total records in the dataset

5.4.2 Data Validity

Data validity measure, as the name suggests, verifies whether the input dataset satisfies the syntax rules set by a domain expert or not. Automation of this process is not possible, as the format of a dataset is what defines a structured dataset. However, constant monitoring by a domain expert is not necessary, as the format for each attribute is only required once, which can then be stored in the metadata of a dataset. Some examples of data format include the length of the data item, type of data, etc. Unlike data completeness, this quality dimension is independent of feature correlation and applies to all the features of the data. Equation 5.2 is used to derive the score of data validity.

$$d_{validity} = \frac{number\ of\ valid\ data\ items}{n * f} \tag{5.2}$$

where, n = number of records
f = total number of features in the dataset

5.4.3 Data Accuracy

Although each quality dimension has its importance in the evaluation of data quality, data accuracy generally dominates the other dimensions, as it verifies the correctness of the information by comparing it with an accurate dataset. Data accuracy is an intrinsic quality dimension that must satisfy the semantic requirements at any given instance. Therefore, we consider all the features of an input dataset for accuracy assessment. A typical accuracy assessment is performed based on the average distance calculated between the input and accurate datasets. However, our SDAQ model performs this operation differently by utilizing word embeddings for non-numeric variables. Equation 5.3 is used to derive the accuracy based on average distances and word similarities.

$$d_{accuracy} = \frac{\Sigma_{i=1}^{m}\Sigma_{j=1}^{n}(1 - \frac{|n_{ij}-c_{ij}|}{max(n_{ij},c_{ij})})}{m * n} \tag{5.3}$$

where, m = number of records in the dataset

n = number of variables in the dataset

$d_{accuracy}$ = intrinsic data accuracy of the new dataset

n_{ij} and c_{ij} represents the data item of i^{th} record and j^{th} variable of the new dataset (N) and the correct dataset respectively

5.4.4 Data Consistency

Consistency is a measure of the integrity of a data value at multiple instances across the dataset. In other words, consistency assessment verifies whether the data satisfies the required semantic constraints or not. As discussed in section 5.2, dependency rules are used in the literature to assess the consistency of the data. However, almost all the studies require predefined dependency rules, which is virtually impossible to provide for all datasets.

Despite the automatic discovery of dependencies that have been proposed in the literature, most of them do not support big data and do not refer to an accurate dataset. In our dissertation, we use Approximate Functional Dependencies (AFDs) to measure the data consistency. Since consistency assessment does not alter the data, using approximate dependencies for evaluation is a sufficient and viable approach.

Mutual Information

In our research, we use Mutual Information (MI) to find approximate dependencies in a dataset. Mutual Information [34] is a statistical quantity that measures the mutual dependence between two variables in a dataset. In other words, it measures how much one variable tells about another. Unlike variable correlation, MI is capable of finding dependencies for both linear and non-linear variables.

MI calculates the reduction of uncertainty about one variable, given the knowledge of another variable. The value of MI is 0 if the variables are independent, whereas, it is inversely proportional to the uncertainty of variables. As MI handles the uncertainty of a variable, it is associated with entropy. Entropy is a measure of available information. The value of entropy is high if the knowledge about a variable is low and vice-versa. Entropy is calculated using equation 5.4.

$$H(X) = -\sum P_X(x) \log P_X(x) \tag{5.4}$$

where, X = a random variable

x = data items in variable X

$P_X(x)$ = probability of appearance of a data item x in variable X

However, mutual information is dependent on both entropy and conditional entropy of variables. Equation 5.5 gives the conditional entropy of variable X after observing variable Y.

$$H(X|Y) = \sum_{y} P_Y(y) \Big[-\sum_{x} P_{X|Y}(x|y) \log(P_{X|Y}(x|y)) \Big] \tag{5.5}$$

where,

$P_{X|Y}(x|y)$ is the conditional probability of x given y

Based on equations 5.4 and 5.5, the mutual information of two random variables X and Y is calculated using equation 5.6.

$$I(X;Y) \equiv H(X) - H(X|Y) \equiv H(Y) - H(Y|X) \tag{5.6}$$

where,

I(X; Y) = Mutual Information between variables X and Y

H(X) = entropy of variable X

H(Y) = entropy of variable Y

$H(X|Y)$ = conditional entropy of X given Y

$H(Y|X)$ = conditional entropy of Y given X

Acquiring dependency rules from the same dataset that is being assessed for consistency will produce inaccurate results. Therefore, we use the accurate dataset obtained from the CDS module. Moreover, since consistency is a contextual quality dimension, we consider only the features generated by the CES model. We calculate MI for all possible pairs of features or variables from the accurate dataset. However, the value of MI is always a non-negative number ranging from 0 to infinity. Thus, we normalize the mutual information score to vary between 0 and 1, where 0 indicates variables are independent and 1 indicates a high dependence between the variables. If two variables are dependent, we add the pair to our list of dependency rules. We consider two variables to be dependent if their MI score is above a certain threshold value. Based on multiple simulations, we observed the median and standard deviation of dependencies for labeled datasets and set the dependency threshold to be 0.93. After identifying the dependencies in both input and accurate datasets, we finally match them and calculate the consistency of the input dataset using equation 5.7.

$$d_{consistency} = \frac{|d_N \cap d_A|}{|d_A|} \tag{5.7}$$

where,

d_N = dependencies in condensed dataset (feature subset defining the context), N

d_A = dependencies in accurate dataset, A

5.4.5 Final Data Quality Score

Unlike the individual estimations of data quality dimensions, the overall quality score of a dataset is defined by the importance of each quality dimension for a particular domain. The relevance and significance of quality dimensions differ by domain. Therefore, a numerical score to determine the overall quality of the data is dependent on a domain expert. Though our model cannot automate this process, obtaining a list of relevant dimensions and their weights in a particular domain is a one-time process. Moreover, the dimensions and their weights do not change from dataset to dataset within the same domain. As discussed earlier in this section, we use the individual scores obtained for each quality dimension and apply them in equation 5.8 to estimate the overall quality of an input dataset.

$$f_s = \frac{\Sigma_{i=1}^{n} d_i * W_i}{n} \tag{5.8}$$

where, n = number of quality dimensions relevant to a particular domain

d_i = individual score of quality dimension i

W_i = weight of quality dimension i

5.5 Experiments and Results

The outcome of our SDQA model is the overall quality score of an input dataset. The validation of the outcome is only possible with sample labeled datasets. In other words, the quality score can only be verified if a dataset is manually assessed for quality. Therefore, it is practically impossible to verify the estimated quality scores without intervention by a domain expert. The following are a few reasons for the outcome to be unverifiable:

- Every dataset is different and correct validation of one dataset cannot be assumed to be true of the other.
- Manual assessment of large datasets for verification (to create labels) is virtually impossible.
- No two assessment approaches can give identical results.

For the above reasons, the outcome of our SDQA framework is non-verifiable. However, without proper observations, a model cannot be determined to be a good model. Thus, we intend to test our approaches that aid in performing the quality assessment. One important step in our quality assessment process is the closest dataset search. We use the dataset provided by the CDS module to assess the accuracy and consistency dimensions. For that reason, we intend to test our CDS module for consistency and accuracy assessment.

The CDS module identifies the closest accurate dataset to the input dataset from a collection of irrelevant and relevant datasets. Aforementioned, we intend to develop a comprehensive DQA model that supports multiple domains. As a result, to test our assessment model, we

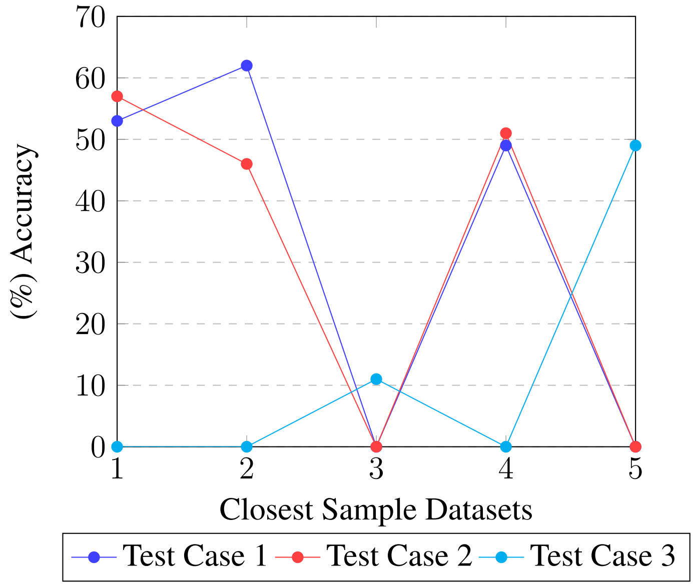

Figure 5.3: Sample datasets vs Data accuracy

considered datasets from multiple domains.

Accuracy Assessment:

First, to represent the group of relevant and irrelevant datasets, we considered a total of 10 datasets, that comprises 7 datasets with climate information of 7 U.S. states and 3 datasets with demographic and employee information of a fictitious firm. We collected the climate data from the National Centers for Environmental Information [45]. The employee information datasets are synthetically generated.

We considered 3 different test cases for the evaluation of data accuracy. Climate data is generally similar within a region or country. So, to test our record linkage module, we considered two test cases of climate data each from different U.S. states (Oklahoma and Vermont). The final test case is an employee dataset similar to those in the dataset collection.

Although the collection of sample datasets contains 10 datasets, we considered the top 5 respective closest datasets to each of the test cases. Figure 5.3 shows the accuracy of 3 test case datasets (Oklahoma climate, Vermont climate, and 1 employee datasets respectively). As shown in the figure, the accuracy of the test cases for each of the sample datasets varies between 0% and less than 70%. The sample dataset which gives the highest accuracy to a particular test case indicates that it is the closest dataset to that particular test case and the numeric accuracy indicates the data accuracy of the test case dataset. The test case datasets that possess 0% accuracy when compared to some sample datasets indicate that the respective sample dataset is irrelevant to the test case dataset. For example, the test case dataset-3 has 0% accuracy to most of the sample datasets. The primary reason for this situation is because 3 sample datasets that are related to employee information compared to 7 sample datasets of climate information. Likewise, this can be observed with other test cases as test cases 1 and 2 have 0% accuracy with respect to some samples, where test case 3 has more than 0% accuracy. This experiment shows the efficiency of the CDS module in the accuracy assessment.

Consistency Assessment:

A similar experiment concerning consistency assessment was performed to evaluate the CDS module. The collection of sample datasets now has 15 datasets, 5 each from three different data sources. The three different data sources are climate data, employee information, and MIMIC-III medical dataset (discussed in chapter III). Note that for every new input test dataset, any dataset from a different data source will be irrelevant. Apart from the group of relevant and irrelevant datasets, we also consider 3 test case datasets, one each from the 3 data sources.

Quality assessment without a reference to an accurate dataset may lead to inaccurate analysis. Thus, we incorporated the closest accurate data search in our model to identify the dependency rules. Using these dependency rules of an accurate dataset, we assess the consistency of a test dataset. The consistency of 3 test case datasets based on their respective relevant datasets is shown in figure 5.4. The consistency of test case datasets changes if

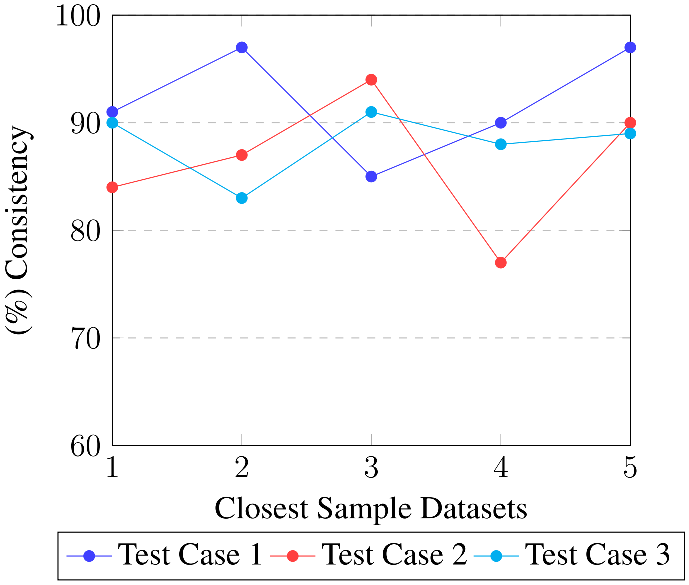

Figure 5.4: Sample datasets vs Data consistency

the dependency rules are compared with irrelevant datasets. This shows the importance of considering an accurate dataset for consistency assessment. A test case dataset can have varying consistency scores for sample datasets within the same source. However, a high consistency does not always mean that it is the true consistency of the test dataset unless the consistency is calculated based on the semantic accuracy of the sample dataset (i.e. accurate dataset).

We discover approximate dependency rules within a dataset to estimate the consistency of a dataset. Our approach supports big and high dimensional datasets by reducing the time taken to discover the dependency rules. Although approximate dependencies vary widely in approximation, our main goal is to assess the consistency, which does not alter the data and change the data accuracy in any way. The existing dependency discovery algorithms cannot support datasets with more than 30 features. Moreover, they discover the dependencies to repair or modify the datasets. We test the execution time of our model by gradually increasing the number of features in a dataset. Fig. 5.5 shows the run time of our model where the x-axis indicates the feature count and the y-axis shows the run time in

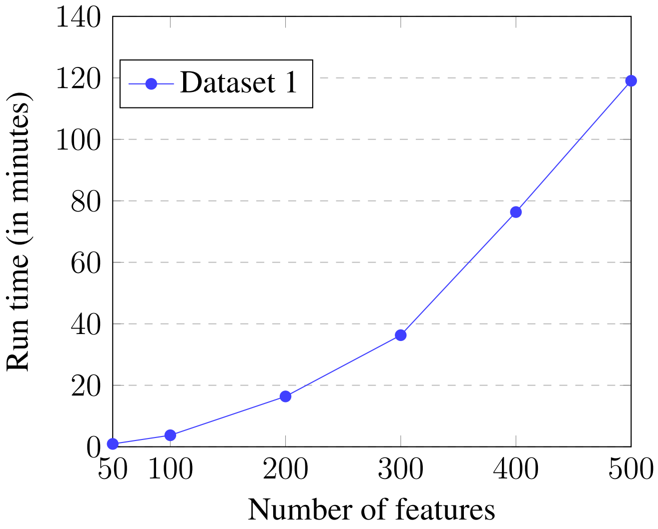

Figure 5.5: Feature count vs Execution time

minutes.

5.6 Summary

Data quality has multiple dimensions such as accuracy, consistency, timeliness. Although every data-driven company does not consider all quality dimensions to measure the overall quality of a dataset, certain dimensions are important in every domain. In this chapter, we devise various approaches to estimate the individual scores of four important quality dimensions namely: completeness, validity, accuracy, and consistency. We then calculate a final data quality score based on the individual scores and their weights which varies by domain.

The estimation of completeness and validity are straightforward as they do not require any additional dataset for comparison. Completeness is a contextual quality dimension that considers only the subset of features produced by the Context Extraction (CES) module. The completeness of a dataset is calculated based on the number of missing values in a dataset. On the other hand, as validity is an intrinsic dimension, it considers all the features in a dataset. Validity verifies the syntactic constraints of a dataset.

Assessment of accuracy and consistency requires an accurate dataset to compare with. We utilize the closest accurate dataset identified by the CDS module. Also, since consistency is a contextual quality dimension, it considers only the features extracted by our CES module. We use a combination of average distances and word similarities for assessing the accuracy of a dataset. For consistency assessment, we identify approximate dependency rules using mutual information.

Our experimental results show the importance and efficiency of our CDS module in the evaluation of accuracy and consistency. We show how accuracy and consistency differ by comparing the input dataset with both relevant and irrelevant datasets.

CHAPTER VI

CONTEXT EXTRACTION IN TEXTUAL DATA (CET)

6.1 Introduction

So far, this dissertation focused on assessing the quality of structured datasets. However, this is only one side of the coin. Other than having a definite structure, data has numerous unstructured forms such as text, image, audio, and video. Unstructured data occupies the major share in present-day big data. Hence, there is also a need to assess the quality of unstructured data. However, quality assessment has to be performed differently for different types of unstructured data. In this dissertation, we focus only on the textual form of unstructured data.

Although the procedure to assess the quality of textual data is different, the approach is analogous to that of structured data. Context is essential for quality assessment, regardless of the type of data. Therefore, context extraction will be the initial step for the quality assessment of textual data. However, unlike in structured data, the perception of context is not simply the importance of features based on a target variable. In textual data, context is not “Where the data is used.” or “How important the data is for a situation.”, but “What is the data explaining.” and “How the topic is explained.” Textual data is very dynamic as a topic can be depicted in multiple ways within the same text. Thus, identifying the context in textual data is a challenging task. In this chapter, we demonstrate our methodology to extract the context from text documents using natural language processing and machine learning techniques.

6.2 Related Work

Context extraction using Natural Language Processing (NLP) has become an important research topic in recent years. Data context plays a key role in performing any kind of analysis on textual data. This is true even in the case of data preprocessing as the action of cleaning and repairing of data depends on implicit information. Many studies in the literature addressed this problem of context extraction from textual data.

As discussed in the previous section, context extraction is a hard problem and there can be no foolproof solution to it. The basic understanding of context is identifying the topic that a text is explaining. For this reason, numerous studies focused on extracting the topic of a text. This technique is popularly termed as Topic Modeling (TM). There are many ways to perform topic modeling, and each has its advantages and disadvantages. One popular method to extract the topic is using n-grams. N-grams are collections of words that represent the topics, phrases, and concepts occurring in a text. There are multiple levels in n-grams, where n represents the number of words in the collection. Jayaraman et. al. proposed a keyword topic model using n-grams [57]. Nikolenko et. al. proposed a topic model using TF-IDF (Term Frequency-Inverse Document Frequency) mechanism which extracts the words that frequently occur in a text [81]. Similar models have been proposed in [8][70].

However, n-grams and TF-IDF mechanisms are naïve approaches to extract topics from a text as they do not necessarily consider the semantics of the words. Therefore, a hybrid approach for topic modeling was proposed by Lee et. al. in [69]. Basic NLP techniques retain a lot of noise in the text which provides redundant information to extract the topic. In [117], Zhang et. al. proposed a topic model that supports a noise filtering mechanism. A slightly more advanced technique using a clustering algorithm was proposed in [116][14].

Nonetheless, these techniques though perform better compared to the basic topic models, they cannot still extract the topic based on the semantics on the text. Sentiment analysis is an application of NLP that focuses on identifying expressions that reflect authors' opinion-based attitude (i.e., good or bad, like or dislike) toward entities (e.g., products, topics, issues) or facets of them (e.g., price, quality) [26]. Even though the semantic relation between words is not considered in sentiment analysis, this method still gives more information than just a topic. Therefore, Sowmiya et. al. proposed a topic model using sentiment analysis in [98]. Similar models have been proposed in [83][15]. These models still do not provide the context of the data.

An advanced topic model provides valuable information about the text. However, the context of a text is more than just a topic. Context not only tells what the text is about but also explains how a topic is expressed and how much importance it carries in the text. Although few studies focused on context-based analysis [76], they can only discern the context to perform prediction such as sentence completion and classification. Therefore, to overcome the problems in the literature, we present an efficient context extraction mechanism using natural language processing and deep neural networks.

6.3 Framework

The framework for Context Extraction in Textual Data (CET) is shown in figure 6.1. There are three main components in our CET model namely topic extraction, sentence matching, and hybrid sentiment analysis. Initially, the topic extraction module derives the essential topics from the raw text. These topics along with the raw text are further processed by sentence matching to obtain the importance of each topic within the text. Finally, the expression of the topics based on the importance is identified by the hybrid sentiment analysis

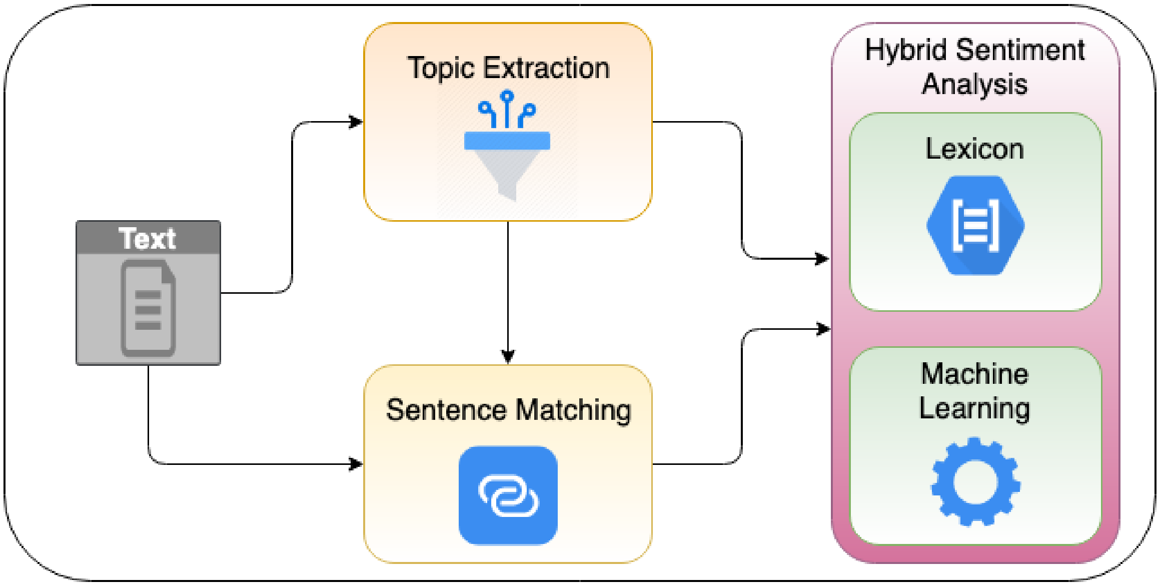

Figure 6.1: Context Extraction Framework for Textual Data

model. The hybrid model consists of two levels to perform sentiment analysis, i.e. lexicon and machine learning. Each of these components of our CET model is discussed in the following sections of this chapter.

6.4 Context Extraction

Context extraction is a complex task regardless of the application. It is a critical phase in the quality assessment of textual data. Though it is possible to determine the quality of a text without the context, it requires a tedious manual interference to comprehend the data. This task becomes practically impossible with large documents. This section describes an automated tool for multi-domain context extraction in textual data.

6.4.1 Topic Extraction

The aforementioned topic extraction is possible with basic NLP techniques such as n-grams, TF-IDF, and filtering tools. However, in this dissertation, since we intend to determine the topic as a part of context extraction, we perform a semantic analysis of raw data. To achieve this goal, we do not perform conventional pre-processing activities such as stemming (obtaining a stem word), and removing of numbers, punctuation, and stop words (is, that, which, etc.), as these techniques may help to extract the topic, but loses valuable information that could be contributing to the context of the data.

In order to avoid the issues caused by case-sensitivity, we first convert the raw text into lower case. There could be multiple instances of a topic in the entire text document. The impact or the relevance of each topic could vary depending on the sentence. Therefore, we then split the lower-case raw text into sentences. Using natural language processing, we perform Parts Of Speech tagging (POS) for every word in each sentence. Each part of speech has its significance in determining the meaning of a sentence. In this dissertation, we consider all types of nouns and verbs to extract the topic from a sentence. However, if needed, other parts of speech can be included with ease. There could be multiple instances of a word or its reference in a document. Since we process the document by each sentence, it is likely to obtain the same word multiple times. Therefore, we remove the duplicates to obtain a unique list of words in a document. At this point, these words are not necessarily topics. Hence, we name them as potential topics.

6.4.2 Sentence Matching

In the previous step, we extract the potential topics from an input text document. Many studies in the literature identify the importance of a topic based on its repetition level in a document or relation to a set of words in a predefined window size. However, we intend to extract the semantic relevance of each potential topic for every sentence in the document. We perform this operation with the help of deep neural networks.

Word2Vec [48] is a collection of neural networks that produce word embeddings from a large corpus of text. Each word in the corpus is represented in a vector space, where semantically related words are close to each other. However, this model cannot compare or find a relation between sentences. For this reason, a new model at paragraph level named Doc2Vec was proposed by Le et. al. in [68]. This model represents the words with an additional dimension in vector space for the specific paragraph it belongs to. Therefore, we use Doc2Vec deep neural networks to identify the relationship between a word and a sentence. However, since Doc2Vec identifies the embeddings at the paragraph level, no sentence, paragraph, or document can be pre-trained as the model considers the context, and the relations are identified dynamically based on Word2Vec.

The raw input text is transmitted to the Doc2Vec network for training. Upon training, each potential topic is then compared to all the sentences in the input document. The similarity score between a topic and a sentence varies between 0 and 1. Where 0 indicates the topic is not relevant to the sentence, 1 indicates that the topic is highly related to that sentence. Since Doc2Vec compares the sentences based on the context and word embeddings, it is highly probable for a topic to be relevant to a sentence that does not contain this particular word or its reference. Therefore, we find the relevance between a topic and all the sentences in the input text. Finally, the average relevancy is calculated for each potential topic. Unlike the models mentioned in the literature, this method extracts the semantic relevance of every topic in the entire document.

6.4.3 Hybrid Sentiment Analysis

Our CET model so far answered the question, “What the data is explaining about.” Although at this stage, there exist multiple topics with a ranking system (relevance score), this alone will not convey the context of a text. We still need to focus on the question, “How the topic is explained.”

This is the fundamental difference between a topic and a context. To address this question, we use a hybrid sentiment analysis model.

Lexicon

The basic methodology to implement sentiment analysis is using a lexicon. A lexicon is a vocabulary of a language (English, in our case). A lexicon for sentiment analysis is a collection of words with labeled sentiment scores. A sentiment score varies from -1 to +1, where -1 indicates that a word is bearing a highly negative sentiment, 0 being neutral, and +1 signifies a highly positive sentiment. However, this lexicon cannot hold all possible words as the sentiment of certain words in a document varies depending on the context. As a result, words (mostly adjectives) that do not change with a context or explain a word (such as a noun) are present in this lexicon.

The topics extracted in our previous step contain only nouns and verbs which do not have a sentiment by itself. Therefore, we extract the sentiment-bearing words associated with the topics from all possible sentences in the document. We then obtain the individual sentiment scores (from the lexicon) for every qualified word associated with each topic. Finally, we calculate the average sentiment score of each topic by aggregating the individual sentiment scores. This process balances the negative and positive scores associated with a topic and thereby provides the overall sentiment of a topic in a given text document.

This concludes the process of context extraction. However, since this model is designed to work for all possible texts in numerous domains, there exist a few drawbacks to this approach. Having a sentiment score solely based on adjectives will not cover all possible contexts of a document. For example, words such as “increasing” and “robbery” can have a sentiment in most cases that necessarily are not associated with an adjective. The word “increasing” can carry both negative and positive sentiments depending on the context, whereas “robbery” carries a negative sentiment in most contexts. This issue cannot be addressed by a lexicon-based analysis containing only adjectives. Therefore, to enhance our sentiment analysis model, we include a machine learning approach to solve this problem.

Machine Learning

A supervised machine learning approach provides a viable solution to identify the sentiment of a word that is not associated with an adjective. We designed a supervised logistic regression model to estimate the sentiment scores of the sentences related to each topic. This model can easily become frail if trained with a smaller dataset. Therefore, we train our machine learning regression model with large datasets [80]. Since this a prediction-based model and margin of error is high compared to the lexicon-based approach, we give less weight to this model.

As mentioned earlier, we calculate the average sentiment of each topic by estimating the individual sentiment scores of every qualified sentence. However, if the topic is not associated with an adjective in a particular sentence, or if an adjective is absent in a sentence, the sentiment score is 0, meaning neutral. In that case, we predict the sentiment scores of only those sentences using a logistic regression model. We then eventually calculate the average sentiment scores of each topic.

6.4.4 Context Evaluation

The context of a text can neither be a single word nor a sentence. In this dissertation, we define context as a topic associated with relevance and sentiment scores. However, a text can contain multiple topics with varying importance. Our primary purpose in using context in this study is quality assessment. We utilize context for automatic identification of a similar text document for quality assessment. Therefore, we use a threshold variable to decide the number of topics (and its scores) to evaluate the context of the input text document. This threshold can be altered depending on the application and its domain. The overview of context extraction for textual data is shown in algorithm 6.1.

Algorithm 6.1 Context Extraction Algorithm for Textual Data

```
Extract the potential topics (nouns and verbs) of the document
for each potential topic do
    for each sentence in the document do
        Calculate the relevance score
    end for
    Calculate the average relevance score
    Calculate the sentiment score based on the Hybrid Sentiment Analysis model
end for
Determine the context of the text document based on its topic, relevance, and sentiment.
```

6.5 Experiments and Results

Although this chapter deals with the context extraction of a text document to assess the quality of the data, it contains multiple aspects associated with the context. Since there is no standard definition for context, there exist no datasets with a context label that suits our definition. Therefore, to test the efficiency of our CET module, we intend to test it based on its components individually, i.e., Topic Extraction, Sentence Matching, and Hybrid Sentiment Analysis.

Each of these components is distinct both in the approach of data comprehension and the purpose of the usage. This situation requires us to use different datasets to experiment with and test their efficiency. Therefore, we use multiple datasets to understand and test our components. However, to the best of our knowledge, there are no labeled datasets available to test the performance of our sentence matching component. Since this component is loosely connected to the similar document extraction (discussed in Chapter VII), we rely on the analysis performed on TDQA to understand the effectiveness of the sentence matching component.

Topic Extraction

Since topic extraction is a popular method to comprehend the data in natural language processing, there are numerous datasets available to test the efficiency of this component. As we aim to develop our quality assessment model for multiple domains, for this experiment, we combine two datasets containing news articles from multiple fields. We combine 20NewsGroup [66] and BBC news datasets [50] to create a collection of more than 21,000 news articles from 5 different domains.

The main metric to determine the efficiency of the topic extraction component is how accurately a model can predict the topic when provided with the text of a document. It is also important that a model predicts the topics not only in one domain but for multiple domains. As shown in Figure 6.2, our topic extraction component predicts the topics with high accuracy in different domains.

Hybrid Sentiment Analysis

The aforementioned, present context extraction mechanisms only consider the topic to determine or understand the situation of the data. As this approach is deficient in terms of comprehending the context, in our work, we use some additional steps to identify the overall data context. It is important to understand the tone of the information in addition to

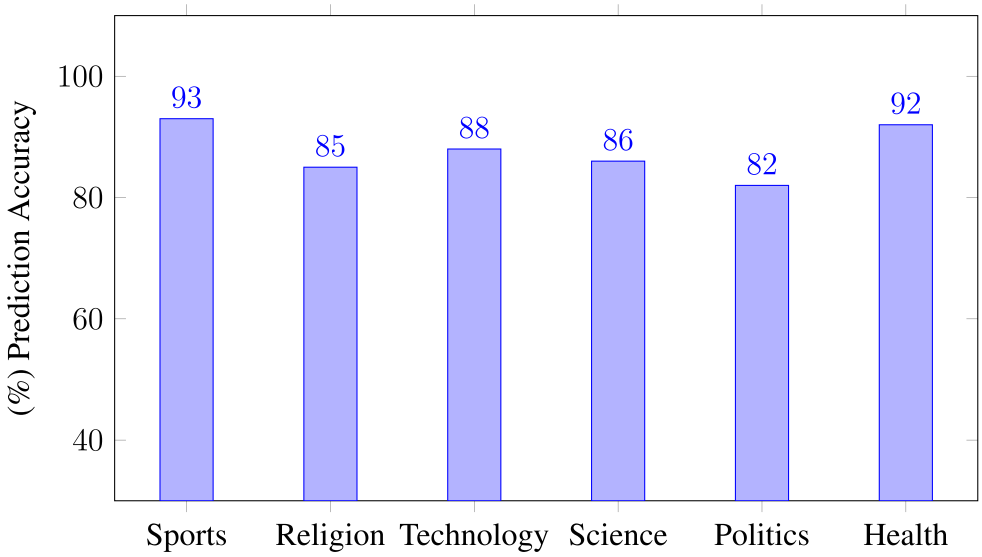

Figure 6.2: Topic Prediction

the topic extracted from the data. As discussed in section 6.4, we achieve this goal by performing sentiment analysis.

Since the previously used datasets do not contain a sentiment score as a label, we use the Amazon Review Dataset [80] to test the effectiveness of our sentiment analysis component. The dataset contains more than 75 million reviews collected from 29 different categories of products sold on Amazon. For experimental purposes, we used a part of this dataset containing around 10 million reviews collected from 5 different categories. Figure 6.3 shows the prediction accuracy of our sentiment analysis component.

As discussed in section 6.4.3, most of the time, sentiment analysis is performed based on a lexicon approach, where the sentiment of a word or a sentence is determined by predefined sentiment scores associated with the adjectives. However, this approach has a shortcoming for certain verbs that carry sentiment in some instances. It is very important to perceive the sentiment and context of the data as accurately as possible, especially because our model is intended to use for data quality assessment. Therefore, we perform a hybrid sentiment analysis to overcome this problem. Figure 6.4 shows the prediction accuracies of

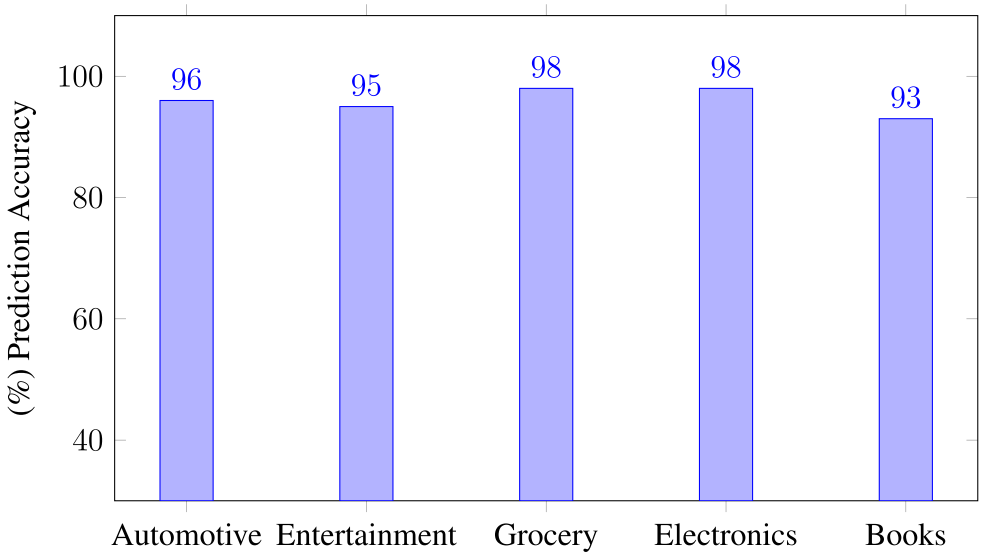

Figure 6.3: Sentiment Prediction

sentiment analysis when performed only using a lexicon versus a hybrid model. Although the accuracy difference is not notable for a few categories, there is a significant difference in most of the categories.

Although we could not test our context extraction module as a whole, nonetheless, the module comprises different components that can be tested individually. The efficiency of our context extraction module can be interpreted from the results obtained from the topic extraction and hybrid sentiment analysis components.

6.6 Summary

No matter what type the data is, context is essential to understand the data better, especially for quality assessment. Therefore, similar to SDQA, context extraction is an essential phase in TDQA. In this chapter, we introduce our Context Extraction for Textual Data (CET) module as the first step in our quality assessment model for textual data.

Unlike structured data, the context of textual data is apparent as to what the data is talking

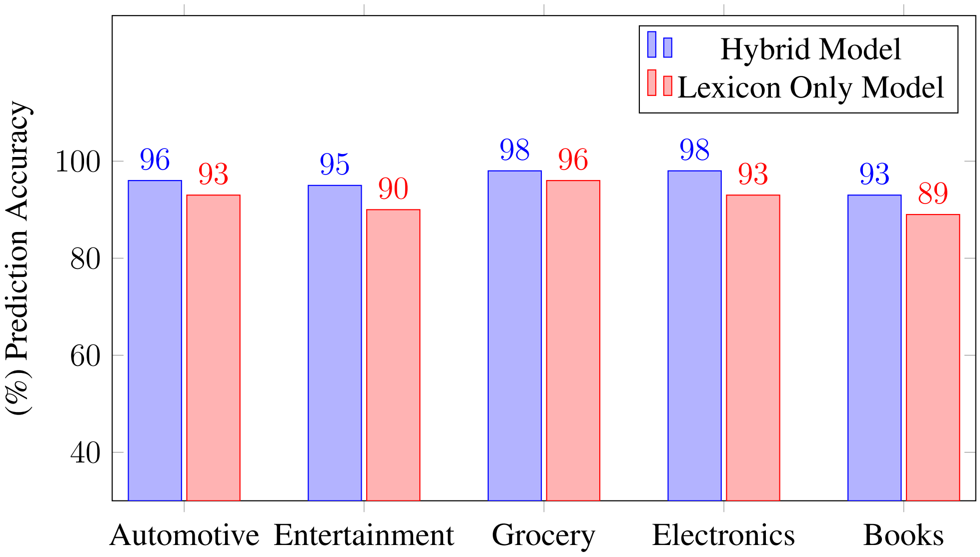

Figure 6.4: Comparison of sentiment analysis models

about, and how it is being presented. However, the automatic extraction of context from textual data is not as simple as it can be defined. Most of the existing models only consider the topic of the document as the context. For quality assessment, this is especially a deficient comprehension as a topic can be depicted in multiple ways.

Our CET module consists of three main components, namely, topic extraction, sentence matching, and hybrid sentiment analysis. The definition of context in our work is defined by these three metrics combined. We perform sentence matching for the topics extracted from a document to identify the relevance of each topic in the entire document. These topics are further analyzed for sentiment as a topic can have either a negative or positive sentiment irrespective of its importance in the document.

Our experimental results show the efficiency of the individual components of our CET module, which indeed reflects the overall performance of our context extraction module for textual data. We also show the importance of hybrid sentiment analysis by comparing different approaches.

CHAPTER VII

TEXTUAL DATA QUALITY ASSESSMENT FRAMEWORK (TDQA)

7.1 Introduction

The lack of a definite structure in the data makes textual data sensitive to changes. The essence of a text in a natural language changes drastically with minor revisions. Accordingly, the quality of the data varies with a change in the meaning of the data. This and intricate prerequisites make a quality assessment of textual data a strenuous task. Since a single dimension cannot be used to interpret the overall quality of the data, our TDQA model examines the two most important quality dimensions of textual data; namely, accuracy and consistency.

Because context in textual data is “What the data is explaining about” and “How the topic is explained,” its application is different in quality estimation. Unlike in SDQA, accuracy assessment also requires the context in TDQA. However, not unlike in SDQA, textual data quality assessment also requires accurate data to compare with. In this chapter, we describe our TDQA model to assess the quality of textual data in two different dimensions.

7.2 Related Work

Analysis of natural language data is extensively studied in the literature. However, since textual data has minimal constraints, data handling techniques must support the dynamic nature of textual data. Moreover, the textual data covering the majority of present-day big data is generated by amateur individuals. This situation escalates the existing quality problem in textual data. Nevertheless, this problem is not well addressed in the literature.

Daniel Sonntag [97] analyzed the quality of text data based on consumer's expectations and classified the quality problems into four major types: Intrinsic, contextual, accessibility, and representational. The authors concluded that automated assessment is possible only on accessibility and representational features of text data.

Cornelia Kiefer [61] identified the indicators to determine the quality of unstructured data. Three quality dimensions namely relevancy, accuracy, and interpretability are observed in this study. The indicators were identified based on the similarity of input data to consumer's expectations and the representation of data in the real world. However, the author did not propose a methodology to assess the indicators of data quality.

Taleb et. al. proposed a quality evaluation model to handle the quality of unstructured big data [104]. The proposed framework is aimed to perform data exploration and feature extraction on textual, media, and web data. This model contains multiple stages ranging from quality requirements gathering to assessment. However, this theoretical framework is still under development, and therefore, does not provide a practical implementation.

To address the issues in the literature, we developed a comprehensive methodology to assess the quality of textual data in two aspects i.e. accuracy and consistency.

7.3 Framework

Figure 7.1 shows the quality assessment framework for textual data. Similar to SDQA, the input data is initially processed in TDQA to extract the context of the data. In SDQA, since context is verifying the data for “fitness for use”, the closest data is identified only after

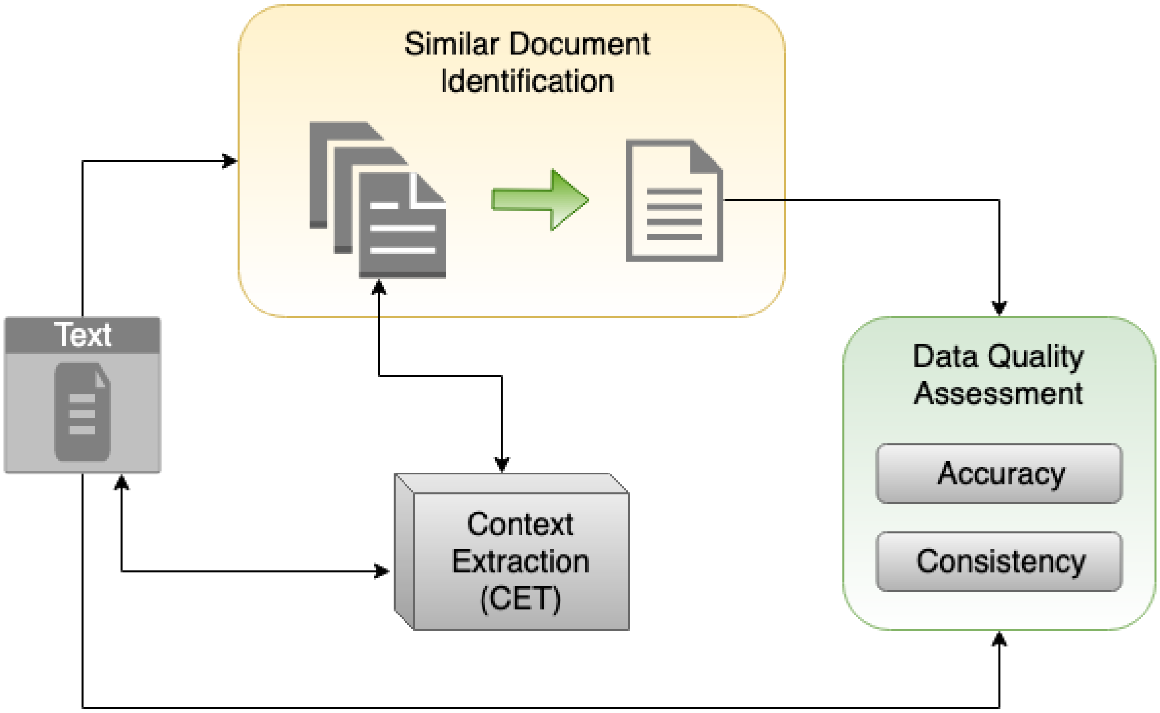

Figure 7.1: Textual Data Quality Assessment Framework

context extraction. However, the document identification is parallelly processed in TDQA with the raw input text, which indeed also consults the context extraction (CET) module to identify a similar document. Finally, the quality assessment module compares the input text with a similar text document to evaluate the accuracy and consistency of the input.

7.4 Identification of Similar Text Document

Since we deal with the accuracy and consistency dimensions of data quality in TDQA, apart from the context, a text document that is being assessed for quality also requires an accurate document to compare with. However, an accurate document matching all the criteria such as words and sentence formation is practically impossible to acquire in every situation. Therefore, we designed our TDQA model to identify a similar document that semantically matches with the input document. We achieve this goal by using the same Doc2Vec neural network model developed for context extraction in chapter VI. Besides paragraph-level, Doc2Vec also performs the comparison at the document-level. The Similar Document Identification module expects manual intervention to provide a collection of related text documents for each domain. Nonetheless, this is a less time-consuming activity and does not require constant monitoring. Upon comparison, the Doc2Vec model generates a shortlist of similar documents.

Two text documents can explain a topic in different ways. For example, a news article can be expressed negatively or positively depending upon the data source. Although our Doc2Vec model trains the documents individually based on the context within a document, it cannot compare the documents based on the context. Thus, we further filter the shortlisted documents by comparing the contexts obtained from our context extraction module (CET). Therefore, apart from the input document, the shortlisted documents are also processed through the CET module before the identification of a similar document. The document that closely matches the context of the input document is then considered as a similar document.

7.5 Data Quality Assessment

The final phase of the TDQA model is quality assessment. As mentioned earlier, unlike in structured data, the accuracy assessment of textual data also requires the context. Therefore, our data quality assessment module uses a similar document obtained from the previous step which is identified based on the context. This section describes the approaches used for the assessment of accuracy and consistency dimensions of data quality.

7.5.1 Data Accuracy

Upon context extraction and accurate dataset search, accuracy assessment in SDQA is a straightforward process by comparing the data values in each record. However, it is not the case with textual data. It is practically impossible to have the same number of words or sentences in any two related documents. Thus, a literal comparison of words or sentences is not feasible for accuracy assessment. In this dissertation, we employ NLP techniques to perform the accuracy assessment. Most of the content in a document depends on the nouns. Since a noun is best described by the adjectives and verbs associated with it, we extract the unique set of objects (adjectives and verbs) related to every noun in both the input and similar documents. We then compare the nouns and their related objects and calculate the accuracy of the input document using equation 7.1.

$$t_{accuracy} = \frac{\sum_{i=1}^{k} \frac{|I_{i_{objects}} \cap S_{i_{objects}}|}{|I_{i_{objects}}|}}{k} \tag{7.1}$$

where,

I = input text document

S = similar text document

i = noun

k = total number of nouns in the input document, I

$i_{objects}$ = all the objects associated with a noun, i

The number of nouns in the input document might differ from that of a similar document. In that case, it is only possible to compare the nouns that are common in both the documents. Therefore, we devise a confidence score for our assessment based on the number of nouns that are compared during the assessment. Equation 7.2 is used to calculate the confidence score of our assessment.

$$\%Confidence = (1 - \frac{|I_{nouns} - S_{nouns}|}{|I_{nouns}|}) * 100 \tag{7.2}$$

where,

I = input text document

S = Similar text document

7.5.2 Data Consistency

The evaluation of data consistency in textual data is similar to that of consistency assessment in SDQA. We use dependency rules to determine the consistency of data. Though the approach is similar, the implementation is quite different. For textual data, we generate the dependency graphs for each sentence in both the input and similar documents. The dependency graphs specify the relationship (such as parent-child) and its type, between all the objects in a sentence. However, since both the documents cannot have the same sentences, we combine all the dependencies of each noun from all instances in respective documents. We then compare the nouns and their dependencies in both the documents to verify if the nouns in the input document have consistent relationships throughout the document. The consistency of the input document is calculated using equation 7.3.

$$t_{consistency} = \frac{\sum_{i=1}^{k} \frac{|I_{i_d} \cap S_{i_d}|}{|I_{i_d}|}}{k} \tag{7.3}$$

where,

I = input text document

S = similar text document

i = noun

k = total number of nouns in the input document, I

i_d = all the dependencies of a noun, i

Similar to accuracy, consistency might also differ based on the number of nouns that are compared during the assessment. Therefore, we use the same equation 7.2 to calculate the confidence score for our consistency assessment.

Finally, to evaluate the overall quality of the input text document, we can use the same methodology used in SDQA. Equation 5.8 in section 5.4.5 is used to calculate the final quality score for text data.

7.6 Experiments and Results

The quality assessment of unstructured data is significantly different from that of structured data. However, the validation of any DQA model irrespective of the data type is similar. Therefore, for the same reasons, as mentioned in section 5.5 for structured data, the direct validation of DQA models for textual data is not possible without labeled datasets.

Nevertheless, examining a DQA model is essential as several important decisions are made based on the quality of the data. For this reason, we aim to test the efficiency of the two main important steps in our TDQA model, viz Similar Document Identification and confidence score. To perform these experiments, it is essential that we use a collection of numerous documents that have correlations with certain documents. Therefore, we use a combination of text documents from All The News [105], and News Aggregator [72] datasets.

Similar Document Identification

The dimensions that we use to assess the quality of textual data in our research require an accurate text document to compare with. Therefore, it is important to identify a closest similar document both in terms of data and its context. Moreover, we need to observe this nature in documents for multiple domains.

There could be numerous documents that are similar to an input document that may or may not have the same context. However, to perform the quality assessment, context matching is

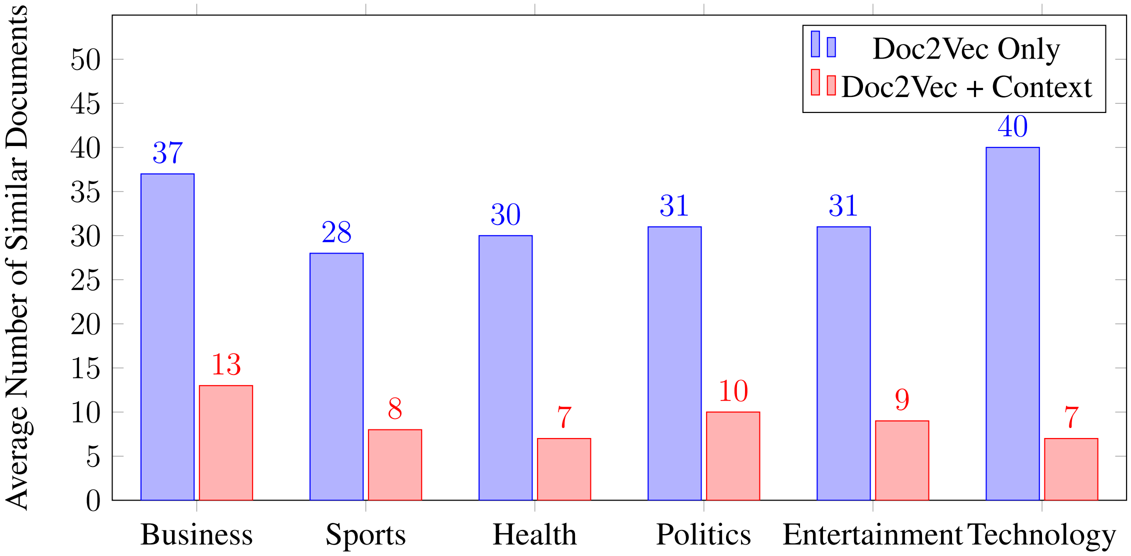

Figure 7.2: Comparison of Similar Document Identification models

crucial. Also, the fewer the number of similar documents we have (with identical context,) the better the quality assessment would be. A simple similarity identification model such as Doc2Vec is useful to obtain a similar document. However, this model does not consider the context of the data. Therefore, we use an enhanced version of this model by combining it with the context extracted from the CET module (chapter VI).

Figure 7.2 shows the importance of the enhanced Doc2Vec model for similar document identification. The graph shows the average number of similar documents obtained to sample input datasets of various fields for Doc2Vec model with and without context. The model with context shows a significant improvement in similar document identification as it considers the context of the data too. In order words, the enhanced model eliminates the chances of choosing a similar document whose context is different from that of the input document.

Since the elimination of irrelevant documents is based on our approach, i.e. context extraction, we intend to determine its efficiency by analyzing the following examples.

To determine whether a document is relevant or not, we use a threshold for relevance and sentiment scores as hyperparameter as this measure varies by each domain. In this experiment, we consider the threshold to be 0.25, i.e. if a document has a difference of relevance and sentiment scores above the threshold, it is considered irrelevant. Otherwise, the respective document is considered relevant and retained in the list of similar documents.

Example 1:

Input document: About most valuable MLB players in 2016

Context:

Topic: Mike Trout; **Relevance:** 0.98; **Sentiment:** 1.0

Topic	Relevance Score	Sentiment Score
Mike Trout	0.95	1.0
Mike Trout	0.92	0.8
Mike Trout	0.88	1.0

Table 7.1: Documents with a similar Context (Example 1)

Topic	Relevance Score	Sentiment Score
Jim Edmonds	0.99	1.0
MVP	0.79	1.0
Mike Trout	0.99	-0.6

Table 7.2: Documents with a different Context (Example 1)

As shown in Table 7.1, all the documents have the same topic. The relevance and sentiment scores are within the threshold; therefore, these documents are considered as relevant documents. On the other hand, Table 7.2 shows irrelevant documents. These documents are considered irrelevant because they have a different context in terms of either topic, relevance, or sentiment scores in the same order of priority. For example, the first two documents have a different topic, and therefore these are considered irrelevant without observing their relevance and sentiment scores. The third document has the same topic as the input document, and the relevance score is close to the input document. However, the sentiment score is far beyond the threshold, and therefore this document is considered irrelevant too.

Example 2:

Input document: About busiest actors of Hollywood in 2014

Context:

Topic: J. K. Simmons; **Relevance:** 0.99; **Sentiment:** 1.0

Topic	Relevance Score	Sentiment Score
J. K. Simmons	0.99	1.0
J. K. Simmons	0.78	0.9
J. K. Simmons	0.94	1.0

Table 7.3: Documents with a similar Context (Example 2)

Topic	Relevance Score	Sentiment Score
Richest	0.90	0.5
Fran Kranz	0.85	1.0
Barefoot	0.95	0.7

Table 7.4: Documents with a different Context (Example 2)

An observation similar to Example 1 is shown in Example 2. Table 7.3 contains the documents with a similar context to that of the input document. Whereas Table 7.4 contains the documents which have a different context although some of the topics have a close association with the input document.

Confidence Score:

As discussed in section 7.5, we use two quality dimensions (accuracy and consistency) to assess the quality of the data. Although the approaches to evaluate these dimensions are

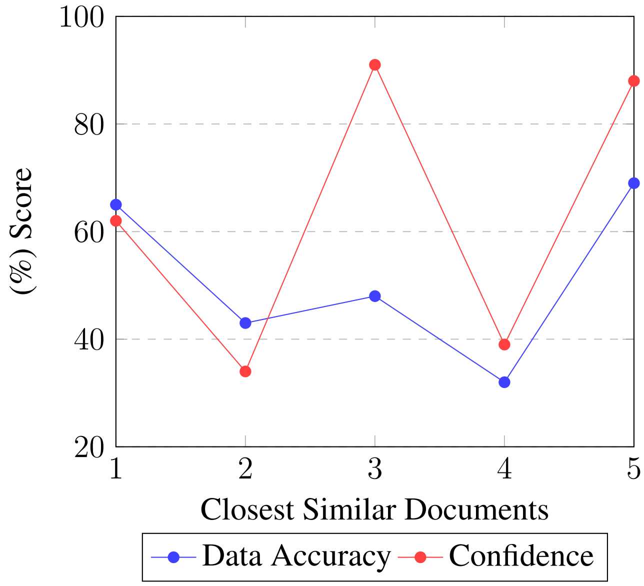

Figure 7.3: Data Accuracy vs Confidence

different, they both have a metric in common, namely, confidence score.

Figure 7.3 shows the data accuracy and confidence scores of a sample input text document when compared with the five closest similar documents. It is important to note that neither of these values decides the quality of the data. The most similar document determines the actual data accuracy of the input document. In this example, document 4 is the most similar, and therefore the accuracy and confidence score associated with that document is what matters, despite the scores are higher when compared to some other documents. In other words, with a 39% confidence rate, the data accuracy of the input document calculated as 32%.

Similar to data accuracy, data consistency is compared with the respective confidence scores in Figure 7.4.

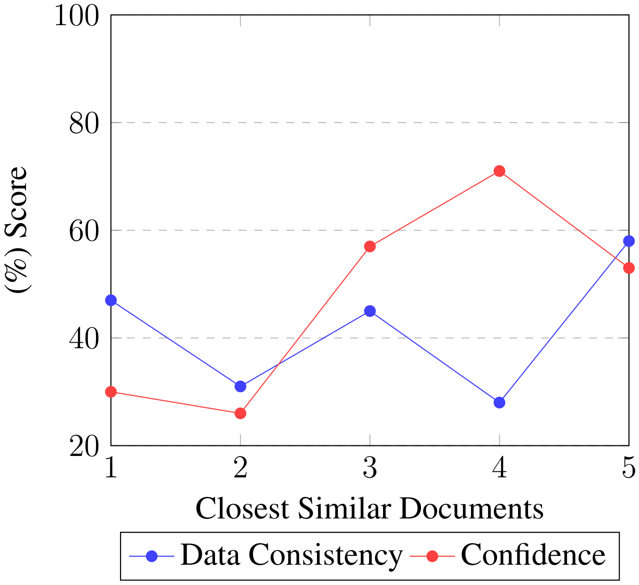

Figure 7.4: Data Consistency vs Confidence

7.7 Summary

Data Quality Assessment is an essential phase of any data analysis task. However, due to the many complications involved in the quality assessment of textual data, there exist no automated DQA models. In this chapter, we present a novel quality assessment model for textual data performed in a couple of dimensions using machine learning.

After the context extraction phase (discussed in chapter VI), the TDQA model performs similar document identification before assessing the quality of the data. The quality assessment is performed in two dimensions, i.e., accuracy and consistency.

It is essential to obtain a similar document to the input document to determine the quality of the data. We use an enhanced version of the Doc2Vec model for similar document identification, which uses a deep learning algorithm in combination with our context extraction model (CET). We use various natural language processing techniques to assess the accuracy and consistency of the textual document. Besides, we use a confidence score metric to determine the usefulness of our quality assessment.

Our experimental results show the efficiency of our enhanced model for similar document identification. In addition, we also present the importance and regulation of confidence scores with respect to data accuracy and consistency.

CHAPTER VIII

CONCLUSION AND FUTURE WORK

Data analysis extracts knowledge and insights from the raw data. However, raw data is not free from errors and therefore cannot be used directly for analysis. Data Quality Assessment (DQA) is an essential application that identifies the data errors and separates the bad quality data from raw data. However, DQA is a challenging task that requires data context and demands a lot of domain knowledge. In this dissertation, we developed a comprehensive quality assessment tool for structured and unstructured data.

As a part of structured data quality assessment (SDQA), we proposed a context extraction framework for structured data (CES) that extracts the context of an input dataset using a collection of feature selection algorithms. We estimated the quality of a dataset in four dimensions; namely, completeness, validity, accuracy, and consistency. Our SDQA models also automate the process of identifying an accurate dataset to compare with the input dataset. This dissertation focuses on the textual type of unstructured data. We proposed a quality assessment model for textual data (TDQA) which contains multiple phases i.e. context extraction, similar document identification, and quality assessment. Using natural language processing and deep neural networks, we developed a context extraction module for textual data (CET) that comprehends the context of an input text based on a topic's relevance and sentiment within a document. We further evaluated the quality of an input text document in accuracy and consistency dimensions of data quality.

Since our quality assessment model is designed to work for multiple domains, we considered datasets from different domains to test the efficiency of our methodologies. Our experimental results show the significance of context-awareness in data quality assessment. We also determine the importance of assessing data quality in multiple dimensions. Manual intervention for DQA is practically impossible for large datasets. Therefore, we developed our automated assessment tools on Hadoop and Spark to support big data.

In addition to our contributions, we observed that the context of structured data changes not only with the change in the purpose of data usage but also with the use of different algorithms for context extraction. Our experimental results in Chapter III show the significance of choosing the best subset for context extraction as the importance of variables varies with different algorithms, which have a direct impact on quality assessment. Besides the context, we observed in Chapter IV that automatic identification closest dataset for quality assessment is essential, as a manual approach is practically impossible and tedious, especially with large datasets. Since the closest dataset has a direct impact on quality assessment, we must choose a reliable and closest dataset possible. Our findings in Chapter V show the changes in quality assessment (in the accuracy and consistency dimensions) for different closest datasets and how important it is to choose the right one.

Although the steps to assess the quality assessment of textual data are no different, i.e., context extraction and quality assessment, the approach is completely different. Our findings in Chapter VI indicate that it is necessary to extract the context from multiple dimensions, creating a fundamental difference between a topic and context. We also observed that a hybrid sentiment analysis model outperforms a lexicon-based approach. In Chapter VII, we observed that an enhanced Doc2Vec model is essential to perform a better quality assessment. Also, we noticed that a confidence score is important to determine the reliability of the quality assessment.

In the future, we aim to enhance our quality assessment models to evaluate other dimensions of data quality, such as timeliness and accessibility. The developed quality assessment models can be extended to support other types of unstructured data, i.e. image, audio, and video. Despite reduction in quality, data obfuscation is an important process that protects sensitive data. We also intend to enhance our assessment tools to identify and keep track of the obfuscation process to provide a more extensive evaluation of the data.

REFERENCES

[1] Netowl - entity extraction, 2017. https://www.netowl.com/2017/08/11/80-worlds-data-unstructured-entity-extraction-must [Last accessed: 09-12-2019].

[2] Apache hadoop, 2019. https://hadoop.apache.org [Last accessed: 09-21-2019].

[3] Apache hadoop, 2019. https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html#Introduction [Last accessed: 09-21-2019].

[4] Apache spark, 2019. https://spark.apache.org/docs/latest/sql-programming-guide.html [Last accessed: 09-21-2019].

[5] Apache spark, 2019. https://spark.apache.org/docs/latest/streaming-programming-guide.html [Last accessed: 09-21-2019].

[6] Naoki Abe, Edwin Peter Dawson Pednault, and Fateh Ali Tipu. Method and apparatus for presenting feature importance in predictive modeling, July 14 2009. US Patent 7,561,158.

[7] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*, 9(12):993–1004, 2016.

[8] Paige H Adams and Craig H Martell. Topic detection and extraction in chat. In *2008 IEEE international conference on Semantic computing*, pages 581–588. IEEE, 2008.

[9] Nitish Aggarwal, Kartik Asooja, and Paul Buitelaar. Deri&upm: Pushing corpus based relatedness to similarity: Shared task system description. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 643–647. Association for Computational Linguistics, 2012.

[10] Samir Al-janabi and Ryszard Janicki. A density-based data cleaning approach for deduplication with data consistency and accuracy. In *2016 SAI Computing Conference (SAI)*, pages 492–501. IEEE, 2016.

[11] Felipe Alonso-Atienza, José Luis Rojo-Álvarez, Alfredo Rosado-Muñoz, Juan J. Vinagre, Arcadi García-Alberola, and Gustavo Camps-Valls. Feature selection using support vector machines and bootstrap methods for ventricular fibrillation detection. *Expert Systems with Applications*, 39(2):1956 – 1967, 2012.

[12] Danilo Ardagna, Cinzia Cappiello, Walter Samá, and Monica Vitali. Context-aware data quality assessment for big data. *Future Generation Computer Systems*, 89:548–562, 2018.

[13] Nicola Askham, Denise Cook, Martin Doyle, Helen Fereday, Mike Gibson, Ulrich Landbeck, Rob Lee, Chris Maynard, Gary Palmer, and Julian Schwarzenbach. The six primary dimensions for data quality assessment. *DAMA UK Working Group*, pages 432–435, 2013.

[14] Hanan Ayad and Mohamed Kamel. Topic discovery from text using aggregation of different clustering methods. In *Conference of the Canadian Society for Computa-*

tional Studies of Intelligence, pages 161–175. Springer, 2002.

[15] Ayoub Bagheri, Mohamad Saraee, and Franciska De Jong. Adm-lda: An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science*, 40(5):621–636, 2014.

[16] Donald P Ballou and Harold L Pazer. Cost/quality tradeoffs for control procedures in information systems. *Omega*, 15(6):509–521, 1987.

[17] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3):16, 2009.

[18] Leopoldo Bertossi and Flavio Rizzolo. Contexts and data quality assessment. *arXiv preprint arXiv:1608.04142*, 2016.

[19] Tony Blakely and Clare Salmond. Probabilistic record linkage and a method to calculate the positive predictive value. *International journal of epidemiology*, 31(6):1246–1252, 2002.

[20] Tobias Bleifuß, Susanne Bülow, Johannes Frohnhofen, Julian Risch, Georg Wiese, Sebastian Kruse, Thorsten Papenbrock, and Felix Naumann. Approximate discovery of functional dependencies for large datasets. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1803–1812. ACM, 2016.

[21] Philip Bohannon, Wenfei Fan, Michael Flaster, and Rajeev Rastogi. A cost-based model and effective heuristic for repairing constraints by value modification. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 143–154. ACM, 2005.

[22] Erik Boiy and Marie-Francine Moens. A machine learning approach to sentiment analysis in multilingual web texts. *Information retrieval*, 12(5):526–558, 2009.

[23] Freddie Bray and D Max Parkin. Evaluation of data quality in the cancer registry: principles and methods. part i: comparability, validity and timeliness. *European journal of cancer*, 45(5):747–755, 2009.

[24] M Kathryn Brohman, Richard T Watson, Gabriele Piccoli, and A Parasurama. Data completeness: a key to effective net-based customer service systems. *Communications of the ACM*, 46(6):47–51, 2003.

[25] Khoo Khyou Bun and Mitsuru Ishizuka. Topic extraction from news archive using tf* pdf algorithm. In *Proceedings of the Third International Conference on Web Information Systems Engineering, 2002. WISE 2002.*, pages 73–82. IEEE, 2002.

[26] Erik Cambria, Dipankar Das, Sivaji Bandyopadhyay, and Antonio Feraco. *A practical guide to sentiment analysis*. Springer, 2017.

[27] Loredana Caruccio, Vincenzo Deufemia, and Giuseppe Polese. A genetic algorithm to discover relaxed functional dependencies from data. In *SEBD*, page 146, 2017.

[28] Peter Christen. Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 151–159. ACM, 2008.

[29] Xu Chu, Ihab F Ilyas, and Paolo Papotti. Holistic data cleaning: Putting violations into context. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 458–469. IEEE, 2013.

[30] William W Cohen and Jacob Richman. Learning to match and cluster large high-

dimensional data sets for data integration. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 475–480. ACM, 2002.

[31] Gao Cong, Wenfei Fan, Floris Geerts, Xibei Jia, and Shuai Ma. Improving data quality: Consistency and accuracy. In *Proceedings of the 33rd international conference on Very large data bases*, pages 315–326. VLDB Endowment, 2007.

[32] Gao Cong, Wenfei Fan, Floris Geerts, Xibei Jia, and Shuai Ma. Improving data quality: Consistency and accuracy. In *Proceedings of the 33rd international conference on Very large data bases*, pages 315–326. VLDB Endowment, 2007.

[33] Russell G Congalton and Kass Green. *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press, 2008.

[34] Thomas M Cover and Joy A Thomas. Entropy, relative entropy and mutual information. *Elements of information theory*, 2:1–55, 1991.

[35] R Crone. Big data veracity assessment: Improving risk assessment by adding high veracity data to existing contents insurance models. 2016.

[36] John Gantz David Reinsel and John Rydning. Idc - the digitization of the world from edge to core, 2018. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf [Last accessed: 09-17-2019].

[37] Jonathan de Bruin. Python record linkage toolkit, 2018. https://recordlinkage.readthedocs.io/en/latest/about.html [Last accessed: 11-28-2018].

[38] Robert Demolombe. Answers about validity and completeness of data: formal defi-

nitions, usefulness and computation technique. In *International Conference on Flexible Query Answering Systems*, pages 138–147. Springer, 1998.

[39] Xiaoou Ding, Hongzhi Wang, Jiaxuan Su, Jianzhong Li, and Hong Gao. Improve3c: Data cleaning on consistency and completeness with currency. *arXiv preprint arXiv:1808.00024*, 2018.

[40] February EPA. Epa guidance for quality assurance project plans. Technical report, EPA QA/G-5, EPA 600/R-98/018, 1998.

[41] Kayode Philip Fadahunsi, James Tosin Akinlua, Siobhan O'Connor, Petra A Wark, Joseph Gallagher, Christopher Carroll, Azeem Majeed, and John O'Donoghue. Protocol for a systematic review and qualitative synthesis of information quality frameworks in ehealth. *BMJ open*, 9(3):e024722, 2019.

[42] Wenfei Fan. Dependencies revisited for improving data quality. In *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 159–170. ACM, 2008.

[43] Wenfei Fan, Floris Geerts, Jianzhong Li, and Ming Xiong. Discovering conditional functional dependencies. *IEEE Transactions on Knowledge and Data Engineering*, 23(5):683–698, 2010.

[44] Filomena Ferrucci, M Kechadi, Pasquale Salza, Federica Sarro, et al. A framework for genetic algorithms based on hadoop. *arXiv preprint arXiv:1312.0086*, 2013.

[45] N. C. for Environmental Information. Climate data online, 2018. `https://www.noaa.gov/` [Last accessed: 11-28-2018].

[46] Christopher Fox, Anany Levitin, and Thomas Redman. The notion of data and its quality dimensions. *Information processing & management*, 30(1):9–19, 1994.

[47] Ted Friedman and Saul Judah. The state of data quality: Current practices and evolving trends, 2013. https://www.gartner.com/en/documents/2636315 [Last accessed: 09-17-2019].

[48] Google. Word2vec, 2013. https://code.google.com/archive/p/word2vec/ [Last accessed: 11-28-2018].

[49] Luis Gravano, Panagiotis G Ipeirotis, Nick Koudas, and Divesh Srivastava. Text joins for data cleansing and integration in an rdbms. In *Proceedings 19th International Conference on Data Engineering (Cat. No. 03CH37405)*, pages 729–731. IEEE, 2003.

[50] Derek Greene and Pádraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 377–384, 2006.

[51] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.

[52] Joseph M Hellerstein. Quantitative data cleaning for large databases. *United Nations Economic Commission for Europe (UNECE)*, 2008.

[53] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the first workshop on social media analytics*, pages 80–88. acm, 2010.

[54] Ihab F Ilyas, Xu Chu, et al. Trends in cleaning relational data: Consistency and deduplication. *Foundations and Trends® in Databases*, 5(4):281–393, 2015.

[55] Aminul Islam and Diana Inkpen. Correcting different types of errors in texts. In *Canadian Conference on Artificial Intelligence*, pages 192–203. Springer, 2011.

[56] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[57] Dhivya Jayaraman. N-gram based keyword topic modelling for canadian longitudinal study on aging survey data. 2018.

[58] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

[59] Beverly K Kahn, Diane M Strong, and Richard Y Wang. Information quality benchmarks: product and service performance. *Communications of the ACM*, 45(4):184–192, 2002.

[60] Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia. *Learning spark: lightning-fast big data analysis*. " O'Reilly Media, Inc.", 2015.

[61] Cornelia Kiefer. Assessing the quality of unstructured data: An initial overview. In *LWDA*, pages 62–73, 2016.

[62] Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, and Doheon Lee. A taxonomy of dirty data. *Data mining and knowledge discovery*, 7(1):81–99, 2003.

[63] Jyrki Kivinen and Heikki Mannila. Approximate inference of functional dependencies from relations. *Theoretical Computer Science*, 149(1):129–149, 1995.

[64] K Krasnow Waterman and Paula J Bruening. Big data analytics: risks and responsibilities. *International Data Privacy Law*, 4(2):89–95, 2014.

[65] Sebastian Kruse and Felix Naumann. Efficient discovery of approximate dependencies. *Proceedings of the VLDB Endowment*, 11(7):759–772, 2018.

[66] Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

[67] Jacek Laskowski. Mastering apache spark, 2019. `https://jaceklaskowski.gitbooks.io/mastering-apache-spark/spark-overview.html` [Last accessed: 09-21-2019].

[68] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[69] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary. Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 251–258. IEEE, 2011.

[70] Sungjick Lee and Han-joon Kim. News keyword extraction for topic tracking. In *2008 Fourth International Conference on Networked Computing and Advanced Information Management*, volume 2, pages 554–559. IEEE, 2008.

[71] Yuhua Li, David McLean, Zuhair A Bandar, James D O'shea, and Keeley Crockett. Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering*, 18(8):1138–1150, 2006.

[72] M. Lichman. News aggregator data set - uci machine learning repository, 2013. `http://archive.ics.uci.edu/ml/datasets/News+Aggregator` [Last accessed: 02-15-2020].

[73] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[74] David Loshin. *The practitioner's guide to data quality improvement*. Elsevier, 2010.

[75] Aida Malaki. *Multidimensional contexts for data quality assessment*. PhD thesis, Carleton University, 2013.

[76] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning generic context embedding with bidirectional lstm. In *Proceedings of the 20th SIGNLL conference on computational natural language learning*, pages 51–61, 2016.

[77] Jorge Merino, Ismael Caballero, Bibiano Rivas, Manuel Serrano, and Mario Piattini. A data quality in use model for big data. *Future Generation Computer Systems*, 63:123–130, 2016.

[78] MicroStrategy. Microstrategy analytics and mobility - 2018 global state of enterprise analytics report, 2018. `https://www.microstrategy.com/getmedia/50ea9c13-feb7-4b9a-b976-8b04ca39abb2/Global-State-of-Enterprise-Analytics-Report-MicroStrategy_2018` [Last accessed: 09-12-2019].

[79] Apra Mishra and Santosh Vishwakarma. Analysis of tf-idf model and its variant for document retrieval. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 772–776. IEEE, 2015.

[80] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

[81] Sergey I Nikolenko, Sergei Koltcov, and Olessia Koltsova. Topic modelling for qualitative studies. *Journal of Information Science*, 43(1):88–102, 2017.

[82] Jack E Olson. *Data quality: the accuracy dimension*. Elsevier, 2003.

[83] Aytug Onan, Serdar Korukoglu, and Hasan Bulut. Lda-based topic modelling in text sentiment classification: An empirical analysis. *Int. J. Comput. Linguistics Appl.*, 7(1):101–119, 2016.

[84] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[85] N Puttkammer, Janet G Baseman, Emily Beth Devine, JS Valles, Nathaelf Hyppolite, France Garilus, Jean-Guy Honoré, Alastair I Matheson, S Zeliadt, Krista Yuhas, et al. An assessment of data quality in a multi-site electronic medical record system in haiti. *International journal of medical informatics*, 86:104–116, 2016.

[86] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[87] Thomas C Redman. *Data driven: profiting from your most important business asset*. Harvard Business Press, 2008.

[88] Andrew P Reimer, Alex Milinovich, and Elizabeth A Madigan. Data quality assessment framework to assess electronic medical record data for use in research. *International journal of medical informatics*, 90:40–47, 2016.

[89] Michael B Richman, Theodore B Trafalis, and Indra Adrianto. Missing data imputation through machine learning algorithms. In *Artificial intelligence methods in the environmental sciences*, pages 153–169. Springer, 2009.

[90] LL Roos and A Wajda. Record linkage strategies. *Methods of information in medicine*, 30(02):117–123, 1991.

[91] Kai-Uwe Sattler. Data quality dimensions. *Encyclopedia of Database Systems*, pages 612–615, 2009.

[92] Monica Scannapieco. *Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications*. Springer, 2006.

[93] George Seif. 3 simple ways to handle large data with pandas, 2019. `https://towardsdatascience.com/3-simple-ways-to-handle-large-data-with-pandas-d9164a3c02c1` [Last accessed: 10-08-2019].

[94] Valerie Sessions and Marco Valtorta. Towards a method for data accuracy assessment utilizing a bayesian network learning algorithm. *Journal of Data and Information Quality (JDIQ)*, 1(3):14, 2009.

[95] Lisa Sokol and Steve Chan. Context-based analytics in a big data world: better decisions. *IBM RedBooks Point-of-View Publication*, 2013.

[96] Marina Soley-Bori. Dealing with missing data: Key assumptions and methods for applied analysis. *Boston University*, 23, 2013.

[97] Daniel Sonntag. Assessing the quality of natural language text data. In *GI Jahrestagung (1)*, pages 259–263, 2004.

[98] JS Sowmiya and S Chandrakala. Joint sentiment/topic extraction from text. In *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, pages 611–615. IEEE, 2014.

[99] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[100] Caroline Sporleder, Marieke Van Erp, Tijn Porcelijn, and Antal Van Den Bosch. Spotting the ‘odd-one-out’: Data-driven error detection and correction in textual

databases. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*, 2006.

[101] Stackla. Bridging the gap - consumer and marketing perspectives on content in the digital age, 2019. `https://stackla.com/resources/reports/bridging-the-gap-consumer-marketing-perspectives-on-content-in-the-` [Last accessed: 09-12-2019].

[102] Daren S Starnes, Dan Yates, and David S Moore. *The practice of statistics*. Macmillan, 2010.

[103] Besiki Stvilia, Les Gasser, Michael B Twidale, and Linda C Smith. A framework for information quality assessment. *Journal of the American society for information science and technology*, 58(12):1720–1733, 2007.

[104] Ikbal Taleb, Mohamed Adel Serhani, and Rachida Dssouli. Big data quality assessment model for unstructured data. In *2018 International Conference on Innovations in Information Technology (IIT)*, pages 69–74. IEEE, 2018.

[105] Andrew Thompson. All the news, 2017. `https://components.one/datasets/all-the-news-2-news-articles-dataset/` [Last accessed: 03-17-2020].

[106] Konstantinos Vassakis, Emmanuel Petrakis, and Ioannis Kopanakis. Big data analytics: applications, prospects and challenges. In *Mobile Big Data*, pages 3–20. Springer, 2018.

[107] Vassilios S Verykios, Ahmed K Elmagarmid, and Elias N Houstis. Automating the approximate record-matching process. *Information sciences*, 126(1-4):83–98, 2000.

[108] Chong Wang and David M Blei. Collaborative topic modeling for recommending

scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM, 2011.

[109] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[110] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[111] D Randall Wilson. Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage. In *The 2011 International Joint Conference on Neural Networks*, pages 9–14. IEEE, 2011.

[112] Philip Mark Woodall, Martin Oberhofer, and Alexander Borek. A classification of data quality assessment and improvement methods. Inderscience, 2014.

[113] Bihan Wu, Gang Wu, and Mengdong Yang. A mapreduce based ant colony optimization approach to combinatorial optimization problems. In *Natural Computation (ICNC), 2012 Eighth International Conference on*, pages 728–732. IEEE, 2012.

[114] Shaomin Wu. A review on coarse warranty data and analysis. *Reliability Engineering & System Safety*, 114:1–11, 2013.

[115] Catharine Wyss, Chris Giannella, and Edward Robertson. Fastfds: A heuristic-driven, depth-first algorithm for mining functional dependencies from relation instances extended abstract. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 101–110. Springer, 2001.

[116] Jianping Zeng, Chengrong Wu, and Wei Wang. Multi-grain hierarchical topic extraction algorithm for text mining. *Expert Systems with Applications*, 37(4):3202–3208, 2010.

[117] Xiaoyan Zhang and Ting Wang. Topic tracking with dynamic topic model and topic-based weighting method. *Journal of Software*, 5(5):482–489, 2010.

[118] Jinling Zhou, Xinchun Diao, and Jianjun Cao. Holistic data accuracy assessment using search & scored-based bayesian network learning algorithms. In *2017 3rd International Conference on Information Management (ICIM)*, pages 432–436. IEEE, 2017.

VITA

Sesha Sai Goutam Sarma Mylavarapu

Candidate for the Degree of

Doctor of Philosophy

Dissertation: CONTEXT-AWARE QUALITY ASSESSMENT OF STRUCTURED AND UNSTRUCTURED DATA

Major Field: Computer Science

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy with a major in Computer Science at Oklahoma State University, Stillwater, Oklahoma in July, 2020.

Completed the requirements for the Master of Science with a major in Computer Science at Oklahoma State University, Stillwater, Oklahoma in May, 2015.

Completed the requirements for the Bachelor of Technology with a major in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad, India in May, 2012.

Experience:

Graduate Teaching Assistant/Associate, Oklahoma State University, 2016 - 2020