DataLDA: Analyzing Data Quality in Big Data Using Traditional Approaches vs. Latent Dirichlet Allocation - Systematic Review

Sayed Abu Sayeed , Naresh Kshetri * , Mir Mehedi Rahman , Samiul Alam , Abdur Rahman

Posted Date: 31 October 2024

doi: 10.20944/preprints202410.2596.v1

Keywords: big data; data quality; topic modeling; latent dirichlet allocation (lda); systematic literature review

Article

dataLDA: Analyzing Data Quality in Big Data Using Traditional Approaches vs. Latent Dirichlet Allocation—Systematic Review †

Sayed Abu Sayeed [1], Samiul Alam [2], Naresh Kshetri [3], Abdur Rahman [4] and Mir Mehedi Rahman [4]

[1] College of Business, Florida Atlantic University, Boca Raton, FL, USA
[2] School of Computing, University of South Alabama, Mobile, AL, USA
[3] Department of Cybersecurity, Rochester Institute of Technology, Rochester, NY, USA
[4] School of Business & Technology, Emporia State University, Emporia, KS, USA
* Correspondence: ssayeed2024@fau.edu
† This paper is accepted at 13th IEEE ISDFS 2025.

Abstract: Big data quality is all about ensuring that the huge amounts of data are accurate, reliable, and useful. It is important because decisions made based on this data can affect many things, like business strategies, customer experiences, and even public policies. Many researchers are studying data quality in the context of big data. This study aims to analyze the research trends and patterns within the domain of big data quality. This method reveals hidden themes in large datasets, helping to understand data quality better. The findings show that Latent Dirichlet Allocation (LDA) is effective for topic modeling in big data, highlighting data quality challenges and opportunities. The research identified key trends and effective assessment methods. For researchers, using LDA in a systematic review helps summarize existing studies and find gaps, guiding future research. Practitioners and future researchers get useful direction for managing data quality and improving efficiency and decisions in various industries.

Keywords: big data; data quality; topic modeling; latent dirichlet allocation (lda); systematic literature review

I. Introduction

In the modern digital age, the utilization of big data takes an important position in facilitating decision-making procedures and influencing many facets of both the commercial realm and society at large. Research by IDC says that the amount of data worldwide will increase significantly from 45 zettabytes in 2019 to 175 zettabytes by 2025 [1]. The 5V's qualities, namely volume, velocity, variety, veracity, and value, are utilized in big data [2]. Table 1 discussed about the characteristics of 5V's.

Table 1. Characteristics Of Big Data.

Volume	Refers to the vast amount of data being generated and collected.
Velocity	Represents the speed at which data is being produced and processed.
Variety	Indicates the diverse types of data, including structured, unstructured, and semi-structured data.
Veracity	Refers to the reliability and accuracy of the data.

 doi:10.20944/preprints202410.2596.v1

Value	Represents the potential insights and benefits that can be derived from analyzing the data.

Due to the quantity and rapidity of data, along with its diverse range of types, the quality of the data is substandard.

In the middle of the great quantity and speed at which data is produced, guaranteeing its quality has become a notable concern. Sometimes, threats like cryptojacking and ransomware also challenge data security, impacting the reliability of big data [3]. The presence of low data quality might result in incorrect conclusions and choices, underscoring the significance of evaluating and upholding data quality, particularly within the field of big data [4].

Despite its importance, there remains a significant gap in systematic methods to identify and analyze research trends and patterns in big data quality. Recent studies focus on improving data quality in big data. Many articles and research papers from databases like Google Scholar, IEEE Xplore, and ACM Digital Library show these efforts. Searches for "data quality in big data" and "research on big data quality" reveal many studies on this topic. New reviews give insights into current knowledge, highlighting ongoing efforts to tackle challenges and opportunities in big data quality.

Researchers often do systematic literature reviews (also known as systematic literature surveys) to keep up with the newest developments in a specific topic [5]. The process consists of three main phases: planning, conducting, and reporting [6]. Many systematic literature reviews have been conducted on the topic of data quality in the realm of big data. However, this manual procedure is difficult and can only identify and examine a restricted quantity of publications [7]. Topic modeling is a widely utilized analytical technique for the assessment of data. Latent Dirichlet Allocation (LDA) has been widely employed by many researchers because of its remarkable topic modeling capabilities. LDA is a powerful statistical tool which is an unsupervised machine learning method. It can be used to find the main topics hidden within a large batch of articles [8]. This research aims to close the gap using LDA's strengths. It is guided by the following research questions:

RQ1: What trends are most prevalent within the domain of big data quality over the years 2010-2024?

RQ2: How do the most prevalent themes change across each year?

RQ3: How does topic modeling, specifically LDA, compare to traditional literature review methods in identifying research trends in big data quality?

The objective of this research is to review past publications and stay updated with current research trends in big data. This research will help in understanding big data quality by reviewing current trends and methods. Using LDA for topic modeling identifies research gaps and future opportunities. Practitioners will gain insights into effective data quality assessment methods. The review will be a great resource for academics and professionals aiming to improve data quality and leverage big data analytics.

II. Literature Review

We explored trends in big data quality using both traditional literature reviews and Latent Dirichlet Allocation (LDA). Traditional reviews gave us a deep analysis of key theories and concepts. LDA helps us find hidden themes in many academic papers. Using both methods together provided a complete understanding of the research landscape. In this study, we manually reviewed a subset of the 683 articles from the Web of Science database. We focused on key themes like data quality dimensions, new methods, and how big data quality is applied in different sectors. Most of the published review papers focus on a specific area of Big Data quality [9].

Elouataoui et al. (2022) examined big data quality and the challenges and solutions for ensuring it. They reviewed existing research on big data quality, Which includes its characteristics, value chain, and quality measures. The paper emphasized that high-quality data is crucial for reliable analytics and decision-making, while poor-quality data leads to errors. They discussed the complexities of big

 doi:10.20944/preprints202410.2596.v1

data, like its large size, fast speed, and variety, and how these make it hard to keep data accurate and complete. The authors reviewed methods to improve data quality, such as data cleaning and validation, and highlighted the need for specific strategies in fields like healthcare, social media, and finance [9].

Ridzuan et al. (2020) looked at data quality in the big data era by reviewing papers published from 2016 to 2020 from sources like the ACM Digital Library, Scopus, and ScienceDirect. They chose 21 papers to examine closely. The authors talked about key aspects of data quality, such as accuracy, completeness, consistency, and timeliness, which are important for making big data useful and reliable. They pointed out that poor data quality can lead to wrong conclusions, so it's crucial for organizations to have strong data management and quality assurance practices. The paper also discussed the challenges industries face in managing data quality, like integrating different data sources and dealing with the fast pace of data generation [10].

Ijab et al. (2019) introduced the Big Data Quality Framework, which includes aspects like Accessibility, Timeliness, Authorization, Credibility, Clarity, Accuracy, Authenticity, Integrity, Consistency, Completeness, Auditability, Fitness for Use, Readability, and Structure. The paper looked at the importance of data quality in Malaysia's public sector. As Malaysia used big data to improve public services, there was a need for a strong data quality framework. The authors created a new framework through a systematic literature review (SLR). This framework aimed to help share high-quality data across government sectors, increase transparency, and encourage innovation. By focusing on data quality from the start, the framework aimed to reduce the time and effort needed to fix data errors, supporting better decision-making in the public sector. The authors reviewed databases like ScienceDirect, IEEE Xplore, Scopus, and AIS Electronic Library from 2000 to 2018 and selected 295 papers, of which they analyzed 16 [11].

Mostafa Mirzaie et al. conducted a systematic literature review on big data quality. They explored three research areas: processing types, active researchers and institutions, and challenges. The study emphasized that the effectiveness of data-driven insights heavily relies on the quality of the data itself. Despite the growing acknowledgment of data quality in big data studies, comprehensive reviews in this area have been limited. We did a systematic literature review to map out the landscape of big data quality research, proposing a research framework that categorizes existing studies based on processing types, tasks, and techniques. They identified and discussed major challenges in big data quality, such as handling vast volumes of data from diverse sources at high velocities, and suggested future directions for research. After selecting 419 studies, they narrowed their focus to 170 for in-depth examination, ultimately analyzing 16 of them [12].

Liu et al. highlighted challenges in handling big data, including issues like inaccurate data collection, incomplete information, and ethical concerns. Their study, based on empirical evidence, underscores the importance of using rigorous scientific methods to ensure reliable findings and address significant data issues. The paper reviewed seven important research papers that explore various types of big data and related research challenges, particularly in geographic and spatial studies. While big data can provide valuable insights, it also poses challenges such as data errors, missing information, inadequate representation, and privacy issues. The authors urge researchers to approach big data use with care to ensure accuracy and respect privacy concerns [13]

Ji et al. (2020) searched key databases and found 83 studies on quality assurance (QA) technologies for big data. They noted a lack of comprehensive reviews on QA methods for large data applications. The article discusses key findings on quality attributes in big data applications and divides QA technologies into two groups: implementation-specific and process-specific. Implementation-specific QA technologies include methods like architectural choices and fault tolerance to prevent faults. Process-specific QA technologies, such as testing and monitoring, ensure applications work correctly throughout their lifecycle. The review highlights important quality attributes affected by big data, like correctness, performance, and reliability. It emphasizes the need for a combined QA approach since no single method can address all issues. The authors call for more research to create new solutions for big data challenges and to test these technologies in real-world situations [14].

 doi:10.20944/preprints202410.2596.v1

Ramasamy & Chowdhury (2020) analyzed 17 research articles to compile a list of big data quality dimensions. They emphasized the importance of traditional factors like accuracy and newer ones like trust, especially for social media and sensor data. The study highlights challenges in managing data quality in big data due to its volume, speed, and variety. It calls for new methods and tools, including real-time analysis and ensuring data confidentiality and credibility, to improve decision-making reliability [15].

These reviews underscore the increasing significance of big data quality research. However, a notable challenge persists: there is a gap in research on the application of LDA in the realm of Big Data Quality. Researchers are leveraging LDA to uncover current research trends across various domains like Blockchain, Cybersecurity, and Machine Learning through topic modeling. LDA facilitates the efficient analysis of vast datasets. As the field of big data quality research has expanded substantially, techniques such as LDA have become increasingly valuable for extracting topics from the extensive documents used in such research. For instance, researchers have employed the LDA technique to predict the future directions of research in blockchain technology. They have identified 17 emerging research trends that are now being implemented and warrant further attention [15].

Also, for LDA's accuracy in topic modeling heavily relies on the quality of the input data. To enhance data validation and integrity, blockchain technology, with its decentralized and immutable properties, provides tamper-proof validation, significantly strengthening the reliability of data used in these analytical outcomes. Its proven effectiveness in securing data across sectors such as healthcare, academia, and defense highlights blockchain's potential to ensure data integrity in big data analytics [16].

Maintaining data quality in big data environments requires a precise and actionable approach to risk assessment. An effective risk assessment method, like AssessITS, provides clear and targeted methodologies for identifying and mitigating data integrity risks, which are crucial for enhancing data reliability. By combining theoretical guidelines with practical, step-by-step processes, this approach helps organizations efficiently address vulnerabilities, reducing the likelihood of data inaccuracies. Such a comprehensive strategy not only supports traditional data quality methods but also strengthens advanced techniques like Latent Dirichlet Allocation (LDA) by ensuring that data integrity is consistently upheld, leading to more accurate and insightful data analysis [17].

III. Big Data And Its Dimensions

The importance of data quality in decision-making has been critical since the advent of the digital age. In 1963, the expression "garbage in, garbage out" was introduced to emphasize its significance in the field of computing. Over time, companies have employed data governance initiatives and tools such as data dictionaries to effectively oversee and regulate data quality. Data warehouses, which were first introduced in the 1980s, presented novel difficulties by consolidating data from many sources, resulting in problems such as the absence or incongruity of data. The advent of the World Wide Web in the 1990s increased the accessibility of data; however, it also presented difficulties due to the absence of data semantics. Initiatives such as the Semantic Web and Schema.org seek to tackle these difficulties by enhancing the compatibility of web data for queries and calculations. As we enter the age of big data—characterized by its velocity, variety, and volume—the fourth and fifth 'Vs'—veracity and value—become increasingly important in guaranteeing the dependability of data and extracting meaningful insights. Strict data quality management is necessary to avoid mistakes that can greatly affect choices and operations, and veracity, or the reliability of data, addresses problems with consistency and quality. Value, on the other hand, highlights the data's prospective insights and advantages, encouraging firms to tackle data quality head-on. An effective strategy for preserving data quality is critical since seemingly little errors can have far-reaching effects, much like the butterfly effect [18]. This does double duty: it protects big data from harm and lets businesses use massive datasets for efficiency, strategy, and regulatory compliance. In other words, it turns data-driven challenges into opportunities. According to IBM, the assessment of data quality pertains to the extent to which a dataset satisfies specific standards in terms of accuracy, comprehensiveness, validity, consistency, uniqueness, timeliness, and suitability for a particular purpose. This aspect

 doi:10.20944/preprints202410.2596.v1

holds significant importance in the context of data governance endeavors inside an organization [19]. According to ISO/IEC, Data Product Quality can be defined as how well the data meets the needs specified by the organization that owns the product [20]. Essentially, these needs are represented in the Data Quality model via various attributes (such as Accuracy, Completeness, Consistency, Credibility, Timeliness, and Accessibility).

Data changes for many reasons, and to improve its quality, it's important to identify and organize these reasons into different categories of data quality dimensions [21]. Data quality dimensions (DQDs) can be described as "a set of data quality attributes that represent a single aspect or construct of data quality" [15]. Common data quality dimensions include Accuracy, Completeness, Consistency, Freshness, Uniqueness, and Validity, with literature also suggesting additional aspects like security, traceability, navigation, and data decay [9]. In the study by Lee et al. (2006), the researchers identified several key dimensions for evaluating data quality (Calvert, 2007). Free of Error, Completeness, Consistency, Credibility, Timeliness, and Accessibility are some of the dimensions that offer a comprehensive framework for assessing data quality, highlighting aspects crucial for organizations. Additionally, Lee et al. propose metrics to quantify these dimensions, further aiding in their practical application. In the context of big data, the challenges related to data quality (DQ) have become ever more complicated. The large quantity of data (volume) poses challenges in ensuring correctness, while the rapid rate at which data is received might undermine its timeliness.

IV. Topic Modelling

Topic modeling is an effective text mining technique derived from Natural Language Processing (NLP) that investigates the relationship between the collected data and the documents. Data mining is an emerging field that aims to extract information from unstructured data [22].

Statistical models are used to find relationships between words and texts within a collection of unorganized textual material. This allows the identification of significant themes or subjects within the data. Natural language processing makes extensive use of topic modeling approaches, especially those based on Latent Dirichlet Allocation (LDA), to sort through unstructured materials for subjects and extract meaning. All sorts of fields can benefit from these techniques, including text mining, data analysis for social media, and information retrieval. The use of LDA-based topic modeling in social media networks, such as Twitter and Facebook, helps to understand online discussions and interactions and allows for the discovery of useful insights and patterns. In addition to their use in social media, topic models have many other applications in areas such as software engineering, geography, politics, and medicine by uncovering the semantics and structures hidden in large datasets [23]. In LDA, a corpus is viewed as a generative probabilistic model. It was introduced by Blei et al. in 2003. Essentially, documents are considered as mixtures of hidden topics, where each topic is defined by a distribution of words. LDA. It is a popular method in topic modeling that represents topics through word probabilities. The words with the highest probabilities in each topic typically provide insight into the topic's content. LDA, being an unsupervised probabilistic approach, assumes that each document can be represented by a distribution of latent topics, with all documents sharing a common Dirichlet prior. Similarly, each latent topic is represented by a distribution of words, with the word distributions also sharing a common Dirichlet prior [24].

Table 2. Steps Involved In LDA.

Step	Description
1	Identify individual words as features in NLP.
2	Utilize Latent Dirichlet Allocation (LDA) method to establish connections between documents.

 doi:10.20944/preprints202410.2596.v1

Step	Description
3	Employ Variational Expectation Maximization (VEM) algorithm to estimate similarities within the corpus.
4	Utilize Bag of Words (BoW) to extract initial few words for LDA model.
5	Represent each document in the corpus as a distribution of topics.
6	Extract topics representing distributions of words.
7	Determine probabilistic distribution of topics for each document through LDA.
8	Obtain comprehensive understanding of relationships between topics.

In natural language processing (NLP), using individual words as features helps find important information quickly. It helps to avoid the need to analyze the entire dataset. The Latent Dirichlet Allocation (LDA) method establishes statistical and visual connections between items. LDA uses the Variational Expectation Maximization (VEM) algorithm to estimate similarities within the corpus. However, LDA often lacks semantic details, so only the first few words from the Bag of Words (BoW) are used. Each document represents a distribution of topics, and each topic represents a distribution of words. LDA reveals the topic distribution in each document, aiding in understanding the relationships between them [22].

V. Methodology

In this section, we included all the steps and activities relating our experiment to data quality in the big data domain. First, we talked about where we got our data from and why we chose a specific database called Web of Science. Then, we describe how we cleaned up the data to make it easier to analyze using a computer program called Python. After that, we introduce LDA. We also talked about how we figured out the best number of topics to look for in our data. Finally, we explain why it's important to give labels to the topics we find and how it helps us understand them better. This part helps you understand how we did our research and why each step was important.

A. Data Collection

In this study, the researchers examined the Web of Science database, known for its extensive research coverage. The Web of Science database contains research articles from various sources such as journals, conferences, books, and more. While other databases exist, including free options like Microsoft Academic and specialized ones like PubMed, the reliability of these alternatives may vary. For a comprehensive analysis and research evaluation, the Web of Science is still considered the most reliable source [23]. The search string designed for this study was (“Big data” AND “data quality”). The terms "big data" and "data quality" were chosen as the focus of this research due to their foundational significance. "Big data" creates challenges that old data management methods can't handle. "Data quality" is crucial for making reliable decisions. Errors in data can cause big problems in business, healthcare, and other fields. The inclusion criteria specify that only research papers published in English and categorized as articles are considered.

We selected the topic category in the search box within the Web of Science databases. In this category, it searches the search string in the title, abstract, and author keywords. To capture the research trends, we considered all articles published related to the string, resulting in a total of 683

 doi:10.20944/preprints202410.2596.v1

papers out of 456 different sources of research articles (as of April 19, 2024). IEEE Access is the primary source of research articles on data quality in big data, housing 17 articles. Following that is Sensors with nine articles, and then Applied Sciences Basel with eight articles. We have these results in an Excel file, with each column containing different data.

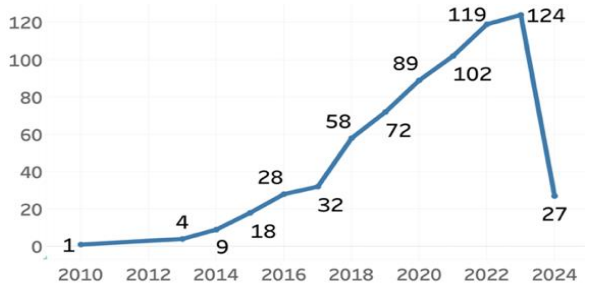

Figure 1. Publication Year.

B. Data Processing

The second step is processing the collected documents. During preprocessing, the aim is to eliminate any unnecessary information present in the material. This process eliminates unwanted words and characters from the gathered corpus or dataset, improving its quality and accuracy for further processing. All these steps are conducted using the Python programming language.

The code imports some useful libraries that provide tools for text processing and topic modeling. These libraries include pandas (for data manipulation) and nltk (for natural language processing). The preprocess text(text) function takes input from a piece of text (like an article title or abstract). This function loads a list of common English words called "stop words" (e.g., "the," "and," "in") that we want to ignore. It uses a lemmatizer to convert words to their base form (e.g., "running" becomes "run") and tokenizes the text (splits it into individual words). It also preprocesses text data by filtering out non-alphabetic and non-numeric words, removing short words and stop words, and ultimately returning a list of cleaned-up words. Afterward, the code tried to load data from the Excel file. The code combines the "Article Title" and "Abstract" columns from the loaded data into a single list called documents. This list contains all the titles and abstracts of research articles. Later, the code creates a dictionary of unique words from the preprocessed documents. It also creates a "bag of words" representation (called a corpus) for each document. The corpus represents how often each word appears in a document. Overall, these preprocessing steps are essential for cleaning and standardizing the text data extracted from input files.

C. Data Analysis

For analysis, we used the LDA method. We used the LDA from Python's Gensim package. There is one problem with LDA it does not give optimal number of topic. The number of topics for the LDA model could be selected using the perplexity, JS divergence, coherence, and stability methods [25]. We chose the coherence method to find the optimal topic number. Latent Dirichlet Allocation (LDA) relies heavily on high-quality data input for accurate topic modeling. The coherence method looks at how well words fit together in a model, checking the chance that words from the top-ranked topics appear together in a document. Our code computed coherence scores for different numbers of topics (ranging from 2 to 20). The optimal number of topics that gave the highest coherence score was 16. The ideal number of words per topic in Latent Dirichlet Allocation (LDA) is not fixed and can vary based on the specific dataset and research objectives. However, too few words per topic may result in vague or overly broad topics as well as too many words per topic can make them difficult to understand. Many studies on topic models typically use the top 5 to 20 words (Choosing Words in a Topic, Which Cut-off for LDA Topics? n.d.). However, we decided to use ten words for each topic.

D. Topic Labeling

 doi:10.20944/preprints202410.2596.v1

The LDA model helps find themes, and each theme is given a label manually by looking at the keywords related to it. It refers to a way of giving descriptive and meaningful labels to the latent topics that are derived from a set of text documents. Researchers try to discover groups of related words (topics) that stand in for general concepts in the corpus when they use topic modeling approaches.

Assigning labels to topics is crucial for understanding and delving deeper into a topic model. It serves not only to aid in interpretation but also to offer qualitative backing when choosing among different models. Labeling topics can highlight which topics are more pertinent to a research query or, conversely, identify less informative ones. This becomes particularly important in models with numerous topics, as distinctions between topics might be more about language structure than actual content; for instance, comparing "eating healthy" with "following a balanced diet" showcases how topic labels can vary based on nuances in wording [26].

E. Comparison with Traditional Literature Review Methods

We used Latent Dirichlet Allocation (LDA) for topic modeling and also did a traditional literature review on some of the same articles. This helped us compare the strengths and weaknesses of both methods. LDA found hidden trends, while the traditional review gave a deeper understanding of key themes. Comparing both methods, we tested how well LDA could find new research trends and checked its accuracy against known literature.

VI. Result

The LDA-Gensim model produced ten words for each of the 16 topics. A coherence value ranging from 0.5 to 0.7 is considered a good score for identifying semantic similarities among the main words of a topic (Evaluation of Topic Modeling: How to Understand a Coherence Value/C_v of 0.4, Is It Good or Bad? n.d.). The coherence score is 0.3357. Since LDA is unsupervised, it might not group words into topics accurately. It works by using the probability of word occurrence in the text. The domain is very diverse, encompassing areas such as marketing, supply chain, social media, algorithm performance, clinical studies, and pollution studies. The top 30 most salient terms are –

Data, Quality, Chain, Supply, Big, Health, Evaluation, Adoption, Analytics, Study, Management, Firm, Model, Research, Performance, System, Medical, Decision, Record, Learning, Information, Science, Database, Method, Big Data Analytics, Based, Patient, Network, Detection, Assessment.

A. Research Area

After extracting key terms from the LDA model, we labeled the resulting topics by scrutinizing the associated keywords.

Table 3. Key Terms And Topic Label.

No.	Key Terms	Topic Label
1.	data, research, study, quality, big, performance, record, firm, process, impact	Data-Driven Decision and Performance Analysis
2.	data, quality, big, method, based, model, proposed, evaluation, process, result	Big Data Evaluation Methods
3.	data, system, analytics, quality, analysis, social,	Social Media Data Analytics and Quality Research

 doi:10.20944/preprints202410.2596.v1

	research, medium, based, line	
4.	data, quality, big, algorithm, database, result, model, performance, processing, process	Big Data Processing and Algorithm Performance
5.	chain, supply, data, performance, big, analytics, model, management, organizational, operational	Big Data Analytics in Supply Chain Management
6.	data, medical, detection, leisure, website, phishing, image, record, big data analytics, agriculture	Big Data Analytics in Medical
7.	data, quality, information, study, model, method, pollution, present, analysis, result	Data Quality Analysis in Pollution Studies
8.	data, decision, big, research, quality, marketing, big data analytics, sensing, technology, analytics	Big Data Analytics for Decision-Making and Marketing
9.	data, health, quality, method, study, big, clinical, research, patient, medical	Big Data in Clinical Research
10.	data, big, quality, study, analysis, research, collection, use, network, using	Network Analysis by Data Quality
11.	data, quality, big, assessment, measure, improvement, label, supplier, decision, using	Data Quality Measures for Supplier Assessment

 doi:10.20944/preprints202410.2596.v1

12.	data, quality, research, management, survey, big, framework, machine, learning, source	Data Quality Management with Machine Learning
13.	adoption, firm, data, big data analytics, study, management, business, theory, perceived, analytics	Business Analytics and Big Data
14.	data, science, big, quality, information, analysis, research, challenge, system, method	Challenges in Data Quality Analysis
15.	data, learning, deep, method, quality, big, analytics, model, machine, approach	Deep Learning Methods for Data Analytics
16.	data, specie, method, area, research, sampling, database, urban, trend, google	Data Sampling Methods

B. Research Trends

These research topics collectively explore the different roles of big data and its quality across diverse fields. For example, “Impact of data-driven decision-making in Lean Six Sigma: an empirical analysis”, “Trusted Decision-Making: Data Governance for Creating Trust in Data Science Decision Outcomes” and “Data Management in Industrial Companies: The Case of Austria” papers discussed data-driven decision and performance analysis related to big data. Likewise, “On Two Existing Approaches to Statistical Analysis of Social Media Data” and “The Rise of NoSQL Systems: Research and Pedagogy” are related to social media data analytics and quality research. We derived those topics using this method. In business and marketing, big data enhances firm performance, supply chain management, and decision-making processes by providing insights and improving operational efficiency. Furthermore, the rise of data-driven decisions has made market intelligence (MI) crucial for marketing [27]. Over the years, extensive research has greatly impacted how businesses use MI for better marketing decisions.

In the realm of technology, the focus shifts to the development and evaluation of algorithms and models that process data more effectively, aiming to boost performance and result accuracy. The application of big data extends to health and environmental sciences, where it aids in clinical research, patient care, and pollution study, underscoring the necessity of robust data quality and sophisticated analytical methods. Additionally, topics like social media analytics and urban planning demonstrate big data's capacity to influence societal trends and public policy. Our analysis of how topics have changed over time shows important trends in big data quality. For example, 'environmental science' didn't appear in the literature until 2016, indicating a growing interest in applying big data quality principles to this area. This shift shows how big data is becoming important in more sectors. Our chart (Table 4) shows the popularity of key topics from 2010 to 2024, revealing

 doi:10.20944/preprints202410.2596.v1

clear shifts in research focus, such as an early focus on 'supply chain management' that later moved towards 'healthcare applications' in big data. Overall, these topics emphasize the critical importance of data quality and the innovative use of analytical methods to harness the potential of big data across various sectors.

VII. Future Research Directions

The literature on data quality in big data analytics is rapidly changing and covers many research domains. However, there are still areas that need more evidence and exploration. We aim to contribute by identifying key research questions for future studies. Our review suggests important questions (Table 4) for improving understanding and addressing challenges in big data analytics across various sectors.

Table 4. Future Research Avenues.

No.	Future Avenues
1.	How can firms use big data to enhance performance and decision-making, and what key factors influence the effectiveness of these data-driven strategies?
2.	What are the most effective methods for evaluating big data quality, and how can these methods be standardized across different industries?
3.	How can social media data analytics ensure high data quality, and what methods can address the challenges of analyzing user-generated content and social trends?
4.	What are the best ways to optimize algorithms in big data processing, and how does data quality affect their results?
5.	How can big data help supply chain management, what challenges and opportunities are there in keeping data quality high?
6.	How does big data analytics help medical detection, diagnosis, how can we keep data quality high in various medical applications?
7.	What are best models, methods for analyzing pollution data accurately?
8.	How can we make marketing analytics data better, and why does good data help make better decisions and marketing plans?

 doi:10.20944/preprints202410.2596.v1

9.	How does good data quality affect clinical research, and how can big data analytics help improve patient care and medical research?
10.	How can we use network analysis to make big datasets better, and what's the best way to collect and use network data in research?
11.	How can we measure data quality when evaluating suppliers, and how does this help in managing them better?
12.	How can we use ML to handle, guarantee data quality in research and surveys, and what are the main challenges in doing so?
13.	What factors affect how much businesses use big data analytics, and how does this affect their performance and decisions?
14.	What are the main challenges in ensuring data quality in big data analysis, and what methods and systems can be developed to address these challenges effectively?
15.	How can deep learning methods be optimized for data analytics, and what are the critical factors affecting the quality and performance of deep learning models?
16.	What are the most effective data sampling methods for ecological research, and how can these methods be applied to urban and species studies to ensure accurate trend analysis?

VIII. Discussions

Data quality in big data is having a major impact across many fields. The research trends from the topics provided illustrate a broad focus on leveraging big data and its quality to enhance various domains. Topics 1 and 4 highlight the impact of big data on business performance and processes. It emphasizes the importance of data quality in improving firm strategies and operational efficiencies. Topic 5 explores how big data analytics transform supply chain management. Topics 7 and 9 focus on environmental sciences and clinical research, respectively. Here, data-driven studies support critical analyses of pollution and patient care, enhancing outcomes through high-quality data models. In technological and methodological advancements, Topics 2, 4, and 15 concentrate on refining algorithms, databases, and deep learning techniques to achieve more accurate results and robust data processing capabilities. Considerable advancements have been made in the detection of intracranial hemorrhage (ICH) with the use of deep learning algorithms in recent study. The research employs attention processes in convolutional neural networks (CNNs) to attain a 99.76% accuracy rate and an AUC-score of 0.9976. This successfully improves the speed and accuracy of neuroimaging analysis, which in turn leads to better outcomes for patients [28]. Topic 3 discusses the integration of data

 doi:10.20944/preprints202410.2596.v1

analytics with social media, aiming to derive insights from user-generated content and social trends. Topics 10 and 12 deal with the collection, management, and analysis of large datasets, using advanced frameworks like machine learning to parse vast information networks. Topic 6 extends to specialized fields such as medical and agricultural analytics, showcasing big data's versatility in detecting and analyzing trends across various sectors. Topic 8 merges big data with marketing and sensing technologies, illustrating its pivotal role in informed decision-making and strategic marketing efforts. Topic 11 focuses on data quality measures for supplier assessment, ensuring high standards in supplier management processes. Topic 13 addresses the adoption of business analytics and big data in firm management, while Topic 14 discusses the challenges in data quality analysis, emphasizing the need for robust systems and methods. Topic 16 highlights data sampling methods in ecological research, particularly in urban and species studies, showcasing the application of big data in environmental monitoring and trend analysis.

Based on the key terms and topic labels, studies are exploring the effects of data accuracy and reliability on business strategies, operational efficiencies, and overall organizational success. There's also a growing interest in developing and refining algorithms, databases, and deep learning techniques to enhance data processing and analysis capabilities. Big data quality research extends across various domains, including healthcare, environmental science, social media analytics, and urban planning. Additionally, Big data analytics is increasingly being used in marketing decision-making processes, and there's a focus on developing frameworks and methodologies for effective data collection, management, and analysis. These trends collectively underscore the importance of data quality in driving innovation, improving decision-making, and addressing societal challenges across diverse fields.

Data quality in big data is having a major impact across many fields. The research trends from the topics provided illustrate a broad focus on leveraging big data and its quality to enhance various domains. For instance, LDA analysis identified "Supply Chain Management" as a significant topic, underscoring the critical role of data quality in optimizing supply chain operations. Similarly, the emergence of "Environmental Science" as a prominent topic through LDA highlights the growing importance of data quality in this field. This finding suggests that as environmental data becomes increasingly integrated into big data analytics, ensuring its accuracy and reliability is vital for effective environmental monitoring and decision-making. LDA showed the importance of "Healthcare Applications,". It reflects the necessity of high-quality data in patient care and clinical research. The implications of poor data quality in this field are particularly severe. Because they can directly affect patient outcomes. These findings are supported by the literature, which emphasizes the need for rigorous data quality frameworks in healthcare. Data quality across different fields illustrates that "big data" and "data quality" are not just theoretical constructs but are deeply interconnected with practical applications across various sectors.

IX. Comparison of Traditional Literature Review Methods with LDA

Using both traditional literature reviews and LDA gives a clear comparison. This approach shows what each method does well and where it falls short. We can see both broad patterns and detailed insights.

A. Depth vs. Breadth

Traditional literature reviews give importance to depth. They examine key concepts, theories, and methods in detail. This approach shows how foundational ideas, like the 5Vs, have evolved and been applied across industries. LDA, on the other hand, focuses on breadth. It analyzes large datasets to find hidden topics. LDA identifies new trends and shifts in research focus. For example, it revealed the rise of "Environmental Science" as an important area. This broad approach helps ensure that no trends are missed.

B. Fact-based vs. Opinion-based

 doi:10.20944/preprints202410.2596.v1

LDA is more fact-based. It tries to find topics based on data patterns, free from human bias. This helps reveal less obvious trends, like big data's links to healthcare and environmental studies. Another aspect is LDA looks at all data equally. It treats every document the same. This avoids focusing too much on famous papers or top journals, which can bias traditional reviews. Traditional literature reviews are often subjective. The analysis depends on the researcher's choices and interpretations. This subjectivity allows for detailed exploration but also could be biased. For example, the researcher emphasized "Data Governance" and "Supply Chain Management." These topics were chosen based on the reviewer's knowledge and popular journals.

C. Theoretical vs. Empirical

The traditional literature review method is deeply rooted in theory. It tells how concepts like data quality have been theoretically framed, debated, and refined over time. This theoretical grounding is essential for understanding the principles that guide research and practice in the field. In our study, the traditional review method allowed us to trace the development of key data quality frameworks, providing a context for interpreting the empirical findings. In contrast, LDA is more empirical. It focuses on the data itself to uncover trends. It identifies what is currently being discussed in the literature. It often highlights areas that received little theoretical attention but are gaining empirical significance. For instance, LDA's identified "Environmental Science" as a significant topic suggests that this area is becoming increasingly important in context of big data quality, even if it has not yet been fully integrated into theoretical discourse.

X. Contributions to Literature and Practitioners

This paper connects theoretical research with practical application, providing valuable insights for both academics and industry professionals. From a literature perspective, this paper reviews the literature on data quality in big data, categorizing methods used to ensure data quality. It helps understand how different aspects of data quality are handled in big data. Furthermore, this study uses LDA to find hidden themes in the literature on big data quality. This novel approach helps identify research trends and gaps, contributing methodologically to the field. Moreover, the paper identifies key areas where more research is needed by analyzing the literature's topics and themes. This highlights different aspects of data quality and emphasizes the need for new methods and tools specifically designed to address the challenges posed by big data. Finally, paper suggests a research framework that highlights current trends, gaps, providing direction for future studies. This framework is beneficial for scholars who want to explore new data quality aspects in big data context.

From a practitioner's perspective, this paper's findings and methods give practical advice for professionals looking to enhance data quality in big data settings. The systematic review and use of LDA offer clear strategies for spotting and tackling data quality problems. Reviewing various methods and frameworks for managing data quality, like data preprocessing, cleaning, and validation, gives professionals a set of best practices. These can be used in different situations to maintain high data quality. The paper shows how data quality management is used in various fields like healthcare, social media, and finance. This industry-specific insight helps professionals customize data quality strategies to fit their particular sector's needs. The paper explores advanced techniques like LDA for topic modeling, offering practitioners innovative tools to analyze large datasets. This guidance is essential for implementing effective data quality assessment and improvement strategies in real-world scenarios.

XI. Limitations and Future Research

Our study included all articles on data quality in big data from the Web of Science (WoS). Future research could use other scholarly literature databases (e.g., Scopus, IEEE Xplore, etc.). Some articles may use inappropriate terms, for example, 'Big Data' and 'Data Quality' in their titles, keywords, or abstracts, but their relevance to the actual research might not align appropriately. This study did not rigorously assess the quality of journals; we considered articles from WoS without discriminating

based on journal rating or relevance. Lastly, our research involved a combination of automated manual analyses.

To address limitations identified on data quality in big data studies, several areas for future work. First, refining search parameters could enhance article selection by ensuring only highly relevant publications are included. Second, implementing a rigorous assessment process for journal quality could improve the overall caliber of sourced articles. Third, enhancing automation techniques could streamline data collection/analysis, reducing reliance on manual efforts. Additionally, integrating multiple search engines may broaden access to information & overcome limitations.

XII. Conclusion

Wrapping up our research on big data quality, it's clear that ensuring data accuracy and reliability is crucial for decision-making. By digging into existing studies and using advanced methods like LDA for topic modeling, we've uncovered some important trends. Firstly, we've seen that big data is growing rapidly, and with need for better data quality. Large amount of data and its different types make maintaining quality a big challenge. Our research highlights the importance of good data in the big data world. While our study faced some challenges, like dealing with a vast research field and relying on both manual & automated analyses, it's kind of like we've set the stage for future explorations in this area. By tackling these challenges head-on & continuing to innovate, we can unlock the full potential of big data to benefit everyone.

References

1. D. Reinsel, J. Gantz, and J. Rydning, “The Digitization of the World From Edge to Core,” Seagate, Nov. 2018.
2. D. Rao, V. N. Gudivada and V. V. Raghavan, "Data quality issues in big data," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 2015, pp. 2654-2660, doi: 10.1109/BigData.2015.7364065.
3. N. Kshetri, M. M. Rahman, S. A. Sayeed and I. Sultana, "cryptoRAN: A Review on Cryptojacking and Ransomware Attacks W.R.T. Banking Industry - Threats, Challenges, & Problems," 2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 2024, pp. 523-528, doi: 10.1109/InCACCT61598.2024.10550970.
4. L. Cai and Y. Zhu, “The challenges of data quality and data quality assessment in the big data era,” *Data Science Journal*, vol. 14, no. 0, p. 2, May 2015, doi: 10.5334/dsj-2015-002.
5. S. Marcos-Pablos and F. J. García-Peñalvo, “Information retrieval methodology for aiding scientific database search,” *Soft Computing*, vol. 24, no. 8, pp. 5551–5560, Oct. 2018, doi: 10.1007/s00500-018-3568-0.
6. B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, “Systematic literature reviews in software engineering – A systematic literature review,” *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Nov. 2008, doi: 10.1016/j.infsof.2008.09.009.
7. R. R. Xiong, C. Z. Liu, and K.-K. R. Choo, “Synthesizing Knowledge through A Data Analytics-Based Systematic Literature Review Protocol,” *Information Systems Frontiers*, Oct. 2023, doi: 10.1007/s10796-023-10432-3.
8. C. Zou and D. Hou, "LDA Analyzer: A Tool for Exploring Topic Models," 2014 IEEE International Conference on Software Maintenance and Evolution, Victoria, BC, Canada, 2014, pp. 593-596, doi: 10.1109/ICSME.2014.103.
9. W. Elouataoui, I. E. Alaoui, and Y. Gahi, “Data Quality in the Era of Big Data: A Global Review,” *Studies in Computational Intelligence*, pp. 1–25, Jan. 2022, doi: 10.1007/978-3-030-87954-9_1.
10. F. Ridzuan, W. M. N. W. Zainon, and M. Zairul, “A Thematic Review on Data Quality Challenges and Dimension in the Era of Big Data,” *Lecture Notes in Electrical Engineering*, pp. 725–737, Sep. 2021, doi: 10.1007/978-981-16-2406-3_56.
11. M. T. Ijab, E. S. M. Surin, and N. M. Nayan, “CONCEPTUALIZING BIG DATA QUALITY FRAMEWORK FROM A SYSTEMATIC LITERATURE REVIEW PERSPECTIVE,” *Malaysian Journal of Computer Science*, pp. 25–37, Nov. 2019, doi: 10.22452/mjcs.sp2019no1.2.
12. M. Mirzaie, B. Behkamal, and S. Paydar, “Big Data Quality: A systematic literature review and future research directions,” *arXiv (Cornell University)*, Jan. 2019, doi: 10.48550/arxiv.1904.05353.
13. J. Liu, J. Li, W. Li, and J. Wu, “Rethinking big data: A review on the data quality and usage issues,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 134–142, Dec. 2015, doi: 10.1016/j.isprsjprs.2015.11.006.

14. S. Ji, Q. Li, W. Cao, P. Zhang, and H. Muccini, "Quality Assurance Technologies of Big Data Applications: A Systematic Literature Review," *Applied Sciences*, vol. 10, no. 22, p. 8052, Nov. 2020, doi: 10.3390/app10228052.
15. A. Ramasamy and S. Chowdhury, "Big Data Quality Dimensions: A Systematic Literature Review," *Journal of Information Systems and Technology Management*, May 2020, doi: 10.4301/s1807-1775202017003.
16. N. Kshetri, R. Mishra, M. M. Rahman and T. Steigner, "HNMblock: Blockchain Technology Powered Healthcare Network Model for Epidemiological Monitoring, Medical Systems Security, and Wellness," 2024 12th International Symposium on Digital Forensics and Security (ISDFS), San Antonio, TX, USA, 2024, pp. 01-08, doi: 10.1109/ISDFS60797.2024.10527226.
17. M. M. Rahman, N. Kshetri, S. A. Sayeed, and M. M. Rana, "AssessITS: Integrating procedural guidelines and practical evaluation metrics for organizational IT and Cybersecurity risk assessment," *arXiv.org*, Oct. 02, 2024. https://arxiv.org/abs/2410.01750
18. S. Sarsfield: The butterfly effect of data quality, 5th MIT IQIS, 2011.
19. IBM, What is data quality? https://www.ibm.com/topics/data-quality
20. "ISO 25012." https://iso25000.com/index.php/en/iso-25000-standards/iso-25012, 2019
21. I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli and C. Bouhaddioui, "Big Data Quality: A Quality Dimensions Evaluation," 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), Toulouse, France, 2016, pp. 759-765, doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122.
22. C. Sharma, I. Batra, S. Sharma, A. Malik, A. S. M. S. Hosen and I. -H. Ra, "Predicting Trends and Research Patterns of Smart Cities: A Semi-Automatic Review Using Latent Dirichlet Allocation (LDA)," in IEEE Access, vol. 10, pp. 121080-121095, 2022, doi: 10.1109/ACCESS.2022.3214310.
23. R. Pranckutė, "Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World," *Publications*, vol. 9, no. 1, p. 12, Mar. 2021, doi: 10.3390/publications9010012.
24. H. Jelodar *et al.*, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, Nov. 2018, doi: 10.1007/s11042-018-6894-4.
25. J. Gan and Y. Qi, "Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example," Entropy, vol. 23, no. 10, p. 1301, Oct. 2021, doi: 10.3390/e23101301.
26. S. J. Weston, I. Shryock, R. Light, and P. A. Fisher, "Selecting the Number and Labels of Topics in Topic Modeling: A Tutorial," Advances in Methods and Practices in Psychological Science, vol. 6, no. 2, p. 251524592311601, Apr. 2023, doi: 10.1177/25152459231160105.
27. M. W. Soykoth, W. Sim, and S. Frederick, "Research trends in market intelligence: a review through a data-driven quantitative approach," Journal of Marketing Analytics, Feb. 2024, doi: 10.1057/s41270-023-00285-9.
28. N. Kumar Kar, S. Jana, A. Rahman, P. Rahul Ashokrao, I. G and R. Alarmelu Mangai, "Automated Intracranial Hemorrhage Detection Using Deep Learning in Medical Image Analysis," 2024 International Conference on Data Science and Network Security (ICDSNS), Tiptur, India, 2024, pp. 1-6, doi: 10.1109/ICDSNS62112.2024.10691276.