IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14417

AN AI FRAMEWORK FOR UNSTRUCTURED DATA QUALITY ASSESSMENT WITH INTEGRATED PII DETECTION

Vanitha A[1], Jumana J[2]

Department of Computer Science, Rathinam College of Arts and Science[1,2]

Abstract: Unstructured data, which includes everything from text, documents, and emails to scanned images and log files, is the dominant type of data in many industries, including the realm of enterprise systems and digital communications. Despite the immense potential this type of data holds for analytics and decision-making processes, the usefulness of this data is hindered by the quality issues it faces, including duplication, inconsistency, incomplete data, and unclear data. Making the situation even worse is the presence of personal data, which creates privacy and compliance issues. The current data quality frameworks, which were initially designed to work with structured data, are inadequate to deal with the challenges posed by unstructured data. As such, this project seeks to address the limitations of the current data quality frameworks by developing an innovative and exhaustive data quality assessment framework. This framework is designed to automatically assess the quality of the data while protecting the privacy of the data. It incorporates anomaly detection techniques for log data, cleaning and normalization techniques for text data, and OCR techniques for image data. Additionally, the framework incorporates transformer-based techniques to automatically identify and mask PII. Data quality is assessed based on different parameters, including completeness, consistency, duplication, semantic correctness, and privacy. Beyond reporting, the system produces a cleaned, privacy preserved dataset that is ready for safe use in analytics and machine learning pipelines. By combining AI driven quality assessment with automated privacy safeguards, this project bridges a critical gap between data reliability and regulatory compliance, offering organizations a scalable solution for managing unstructured data with confidence.

Keywords: Unstructured Data, Data Quality Framework, Automation, PII Detection, Text Analytics, Image Quality, Log Analysis, AI-driven Data Cleaning

I. INTRODUCTION

Data has become the lifeblood of modern organizations in terms of analytics, machine learning, and decision making. However, data quality is what makes data valuable. High-quality data can lead to accurate insights and predictions as well as regulatory compliance, while low-quality data can lead to costly errors and reputational damage. In the real world, there has been a rapid increase in unstructured data in various domains, including e-commerce, healthcare, and enterprise systems, in the form of text documents, scanned images, and system logs. Unlike structured data, unstructured data is messy and inconsistent in nature and may contain personal information, making it difficult and risky to process. The problems associated with unstructured data include missing values, duplication, inconsistency, ambiguity in data content, and abnormalities in log data. These problems make data difficult to use and create a lack of confidence in data analysis results. However, there is a problem in terms of data privacy due to the presence of personal identifiable information (PII) in data content in the form of names, addresses, phone numbers, etc. The data quality framework is not designed to handle these problems in terms of unstructured data. To overcome these problems associated with data quality in terms of unstructured data, a comprehensive automated framework is proposed in this project. The objective is clearly defined in terms of quality assessment of data content in the form of text data, images, and log data; generating a report highlighting data quality problems and recommending solutions; and automatically cleaning data based on recommendations. A critical aspect of data quality assessment is identifying and securing data by detecting PII in data content and applying techniques such as data masking and data anonymization. By applying anomaly detection techniques, text data cleaning techniques, OCR techniques. the framework first generates a detailed quality report based on the performance evaluation in this report, the system then applies recommended remediation steps to automatically clean the data and secure any sensitive fields.

II. PROBLEM STATEMENT

Recent studies have shown some promise in enhancing the quality of unstructured data through the application of automated methods to identify inconsistencies and improve reliability. For example, transformer-based models have

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14417

shown promise in identifying sensitive information within text, and adaptive frameworks have also been proposed to assess data quality within a multimodal environment. However, these solutions have shown limitations, as they are mostly applicable to text-based data and cannot support current data environments, where information is often scattered across multiple modalities such as images, documents, and system logs. These different types of data require different methods to be extracted and processed, and current solutions have failed to provide a unified framework to handle them. The next important aspect where current solutions have failed to deliver is in terms of protecting data privacy. While current solutions have shown application promise in improving data quality, they fail to address issues such as the automatic detection and protection of Personally Identifiable Information (PII). This is a matter of concern, as an organization might be able to improve the quality of their data using such solutions, but they will still be at risk of suffering from issues such as privacy breaches and noncompliance with regulatory policies. This is a matter of concern, as current data environments heavily depend on reliable and secure unstructured data to support analytics, machine learning, and decision-making processes. In the absence of proper quality management techniques to address these issues in different formats and to ensure the non-disclosure of sensitive data, the data-driven approaches may lead to erroneous results and the disclosure of sensitive data. Furthermore, the current techniques may not effectively address the quality issues, such as duplication, inconsistency, missing data, and data anomalies, which affect the quality of the data. Moreover, the privacy issues may not be addressed adequately, with less emphasis on the incorporation of techniques to automatically identify Personally Identifiable Information (PII). Therefore, to overcome the aforementioned limitations, this project is developed to propose a comprehensive framework to address the quality issues using the combined approaches of multimodal extraction techniques and privacy-preserving data detection. The proposed framework can effectively identify the quality issues with diverse data formats, automatically generate the quality report with performance evaluation and recommendations, and subsequently clean the data based on the suggested steps. One of the important features of the proposed framework is the ability to ensure the non-disclosure of sensitive data by automatically detecting and masking PII.

III. SYSTEM ARCHITECTURE

A. System Overview

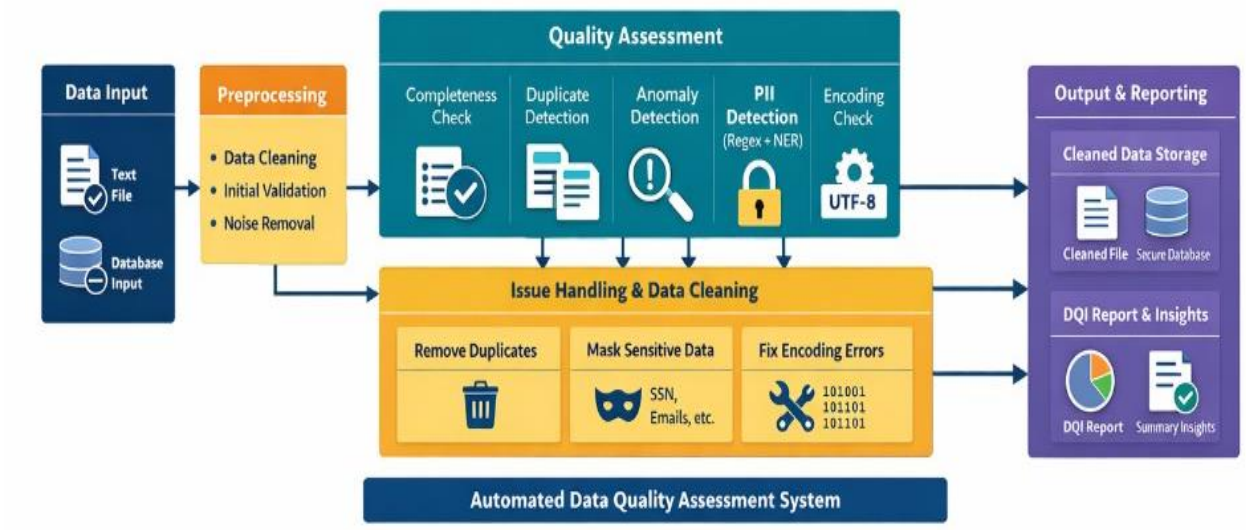

Fig. 1. The workflow architecture

cleaning of unstructured data by integrating various processing stages into a cohesive system. The system can accept various types of data as

1) **Data Ingestion Layer:** The first stage of the system involves collecting unstructured data from the user through a web-based interface. This step is essential because real-world data exists in multiple formats such as text files, CSV datasets, system logs, chat transcripts, and images. Supporting multiple formats ensures that the framework is flexible and applicable to real-world enterprise scenarios. The uploaded data is stored temporarily in memory or local storage to avoid persistence risks and to maintain privacy. This stage acts as the entry point of the pipeline and ensures that all subsequent processes receive standardized input regardless of the source format.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14417

2) **Preprocessing Layer:** The Preprocessing Layer focuses on preparing the raw data for quality assessment by performing essential initial transformations. This includes basic data cleaning operations such as noise removal, elimination of irrelevant characters, and initial validation checks to ensure that the data meets minimum quality requirements. The layer may also involve normalization steps, such as converting text to a standard encoding format. By addressing low-level inconsistencies at an early stage, this layer improves the effectiveness and accuracy of The proposed system architecture is based on a framework of AI-driven data quality assessment and the downstream quality assessment processes.

3) **Data Quality Assessment Layer:** In this stage, the system evaluates the quality of the data using multiple dimensions:

$$DQI = w1C + w2Co + w3U + w4V \quad (1)$$

- **Completeness:** it measures the proportion of non-missing values:
- **Consistency:** it checks whether data follows predefined formats (e.g., email regex validation).
- **Uniqueness:** it identifies duplicate records:

Uniqueness = (Number of unique records) / (Total records)

- **Validity:** it ensures that values fall within acceptable ranges or formats.
- **Anomaly Detection:** it is performed using machine learning models such as Isolation Forest, which detects outliers based on data distribution.

average path length: h(x) = average path length of point(x)

Expected Path Length: c(n) = 2H(n1)n2(n1)/n

Anomaly Score:

s(x) = 2c(n)h(x)/c(n)

- **Privacy Detection (PII):** Sensitive data such as emails, phone numbers, credit card numbers, and IP addresses are detected using regex and transformer-based Named Entity Recognition (NER). This stage is critical because it quantifies the reliability and usability of the data before any cleaning is performed.

(a) **Regex based detection(Rule-based):**

PII regex = No of values matching PII patterns / Total values

(b) NER-based Detection (ML-based):

PII ner = No. of detected sensitive entities/total value

4) **Issue Handling and Data Cleaning Layer**: The Issue Handling and Data Cleaning Layer is responsible for resolving the data quality issues identified in the previous stage. It applies targeted cleaning operations such as removing duplicate records, masking or anonymizing sensitive information to ensure privacy compliance, and correcting encoding errors to maintain consistency. The cleaning strategies are adaptive and depend on the type and severity of the detected issues. This layer ensures that the output data is accurate, consistent, and ready for analysis or storage, thereby significantly improving overall data reliability.

5) **Data Storage and Reporting Layer:** After cleaning, the system generates a comprehensive report that includes the Data Quality Score (DQI), issues detected, column-level insights, and a summary of cleaning actions performed. This report provides transparency and actionable recommendations for further improvement. Finally, users can download both the cleaned dataset and the quality report in formats such as text or PDF. This dual output ensures that organizations not only receive improved data but also gain insights into the quality issues and remediation steps taken.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14417

IV. LITERATURE REVIEW

The increasing reliance on unstructured data across domains such as e-commerce, healthcare, and enterprise systems has significantly accelerated research in data quality assessment, anomaly detection, data cleaning, and privacy preservation. Traditional data quality approaches have largely focused on structured datasets, utilizing rule-based validation, schema enforcement, and duplication checks; however, these methods natural language processing (NLP) for text and optical character recognition (OCR) for images; however, existing frameworks provide limited automation and fail to incorporate privacy aware remediation.

In the domain of sensitive information detection, recent studies such as the DEXA 2024 comparative analysis have highlighted the superiority of transformer-based models like BERT and RoBERTa over conventional regex-based methods for identifying personally identifiable information (PII) in text, though these approaches remain constrained to single data modalities. Despite these advancements, significant gaps persist, particularly the lack of a unified, end-to-end framework that integrates multimodal data quality assessment, anomaly detection, automated cleaning, and privacy preservation. Addressing these limitations, the proposed work introduces a comprehensive and scalable architecture that extends beyond existing solutions by incorporating adaptive preprocessing, integrated anomaly detection, and advanced transformer-based PII detection and masking across diverse data types, including text, logs, and images. This holistic approach not only enhances the accuracy and robustness of data quality evaluation but also ensures privacy compliance, thereby providing a more effective and practical solution for real-world unstructured data management.

Comparison table of existed work and proposed work:

Aspect	Previous research	My project (Unified framework)
Focus	Transformer model used for PII & adaptive scoring for multimodal DQ	Unified multimodal quality + privacy framework
Privacy handling	Detects PII in text only and limited privacy consumption	Automated PII detection & masking across modalities(text, images, log)
Cleaning mechanism	Basic remediation like duplicates, missing values..	Adaptive cleaning + privacy-aware remediation
Outcome	Quality index and adaptive assessment	Cleaned privacy- preserved dataset + quality report

From this table, the previous research has advanced either data quality assessment or PII detection individually, this project enhances both by integrating them into a single multimodal framework. By combining anomaly detection, adaptive cleaning, and transformer-based PII detection, the proposed system delivers a dual outcome: a comprehensive quality report and a cleaned, privacy-preserved dataset ready for analytics and machine learning.

V. RESULT AND DISCUSSION

The implementation of the proposed framework produced clear improvements in the overall quality and usability of unstructured data. During the initial assessment, the uploaded datasets exhibited common issues such as missing values, duplicate records, encoding errors, and the presence of sensitive information like email addresses and phone numbers. These problems were quantified through the Data Quality Index (DQI), which provided a baseline score reflecting the reliability of the raw data. After applying preprocessing and cleaning techniques, the DQI scores showed a significant increase, demonstrating the effectiveness of the framework in enhancing data completeness, consistency, and validity. For example, duplicate records were successfully removed, missing values were either filled or eliminated, and encoding errors were corrected. The anomaly detection module, powered by Isolation Forest, accurately flagged irregular patterns in log data, which were then handled through adaptive cleaning strategies. A critical outcome was the system's ability to detect and secure Personally Identifiable Information (PII). Using regex and transformer-based Named Entity Recognition (NER), sensitive fields such as emails, phone numbers, and IP addresses were automatically identified and masked.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14417

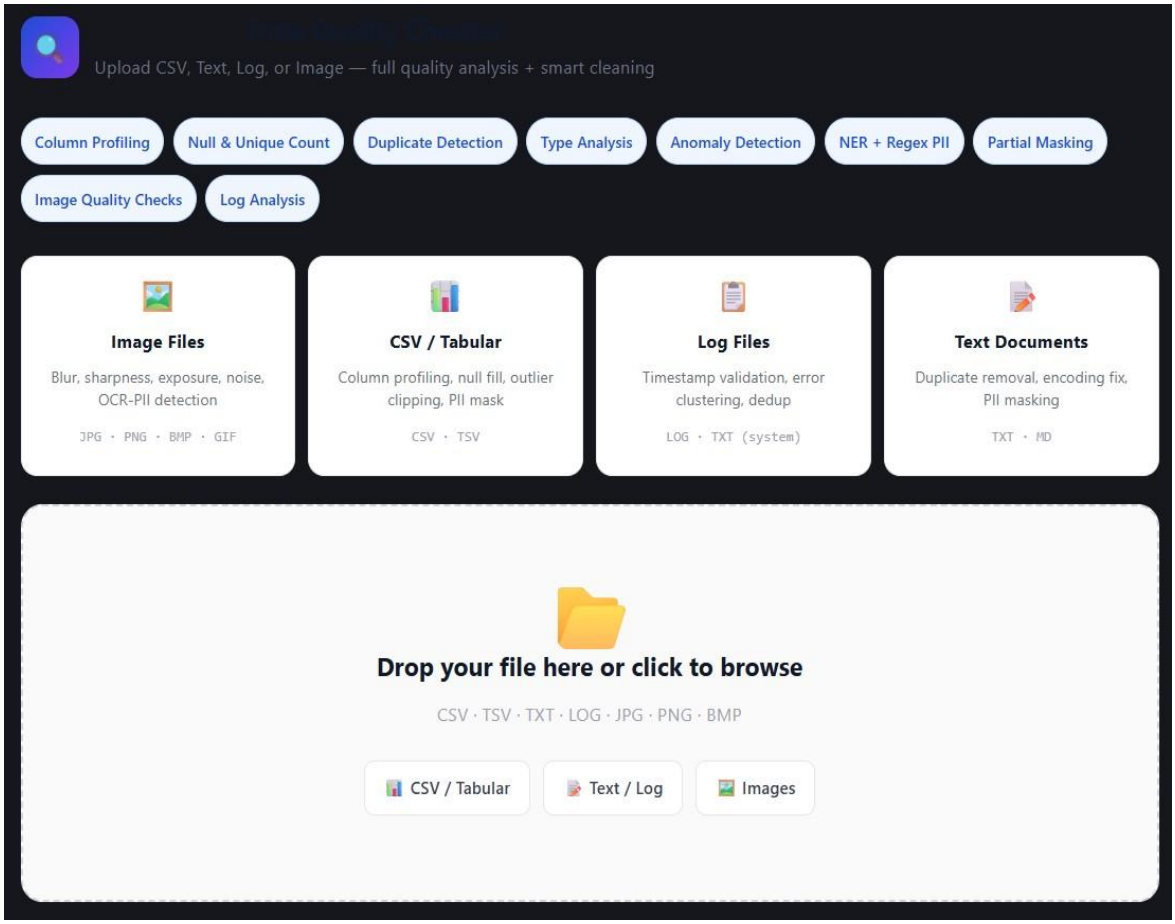

Fig. 2. the Quality checker site

This ensured that the cleaned dataset not only met quality standards but also complied with privacy regulations. The masking process preserved the usability of the data while eliminating risks of exposure. The generated quality report provided detailed insights into the issues detected, their frequency, and the corrective actions applied. Column-level analysis highlighted specific areas of weakness, while the summary offered actionable recommendations for further improvement. The cleaned dataset, exported alongside the report, represents a privacy-preserved and reliable resource ready for analytics and machine learning applications. From a performance perspective, the system achieved high detection accuracy for anomalies and PII, with precision and recall values indicating robust identification across diverse data types. The preprocessing techniques, including UTF-8 encoding for text and OCR for images, proved effective in standardizing heterogeneous inputs. Overall, the results confirm that integrating multimodal quality assessment with privacy-aware detection provides a scalable solution for managing unstructured data. the framework introduces adaptive cleaning strategies powered by machine learning models that learn from historical cleaning patterns to recommend the most appropriate remediation techniques for new datasets, allowing context-aware decisions such as whether to impute, delete, or substitute missing values. The system also emphasizes improved interpretability by providing visual representations of data quality improvements, including before-and-after comparisons of Data Quality Index (DQI) scores, issue counts, and PII detection results, which help users quickly assess the effectiveness of the cleaning process. Furthermore, domain-specific quality rules are incorporated to tailor data validation and cleaning operations according to industry requirements, such as those in healthcare, finance, or e-commerce, ensuring compliance with relevant standards and regulations. Finally, the framework supports continuous monitoring through a real-time processing mechanism, where incoming data is automatically evaluated and cleaned, complemented by a feedback loop that allows users to report overlooked issues, thereby enabling the system to continuously improve its accuracy and performance over time.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 4, April 2026

DOI: 10.17148/IJIREEICE.2026.14417

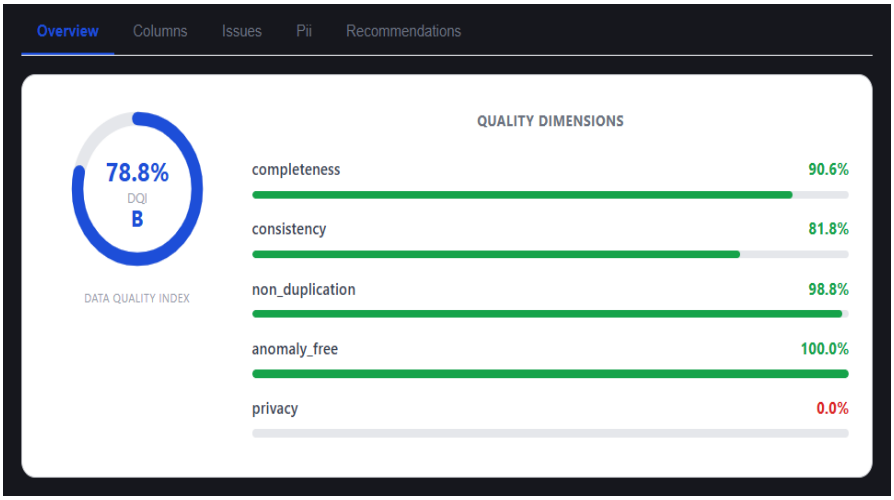

Fig. 3. Data Quality scoring Dimensions

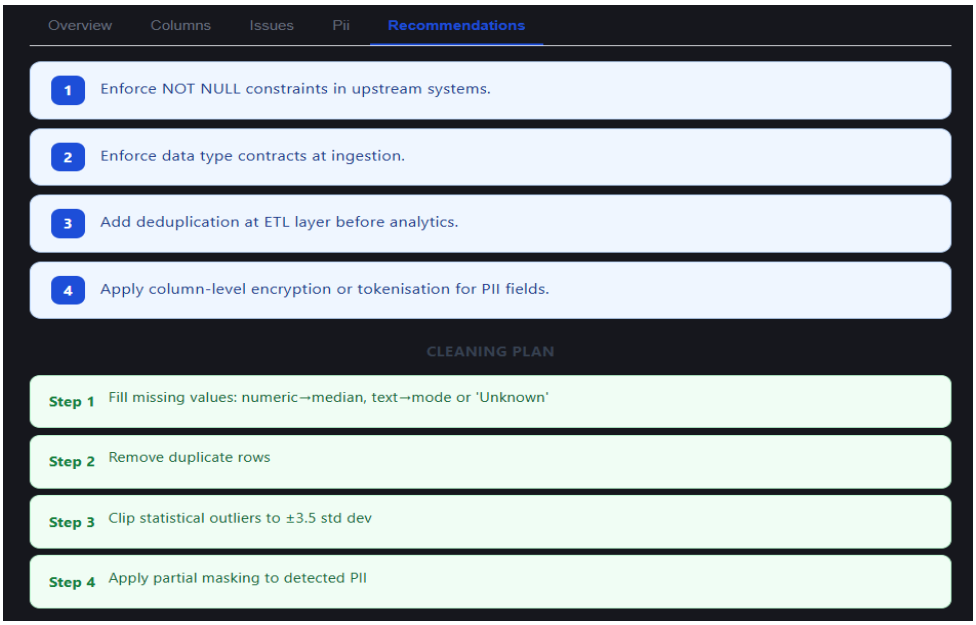

Fig. 4. Recommendations for further improvement

VI. FUTURE ENHANCEMENT

The proposed framework can be further strengthened by integrating advanced AI-driven capabilities that go beyond traditional rule-based remediation, leveraging deep learning models to learn context-aware cleaning strategies and enable intelligent operations such as predictive imputation for missing values and adaptive correction of anomalies based on historical data patterns. In addition, the system can be extended to provide comprehensive multimodal privacy protection by enhancing PII detection mechanisms to identify sensitive information not only in text but also in images (such as scanned identification documents), audio transcripts, and video frames, thereby ensuring robust privacy compliance across diverse data formats. Furthermore, the inclusion of interactive visualization dashboards would allow real-time monitoring of Data Quality Index (DQI) scores, issue distributions, and applied cleaning actions, offering greater transparency and enabling users to better understand the data transformation process while fostering trust in the system's effectiveness.

VII. CONCLUSION

This project set out to address the dual challenges of ensuring data quality and protecting privacy in unstructured multimodal datasets. By designing and implementing an automated framework, the system successfully demonstrated its ability to detect and remediate common data quality issues such as duplication, inconsistency, missing values, anomalies,

DOI: 10.17148/IJIREEICE.2026.14417

and encoding errors. At the same time, it integrated transformer-based techniques and regex methods to identify and mask Personally Identifiable Information (PII), thereby ensuring compliance with privacy regulations. The framework's layered architecture—spanning data input, preprocessing, quality assessment, cleaning, multimodal handling, and reporting—proved effective in producing both a cleaned dataset and a comprehensive quality report. The Data Quality Index (DQI) provided a quantifiable measure of improvement, while anomaly detection and PII masking reinforced the system's robustness and trustworthiness. The results confirm that combining multimodal extraction with privacy-aware detection significantly enhances the reliability, usability, and security of unstructured data. Beyond immediate outcomes, the project contributes a unified approach that bridges a critical gap in existing research, which often treats data quality and privacy as separate concerns. By integrating them into a single automated pipeline, the framework offers organizations a scalable solution for managing diverse data sources with confidence. In conclusion, this work demonstrates that unstructured data can be transformed into a reliable and privacy-preserved resource through systematic quality assessment and cleaning. The framework not only improves data readiness for analytics and machine learning but also safeguards sensitive information, making it a valuable contribution to both academic research and practical applications in data-driven environments.

REFERENCES

[1]. IEEE, "Adaptive Framework for Comprehensive Quality Assessment in Unstructured Big Data," *IEEE Conf.*, 2025.
[2]. S. Author et al., "Identifying Personally Identifiable Information (PII) in Unstructured Text: A Comparative Study on Transformers," in *Proc. DEXA*, 2024.
[3]. J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[4]. Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019.
[5]. K. El Emam et al., "A Systematic Review of Re-identification Attacks on Health Data," *PLoS ONE*, 2011.
[6]. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley, 2011.
[7]. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[8]. M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2016.
[9]. G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. Johns Hopkins Univ. Press, 2013.
[10]. R. Kimball and M. Ross, *The Data Warehouse Toolkit*, 3rd ed. Wiley, 2013.
[11]. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012.
[12]. M. Goldstein and S. Uchida, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms," *Knowl. Inf. Syst.*, vol. 28, pp. 363–387, 2011.
[13]. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.
[14]. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. ICDM*, 2008, pp. 413–422.
[15]. M. Stonebraker et al., "Data Curation at Scale: The Data Tamer System," in *Proc. CIDR*, 2013.
[16]. X. Chu, I. F. Ilyas, and P. Papotti, "Holistic Data Cleaning: Putting Violations into Context," in *Proc. ICDE*, 2013, pp. 458–469.
[17]. T. Dasu and T. Johnson, *Exploratory Data Mining and Data Cleaning*. Wiley, 2003.
[18]. 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. 2, pp. 1–10, 2015.
[19]. C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*. Springer, 2006.
[20]. E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.
[21]. A. Vaswani et al., "Attention Is All You Need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
[22]. S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly, 2009.
[23]. R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *J. Manage. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.
[24]. R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley, 2011.
[25]. P. J. Rousseeuw and M. Hubert, "Anomaly Detection by Robust Statistics," *WIREs Data Mining*, vol. 1, pp. 73–79, 2011.
[26]. Z. Abedjan, X. Chu, I. F. Ilyas, M. Ouzzani, P. Papotti, M. Stonebraker, and N. Tang, "Detecting Data Errors: Where Are We and What Needs to Be Done?," *Proc. VLDB Endowment*, vol. 9, no. 12, pp. 993–1004, 2016.