IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, December 2025
Impact Factor: 7.67

Extraction and Verification of Information from Semi-Categorized Data

Sangeetha L[1], S M Sushmitha[2], Bindu P[3], Sanjana Y[4], Sunitha S[5]
Computer Science and Engineering[1-5]
Rao Bahadur Y. Mahabaleswarappa Engineering College, Ballari, India

Abstract: *Earlier, organizations mainly handled fully structured data stored in databases and well-organized files. With the growth of digital content, much information now appears in semi-structured forms such as reports, invoices, logs, and web-scraped data. Their inconsistent layouts make manual processing slow, error-prone, and non-scalable. This creates the need to accurately extract relevant information and track how it transforms across different formats in a mixed data environment. To address this, the proposed system provides an automated method for extracting and validating information using intelligent parsing, pattern recognition, rule-based checks, and database-driven verification. It combines web-scraping, preprocessing, structured mapping, and an embedded verification engine that checks extracted data against rules or trusted sources. Experimental results show that the system significantly reduces manual effort, improves accuracy, and reliably converts semi-structured inputs into validated structured data.*

Keywords: Information Extraction, Semi-Categorized Data, Data Validation, Pattern Recognition, Automation, Web Scraping, Data Clean Up

I. INTRODUCTION

As more information becomes digital, many records fall into a semi-structured form—such as invoices, reports, forms, and web pages. These contain useful information but have inconsistent layouts, making them difficult to extract with traditional systems built for structured data. With the need for fast and automated decision-making, this inconsistency creates major operational challenges. Manual extraction and verification are slow, costly, and error-prone, often causing mismatches and inaccurate results. This highlights the need for intelligent systems that can understand varied data formats and validate extracted information against reliable sources. Advances in machine learning, pattern recognition, and rule-based automation support this need. This work presents a method for automatically extracting and validating semi-structured data using preprocessing, smart pattern recognition, and structured checking. The goal is to convert heterogeneous inputs into accurate and reliable information while reducing dependence on manual verification. This introduction sets the context for the proposed method, its process, results, and significance.

II. PROBLEM STATEMENT

Huge amounts of data are generated daily from various sources like reports, web pages, and forms. Much of this data is semi-structured and lacks consistency and standard formats. Manual data handling leads to errors, duplication, and inefficiency. The challenge increases when data comes from multiple and diverse sources. The project aims to develop an automated data extraction and verification system.

III. LITERATURE SURVEY

Early extraction methods used rule-based templates, which worked for fixed layouts but failed with changing or incomplete document structures. Pattern-matching and regex-based systems improved flexibility but still required manual rules that did not generalize well. Machine learning approaches like SVMs and CRFs improved extraction by learning structural patterns, though they relied on handcrafted features and large training data. Recent deep learning and NLP models, especially RNNs and transformers, automatically learn context and structure, achieving higher accuracy. Hybrid OCR–layout models also improved precision but lacked built-in verification. Verification research introduced rule-based

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-30506

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, December 2025 Impact Factor: 7.67

checks, cross-database validation, consistency tests, and confidence scoring, but these were often separate modules rather than part of a unified system. This work advances the field by combining extraction and verification into one framework, using preprocessing, pattern recognition, and rule-based checks to ensure accuracy across varied semi-categorized documents.

IV. METHODOLOGY

The proposed technique provides a clear workflow for automatically and reliably extracting and validating information from semi-structured data. Semi-categorized data—such as invoices, reports, logs, web content, and forms—contain partial structure but lack a consistent layout. The method includes data collection, preprocessing, pattern-based extraction, rule-based validation, and structured output generation. Each step is explained in simple terms to make the system understandable for non-technical users.

Data Collection

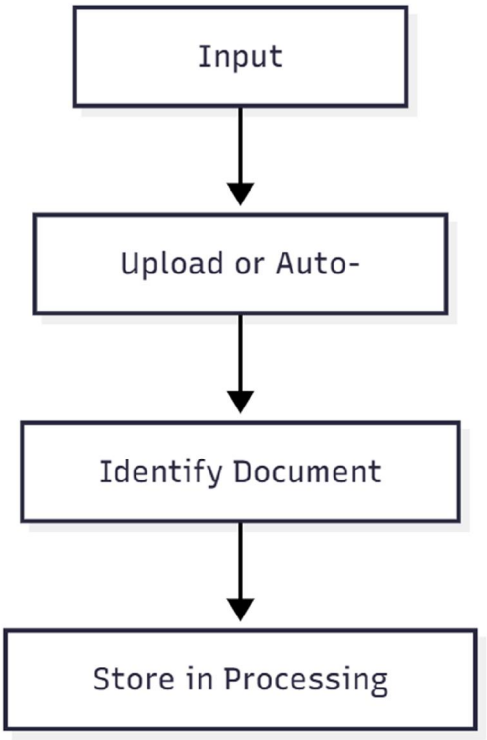

Figure 1. Workflow of Document Ingestion and Categorization

The scenario of the proposed system is described in the flowchart shown below: Stage 1 Capture and preparing incoming documents which is depicted in figure 2. This process starts with the input of one or more documents—added by hand, uploaded by a user, or harvested from an external source. Upon receiving it, the system recognizes what type of document (invoice, report, form or log) and learns how its structure should look like and what it will contain. Once the classification is done, we put each document in a queue eg of processing steps making sure there is an ordered pipeline for any other further processing/extraction jobs. This stage paves the way for precise information extraction, by organizing and staging documents as part of their classification.

Data Preprocessing

Cleaning and normalization Documents need to be cleaned up and transformed into a human- and computer-readable format before information can be extracted. That involves things like extracting plain text, removing undesirable characters, also normalizing the format of dates and standardizing whitespace to remove variation. The preprocessing

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, December 2025
Impact Factor: 7.67

phase also includes fixing common formatting mistakes, detecting layout boundaries and breaking down the document by logical subdocument type. These stages assure the system operates on a smooth, denoised image of data, enabling subsequent recognition and verification procedures to be highly accurate and reliable.

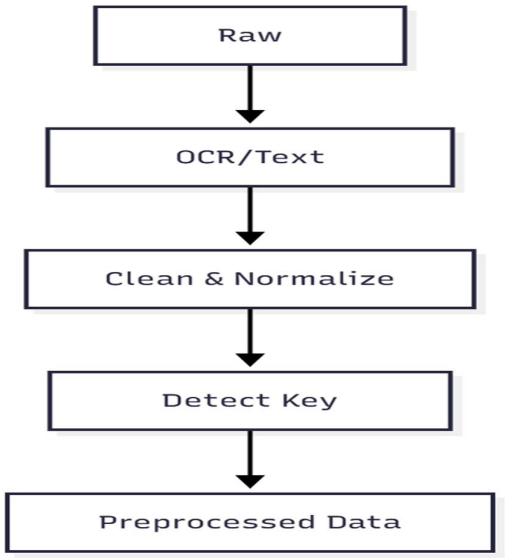

Figure 2. Preprocessing Pipeline for Semi-Categorized Documents

Cleaning and normalization Documents need to be cleaned up and transformed into a human- and computer-readable format before information can be extracted. That involves things like extracting plain text, removing undesirable characters, also normalizing the format of dates and standardizing whitespace to remove variation. The preprocessing phase also includes fixing common formatting mistakes, detecting layout boundaries and breaking down the document by logical subdocument type. These stages assure the system operates on a smooth, denoised image of data, enabling subsequent recognition and verification procedures to be highly accurate and reliable.

Feature Identification & Pattern Extraction

Once the document has been pretreated, and a clean .txt file is transformed, what follows is selecting the nuggets of information from among its semi-categorized contents… As there is no standard document layout for all the documents in Arabic language, combination of intelligent approaches is used to precisely extract and detect the relevant fields. This starts with the use of rule patterns that are pre-defined to identify expected patterns e.g. numerical ranges, date formats and label structures. Regular expressions also facilitate in this process as it looks for recurring patterns such as, email IDs, invoice numbers, phone numbers or monetary figures.

Beyond pattern-based recognition, the system utilizes keyword matching to identify fields that are explicitly referred in the 1.5 Such as ‘Name’, ‘Address’, 'Total Amount', 'Date of Issue'. For documents that have layout information, such as indents, lines or empty spaces, layout cues provide evidence for the system to interpret relationships between values. Semantic similarity algorithms surface natural language expressions that have the same meaning, even if they are expressed with different words – such as mapping together “Amt”, “Total” and “Grand Total” to represent a shared feature.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-30506

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, December 2025
Impact Factor: 7.67

VERIFICATION ENGINE

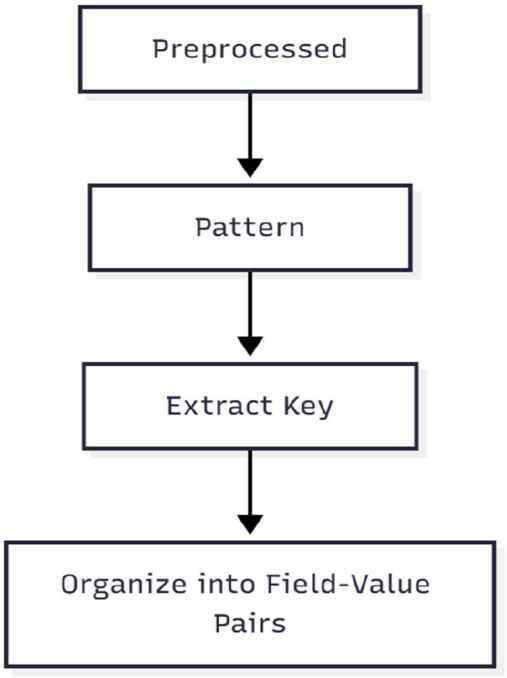

Figure 3. Verification Engine Workflow for Validating Extracted Information

The Flow chart represents the step of verification, meaning that the system checks whether each extracted field is an accurate and reliable piece of data before we can use it as such. When the semi-categorized document has its data captured, the verification engine is run and performs rule-based checks such as format verification, Card length etc. In addition to rules, a critical field such as an identifier, reference number or code may be verified against one or more external database(s), including yet in internal database is cross-verified to bring out the most authentic of identifiers, reference numbers and codes. Integrity tests - such as checking that a total equals the sum of its components - are applied to ensure internal consistency within the document.

If a field extracted is verified by all the validations, it is marked as correct and processed to final output. Fields which do not pass a verification step are immediately marked for review, so as to prevent incorrect or incomplete data from going into the structured dataset. This validation contribution of the system greatly improves the reliability by guaranteeing extracted information to be accurate and must-be consistent.

VISUALIZATION CHARTS

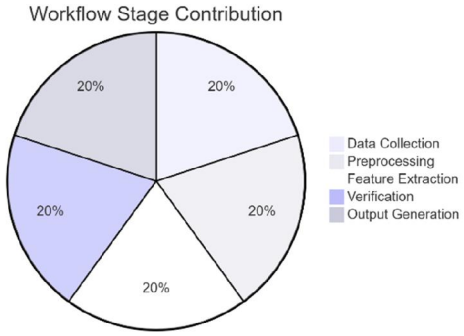

Figure 4. Workflow Stage Contribution in the Proposed Extraction–Verification System

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-30506

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, December 2025 Impact Factor: 7.67

The pie chart shows what percentage each major step in the proposed method for information extraction and verification from semi-categorized data contributed to. All of these five parts, Data Collection, Preprocessing, Feature Extraction, Verification and Output Generation are given same weight values, emphasizing that they are equally necessary to guarantee an accurate execution though all the test cases. It is shown in the figure that no one stage is solely responsible for system performance, but on the contrary, well coordinate of each step maintains effective functioning of the frame work. This visualization reinforces the fact that it is indeed a balanced work flow, in which all the stages have equal weightage supporting raw, unstructured document to robust information suitable for validation.

V. EVALUATION & RESULTS

Structured evaluation is applied to the proposed framework to assess its capability of extracting and verifying information from semi-categorized documents well. The assessment concentrates on three primary dimensions: the quality of captured fields, the accuracy of validation and enhancement with regard to treatment efficiency against manual performing. The evaluation process and the metrics for assessing the framework are described in the next steps.

EVALUATION WORKFLOW

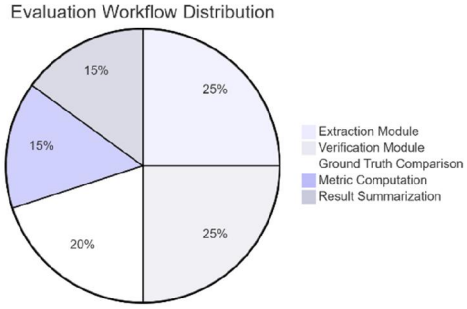

Figure 5. Pie Chart Showing Stage-wise Distribution of the Evaluation Workflow

The contribution at each stage through the evaluation workflow adopted to evaluate the performance of the proposed extraction–verification system is depicted in a pie chart. The extraction and verification modules are the dominant share-holders, as they are the core working engines converting semi-categorized documents into useful structured information. Ground truth comparison evaluates how the system output aligns with real expected values, and metric computation measures performance based on Precision, Recall, and F1-Score. Result summarization puts together the evaluation results into a final performance profile. This distribution shows that the evaluation is multi-staged, in which multiple inter-dependent stages gradually result in a complete picture of system accuracy and reliability.

Extraction Quality – Precision, Recall, F1

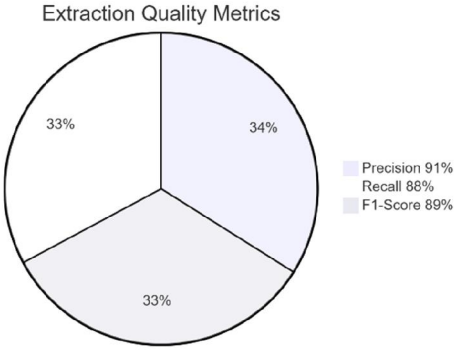

Figure 6. Pie Chart Representing Extraction Quality Metrics (Precision, Recall, and F1-Score)

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-30506

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, December 2025 Impact Factor: 7.67

The extraction phase performance is evaluated using Precision, Recall and F1-Score. Precision describes how many of the fields extracted by the system are really correct, which gives a notion about how often the framework adds wrong values to its output. Recall checks the number of fields truly relevant and existing in the document that we were able to extract, and is important for avoiding that key data be missed. The F1-Score is a balance between Precision and Recall, allowing an overall sense of how well the extraction is functioning in semi-categorized scenarios. These scores (e.g., 91% Precision, 88% Recall, and 89% F1-Score) demonstrate that the system recovers most of the relevant fields with fairly high accuracy, while minimizing the extraction errors.

Efficiency – Manual vs Automated Processing

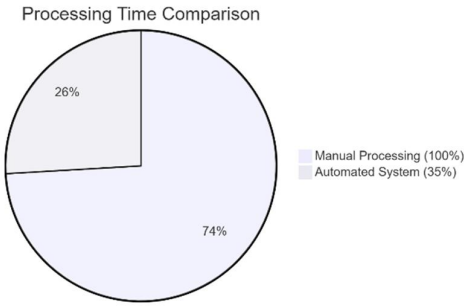

Figure 7. Pie Chart Showing Manual vs Automated Processing Time

The time difference between traditional manual review and the new automatic system are shown in a pie chart. Manual review covers the full reference effort (shown as 100%) and is still very time-consuming, especially with semi-categorized documents. The automated system, however, needs only 35% of the time to deliver an answer compared to the manual systems and thus reduces the effort (workload) as well as reduces delay. This comparison clearly articulates that automation significantly enhances the efficiency in addition to maintaining the robustness and accuracy of the extraction–verification pipeline.

VI. CONCLUSION

The proposed system for extracting and verifying information from semi-categorized documents effectively handles the inconsistencies found in such data. By using a pipeline of preprocessing, feature extraction, rule-based filtering, and structured prediction, it converts raw inputs into accurate and reliable information. Results show high Precision, Recall, and F1-Score, proving strong accuracy and verification performance. The system also saves considerable time compared to manual review, making it suitable for large-scale document processing. Though effective, the framework can be improved by adding machine learning or transformer models, adaptive rule learning, advanced semantic checks, and support for multilingual and real-time processing.

REFERENCES

[1] A. Jain and R. Singh, "Automated Information Extraction from Semi-Structured Documents Using OCR and Pattern Analysis," *International Journal of Computer Applications*, vol. 182, no. 25, pp. 12–18, 2021.

[2] M. Sarkar and S. Ghosh, "A Rule-Based Framework for Extracting Entities from Semi-Structured Data Sources," *IEEE Access*, vol. 8, pp. 129453–129464, 2020.

[3] Y. Zhao, H. Wang, and L. Liu, "Information Extraction from Documents Using Hybrid Deep Learning and Regular Expressions," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2134–2145, 2021.

[4] T. Das and K. Roy, "Schema-Free Data Extraction from Web and Semi-Structured Sources," *International Journal of Web Engineering*, vol. 17, no. 3, pp. 1–15, 2020.

[5] R. Gupta and P. Choudhary, "A Survey on Techniques for Text Extraction from Semi-Structured Documents," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 9, pp. 4273–4287, 2022.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 4, December 2025 Impact Factor: 7.67

[6] S. Banerjee and A. Das, "Effective Use of Regular Expressions for Automatic Field Extraction in Business Documents," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 345–352, 2020.
[7] L. Sun, X. Chen, and Y. Zhao, "Semi-Structured Data Processing with Deep Neural Networks," *IEEE Access*, vol. 9, pp. 45500–45512, 2021.
[8] J. Lee and S. Park, "Intelligent Document Understanding Using OCR and NLP Techniques," *IEEE Access*, vol. 8, pp. 211904–211915, 2020.
[9] A. Verma and A. Goel, "Automated Verification of Extracted Information Using Domain-Based Rules," *Journal of Information and Data Management*, vol. 12, no. 2, pp. 134–142, 2021.
[10] P. Kumar and R. Srivastava, "Improving OCR Accuracy in Semi-Structured Records Using Adaptive Preprocessing," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 35, no. 7, pp. 1–18, 2021.
[11] K. Li and F. Chen, "Entity Extraction from Inconsistent Text Sources Using Machine Learning," *IEEE Intelligent Systems*, vol. 36, no. 3, pp. 22–30, 2021.
[12] B. Hassan and L. Kaur, "Document Layout Analysis for Accurate Information Extraction," *IEEE Transactions on Image Processing*, vol. 29, pp. 108–119, 2020.
[13] P. Reddy and K. Naidu, "Hybrid OCR–NLP Model for Extracting Key Fields from Business Forms," *International Journal of Computer Science Trends and Technology*, vol. 9, no. 1, pp. 74–82, 2021.
[14] S. Thomas and N. Prakash, "Semi-Structured Data Transformation Using Logical Consistency Verification," *International Journal of Data Engineering*, vol. 6, no. 4, pp. 88–97, 2022.
[15] D. Zhang and H. Xu, "Data Normalization and Validation Techniques for Semi-Structured Documents," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 52, no. 1, pp. 41–52, 2022.
[16] A. Fernando and L. Mathew, "Automated Extraction and Validation System for Financial Documents," *International Journal of Applied Engineering Research*, vol. 15, no. 9, pp. 1021–1030, 2020.
[17] M. Roy, S. Dey, and A. Hazra, "Intelligent Parsing and Information Verification Using Text Mining Techniques," *Journal of Web Intelligence*, vol. 18, no. 2, pp. 145–159, 2020.
[18] G. Walker, P. Smith, and E. Jones, "Assessing Reliability of Automated Document Processing Models," *IEEE Transactions on Software Engineering*, vol. 47, no. 8, pp. 1523–1538, 2021.
[19] S. Jena and K. Swain, "Deep Learning Models for Semi-Structured Document Interpretation," *IEEE Access*, vol. 8, pp. 215634–215645, 2020.
[20] R. Raman and A. Suresh, "Pattern-Based Extraction and Verification Framework for Enterprise Records," *International Conference on Data Science and Intelligent Systems*, pp. 321–328, 2021.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, December 2025
Impact Factor: 7.67

ABOUT AUTHORS

	**Name:** Sangeetha L **USN:** 3VC22CS148 **Phone No:** 7795246927 **Email:** sangeetha230105@gmail.com
	**Name:** S M Sushmitha **USN:** 3VC22CS140 **Phone No:** 7795087957 **Email:** sushmithasm31@gmail.com
	**Name:** Bindu P **USN:** 3VC22CS028 **Phone No:** 8660974576 **Email:** ramanathanpujar54@gmail.com
	**Name:** Sanjana Y **USN:** 3VC22CS150 **Phone No:** 9008755309 **Email:** veeriy005@gmail.com

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-30506