Intelligent Framework for Automated Structured Data Generation from Unstructured Text, Images, and PDF Documents

TUSHAR T S
Tusharts291421@gmail.com
Student. Department of Computer Engineering,
University B D T College of Engineering,
Davangere, Karnataka, India

Naveen Kumar B
navinaveen_b@yahoo.co.in
Assistant Professor
Department of Computer Engineering,
University B D T College of Engineering,
Davangere, Karnataka, India

How to Cite this Article:

S, T. T. (2026). Intelligent Framework for Automated Structured Data Generation from Unstructured Text, Images, and PDF Documents. International Journal of Creative and Open Research in Engineering and Management, 02(6). https://doi.org/10.55041/ijcope.v2i6.119

https://doi.org/10.55041/ijcope.v2i6.119

Abstract: The rapid growth of digital information has led to the generation of large amounts of unstructured data in formats such as text documents, images, and PDF files. Since this type of data does not follow a fixed structure, extracting useful information from it remains a significant challenge for conventional data processing systems. This paper presents a unified framework for transforming unstructured content into structured and machine-readable data. The proposed approach combines Natural Language Processing (NLP) and Optical Character Recognition (OCR) techniques to handle information from textual documents, scanned records, images, and PDF files. NLP methods are employed to identify important entities, keywords, and contextual information from textual content, while OCR technology is used to extract text embedded within images and document files. The extracted information undergoes preprocessing, cleaning, and normalization before being organized into structured formats such as JSON and CSV. The developed framework improves data accessibility, simplifies information management, and reduces the effort required for manual data extraction. The solution can support various applications, including document management systems, business intelligence, healthcare record processing, and information retrieval platforms. Experimental evaluation indicates that the framework provides an efficient and reliable method for converting heterogeneous unstructured data into structured datasets suitable for analysis and decision-making.

Keywords— Unstructured Data, Structured Data Generation, Natural Language Processing, Optical Character Recognition, Document Processing, Data Extraction, Information Retrieval, Artificial Intelligence.

INTRODUCTION

The continuous growth of digital technologies has resulted in the generation of vast amounts of information from sources such as websites, online platforms, electronic documents, images, scanned records, and PDF files. A large proportion of this information is available in unstructured form, meaning that it does not follow a fixed format or predefined database structure. As a result, managing, searching, and analyzing such data becomes a complex task for traditional information systems. Unstructured data contains valuable information that can support decision-making, knowledge discovery, and business operations. However, extracting useful content from raw documents often requires significant manual effort. Organizations dealing with large collections of reports, forms, invoices, research documents, and image-based records face challenges related to data accessibility, consistency, and efficient information retrieval. Manual extraction methods are not only time-consuming but may also introduce errors and inconsistencies. Recent developments in data processing technologies have provided effective methods for handling unstructured information. Natural Language Processing (NLP) enables the identification of important terms, entities, and contextual relationships from textual content. Similarly, Optical Character Recognition (OCR) allows text embedded in images, scanned documents, and PDF files to be converted into editable and searchable formats. These technologies make it possible to automate information extraction and improve the usability of digital content. This research presents a framework for converting unstructured data into structured and machine-readable formats. The proposed system focuses on processing textual documents, images, and PDF files through an integrated workflow. The framework automatically identifies the input type, performs the necessary preprocessing operations, extracts relevant information, and organizes the output into structured formats such as JSON and CSV. This approach simplifies data storage, retrieval, and analytical processing.

The primary objective of the proposed framework is to improve the efficiency of information extraction while minimizing manual intervention. By transforming unstructured content into structured datasets, the system supports faster access to information and enhances data management capabilities. The resulting structured data can be utilized in areas such as digital document management, business analytics, healthcare record processing, academic research, and information retrieval systems.

II. LITERATURE SURVEY

The conversion of unstructured data into structured formats has become an important research area due to the rapid growth of digital information across multiple domains. Researchers have proposed various methods using Artificial Intelligence (AI), Natural Language Processing (NLP), Optical Character Recognition (OCR), Machine Learning (ML), and Deep Learning techniques for extracting meaningful information from heterogeneous data sources.

A study titled “Explicit and Implicit Section Identification from Clinical Discharge Summaries” (2022) focused on processing medical discharge summaries using Natural Language Processing techniques. The researchers used machine learning-based text segmentation and section identification methods to organize clinical documents into structured formats. The system improved medical information retrieval and healthcare document analysis. However, the work was limited mainly to textual clinical data and did not support multimedia inputs such as audio or video.

Kaixin Sun et al. (2025), in their work “Flexible Data Extraction from Unstructured Measurement Reports using a Template-Driven Approach,” proposed a template-based extraction framework for processing measurement reports. The system used Template Definition Files (TDF), coordinate-based mapping, and regular expression techniques to extract information from Word documents automatically. Although the approach provided good accuracy for predefined templates, it lacked flexibility for handling highly dynamic and multimodal unstructured data.

Another important contribution was made by Xiangfeng Liu et al. (2024) in the research “Research and Applications of Large Language Models for Converting Unstructured Data into Structured Data.” The study

utilized transformer-based Large Language Models (LLMs) with self-attention mechanisms for structured information extraction. The authors also explored multimodal processing using deep learning approaches. Their work demonstrated the effectiveness of AI-based semantic understanding but required large computational resources and extensive training datasets.

Irena Valova et al. (2025) proposed an OCR-based framework in "Automatic Extraction and Analysis of Text and Stylistic Features of PDF Documents." The system used Tesseract OCR to extract textual information from PDF files and convert it into structured CSV outputs. The research improved automated document analysis but focused mainly on PDFs and image-based text extraction without integrating audio or video analysis capabilities.

Kara Schatz et al. (2023), through the BUILD-KG framework, introduced a knowledge graph-based approach for integrating heterogeneous datasets into analytics-enabled graph structures. The framework used Named Entity Recognition (NER), ontology mapping, and Neo4j graph databases to represent structured relationships among extracted entities. Although the system improved semantic data organization, it mainly focused on knowledge graph generation rather than multimodal data extraction.

Samuel Rodrigues et al. (2025), in their work "Cricket Player Performance Analysis Using Deep Learning," extracted structured information from PDFs and multimedia sports data using deep learning models. The system generated structured datasets for predictive sports analytics. The research highlighted the importance of multimodal data processing but was domain-specific and not generalized for broader applications.

III. PROBLEM STATEMENT

The increasing use of digital technologies has resulted in the accumulation of large volumes of unstructured information in formats such as text documents, images, scanned files, and PDF records. Unlike structured databases, these data sources do not follow a standardized format, making information retrieval and analysis more difficult. As the amount of unstructured content continues to grow, organizations face challenges in efficiently managing and utilizing the information contained within these documents. Conventional data processing methods are often limited when dealing with diverse document formats. Extracting relevant information manually from multiple sources requires considerable time and effort, and the process may lead to inconsistencies and human errors. These limitations reduce the efficiency of data management and hinder effective decision-making. To address these challenges, there is a need for an automated framework capable of extracting, organizing, and transforming information from unstructured sources into structured and machine-readable formats. Such a system should be able to handle different document types, identify relevant content, and generate organized datasets suitable for storage, retrieval, and analysis.

IV. PROPOSED METHODOLOGY

A. System Overview

The proposed framework is designed to transform unstructured information into structured and machine-readable data through a sequence of processing stages. The system follows a modular architecture in which each module performs a specific task, from data acquisition to structured output generation. The framework focuses on textual documents, images, and PDF files.

The major stages of the framework are:

1. Data Collection
2. Data Preprocessing
3. Content Analysis
4. Information Extraction
5. Data Organization
6. Structured Output Generation

A.1 Data Collection

The framework accepts different forms of document-based input commonly used in digital environments. These include:

- Text Documents (.txt, .docx)
- PDF Files (.pdf)
- Image Files (.jpg, .jpeg, .png)

After submission, the files are temporarily stored and routed to the appropriate processing module for further analysis.

A.2 Data Preprocessing

Preprocessing improves data quality and increases the effectiveness of information extraction techniques.

1. Text Preprocessing

For textual content, the following operations are performed:

- Conversion of text into a uniform format
- Removal of unnecessary symbols and special characters
- Token generation for text analysis
- Elimination of irrelevant words
- Text normalization

2. Image and PDF Preprocessing

For scanned documents and images, preprocessing includes:

- Image enhancement
- Grayscale transformation
- Noise filtering
- Contrast adjustment
- Resolution optimization

These steps improve the accuracy of text recognition and information extraction.

B. Text Analysis Using Natural Language Processing

Natural Language Processing (NLP) is employed to identify meaningful information from textual content. The objective is to extract important entities, keywords, and contextual information from documents.

Techniques Used

- Tokenization
- Part-of-Speech Analysis
- Named Entity Recognition (NER)
- Keyword Identification

Software Libraries

- SpaCy
- NLTK

Example

Input Text:

"John works at Google in California."

Person	Organization	Location
John	Google	California

The extracted entities can be further organized into structured datasets.

C. OCR-Based Image and PDF Processing

Optical Character Recognition (OCR) is utilized to convert text embedded within images and scanned PDF documents into editable and searchable content.

OCR Workflow

1. Document Enhancement
2. Text Region Detection
3. Character Recognition
4. Text Reconstruction
5. Data Validation

Tool Used

- Tesseract OCR

Example

Invoice Number	Date
1024	12/06/2024

The recognized text is forwarded to the information extraction module for further processing.

D. Data Organization and Structuring

Information collected from NLP and OCR modules is consolidated and transformed into structured formats. The system categorizes extracted information into predefined fields, making it suitable for storage and analysis.

Output Formats

- JSON
- CSV
- Relational Database Tables

Example JSON Output

```
{
 "document_type": "invoice",
 "vendor": "ABC Supermarket",
 "date": "12/05/2024",
 "amount": "45"
}
```

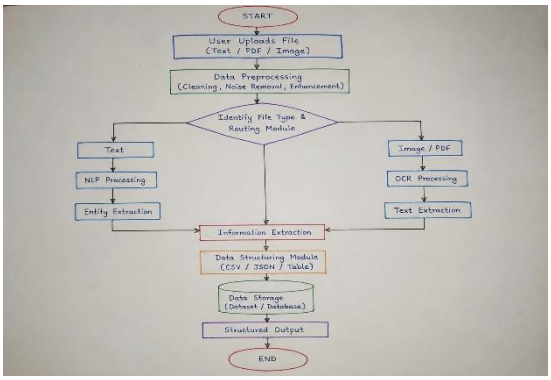
Structured outputs generated by the framework can be directly used for reporting, analytics, search operations, and decision-support applications.

E. Working of the Proposed Framework

The overall workflow begins with document acquisition, followed by preprocessing and content extraction. NLP techniques process textual documents, while OCR handles images and scanned PDFs. The extracted information is then cleaned, validated, and organized into structured formats. This approach enables efficient management of large volumes of unstructured data and reduces the dependency on manual data entry.

The proposed methodology provides a scalable and automated solution for converting heterogeneous document-based information into structured datasets that can be easily stored, searched, and analyzed.

V. SYSTEM ARCHITECTURE



The proposed framework is organized into multiple functional layers that work together to convert unstructured information into structured and machine-readable data. Each layer performs a specific operation, ensuring efficient data processing, extraction, and storage. The architecture primarily focuses on textual documents, images, and PDF files.

1. Input Layer

The input layer serves as the entry point of the system. It accepts different types of unstructured documents submitted by users, including:

- Text Files (.txt, .docx)
- PDF Documents (.pdf)
- Image Files (.jpg, .jpeg, .png)

The uploaded files are validated and forwarded to the subsequent processing stages.

2. Preprocessing Layer

The preprocessing layer improves the quality of input data before analysis. Depending on the file type, various preprocessing operations are performed.

Text Preprocessing

- Text normalization
- Removal of unwanted symbols and characters
- Token generation
- Elimination of irrelevant words

Image and PDF Preprocessing

- Grayscale conversion
- Noise reduction
- Contrast enhancement
- Image quality improvement

These operations help improve the accuracy of information extraction.

3. Data Processing Layer

This layer contains specialized modules responsible for analyzing different forms of input data.

NLP Processing Module

Processes textual information and identifies meaningful content such as entities, keywords, and contextual relationships.

OCR Processing Module

Extracts textual information from images and scanned PDF documents and converts it into machine-readable text.

The processing modules operate independently and produce standardized outputs for further analysis.

4. Information Extraction Layer

The extracted text from NLP and OCR modules is analyzed to identify relevant information. Commonly extracted elements include:

- Names of persons
- Organizations
- Locations
- Dates
- Keywords
- Numerical values
- Document-specific attributes

This layer transforms raw content into meaningful information units.

5. Data Structuring Layer

The extracted information is organized into predefined fields and converted into structured datasets. Data validation and formatting operations are performed to ensure consistency and accuracy.

Supported structured formats include:

- JSON
- CSV
- Database Tables

This organization facilitates efficient storage and retrieval of information.

6. Output Layer

The final layer generates user-accessible outputs from the processed data. Structured information can be exported, stored, or integrated with other applications for reporting and analysis.

The output layer provides:

- Downloadable structured datasets
- Searchable records
- Data summaries
- Analytical reports

By organizing extracted information into structured formats, the system enables effective data management and supports informed decision-making.

.VI. RESULTS AND DISCUSSION

The proposed framework was evaluated using a collection of text documents, images, and PDF files obtained from different sources. The system successfully extracted relevant information and converted it into structured formats such as JSON and CSV.

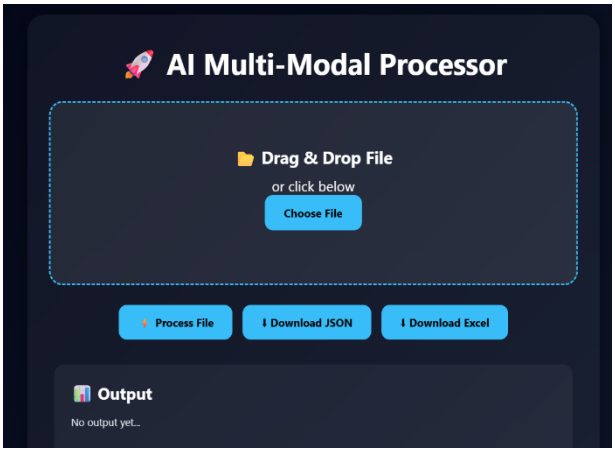


Fig1: Frame work Desktop

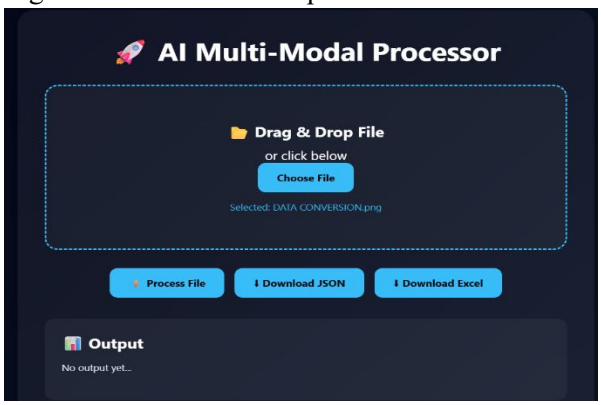


Fig 2: image Uploaded

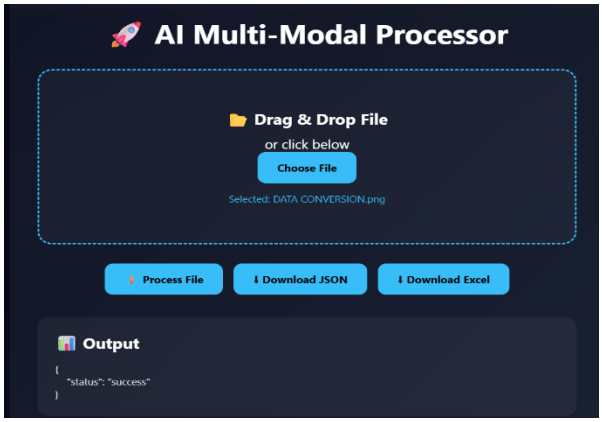


Fig 3: Fig 4: File Processed

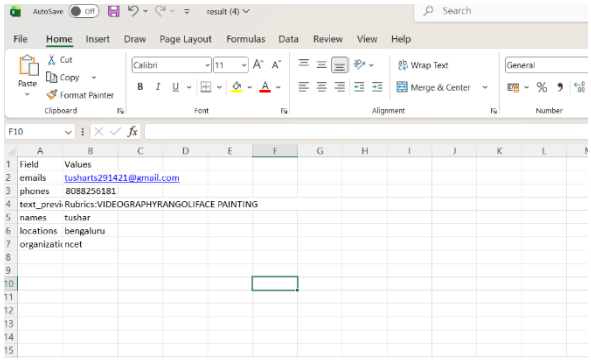


Fig 4: Converted Excel dataset

The NLP module effectively identified important entities, keywords, and contextual information from textual content. Similarly, the OCR module accurately recognized text from images and scanned documents after preprocessing operations such as noise reduction and image enhancement. The PDF processing module efficiently extracted textual content from digital documents without requiring manual intervention.

The generated structured datasets improved information accessibility and simplified data management tasks. Compared with manual extraction methods, the proposed framework reduced processing time and minimized human errors. The modular architecture also demonstrated flexibility in handling different document formats within a single workflow.

The experimental results indicate that the framework provides an effective solution for transforming unstructured information into organized datasets suitable for storage, retrieval, reporting, and analytical applications. The system can be applied in areas such as document digitization, business analytics, healthcare record management, academic research, and information retrieval systems.

VII. FUTURE ENHANCEMENTS

The proposed framework establishes a foundation for automated extraction and structuring of information from unstructured documents. Several improvements can be incorporated in future versions to enhance functionality and performance.

1. Integration of advanced language models for improved text understanding and contextual information extraction.
2. Support for multilingual document processing to handle content written in different languages.
3. Development of cloud-based deployment models for large-scale document processing and storage.
4. Implementation of intelligent document classification techniques for automatic categorization of files.

These enhancements can further improve the scalability, accuracy, and applicability of the framework across various domains.

VIII. CONCLUSION

This research presented a unified framework for transforming unstructured information into structured and machine-readable data. The proposed system combines Natural Language Processing (NLP) and Optical Character Recognition (OCR) techniques to process textual documents, images, and PDF files within a single workflow.

The framework automatically extracts relevant information, organizes the extracted content, and generates structured outputs in formats such as JSON and CSV. By reducing manual effort and improving data accessibility, the system provides an efficient approach for handling large volumes of unstructured information.

Experimental evaluation demonstrated that the framework can effectively process different document formats and produce organized datasets suitable for storage, retrieval, and analytical applications. The modular design also enables future expansion and integration with emerging technologies.

Overall, the proposed framework offers a practical solution for document digitization, information

management, business analytics, healthcare record processing, academic research, and intelligent information retrieval systems. The study highlights the importance of automated data extraction techniques in improving the usability and value of unstructured digital content.

REFERENCES

[1] K. Sambrekar, V. S. Rajpurohit, and J. Joshi, "A Proposed Technique for Conversion of Unstructured Agro-Data to Semi-Structured or Structured Data," Proceedings of IEEE ICCUBEA, 2018.

[2] N. I. Abo Dabowsa, A. M. Maatuk, S. M. Elakeili, and M. A. Ali, "Converting Relational Database to Document-Oriented NoSQL Cloud Database," Proceedings of IEEE MI-STA, 2021.

[3] I. Valova, T. Kaneva, and T. Halacheva, "Automatic Extraction and Analysis of Text and Stylistic Features of PDF Documents," Proceedings of IEEE EE&AE, 2025.

[4] K. Schatz, P.-Y. Hou, A. V. Gulyuk, Y. G. Yingling, and R. Chirkova, "BUILD-KG: Integrating Heterogeneous Data Into Analytics-Enabling Knowledge Graphs," Proceedings of IEEE BigData, 2023.

[5] S. Rodrigues, A. Mhatre, K. Kuwar, and S. Borde, "Cricket Player Performance Analysis Using Deep Learning," Proceedings of IEEE CONIT, 2025.

[6] IEEE IMCOM Authors, "Explicit and Implicit Section Identification from Clinical Discharge Summaries," Proceedings of IEEE IMCOM, 2022.

[7] K. Sun et al., "Flexible Data Extraction from Unstructured Measurement Reports Using a Template-Driven Approach," IEEE Access, 2025.

[8] Q. Zhai et al., "High Efficient Efuse Full Process Burning Solution Based on ATE," IEEE Semiconductor Testing Research, 2025.

[9] L. M. Hoi et al., "Manipulating Data Lakes Intelligently With Java Annotations," IEEE International Conference on Big Data, 2024.

[10] X. Liu et al., "Research and Applications of Large Language Models for Converting Unstructured Data into Structured Data," IEEE Research Publication, 2024.

[11] Jurafsky, D., & Martin, J. H., Speech and Language Processing, 3rd Edition, Pearson, 2024.

[12] Bird, S., Klein, E., & Loper, E., Natural Language Processing with Python, O'Reilly Media, 2009.

[13] Smith, R., "An Overview of the Tesseract OCR Engine," International Conference on Document Analysis and Recognition (ICDAR), pp. 629–633, 2007.

[14] Reimers, N., & Gurevych, I., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proceedings of EMNLP, 2019.

[15] SpaCy Documentation, Available: https://spacy.io

[16] NLTK Documentation, Available: https://www.nltk.org

[17] Tesseract OCR Documentation, Available: https://github.com/tesseract-ocr/tesseract

[18] PDFPlumber Documentation, Available: https://github.com/jsvine/pdfplumber

[19] Russell, S., & Norvig, P., Artificial Intelligence: A Modern Approach, 4th Edition, Pearson, 2021.

[20] Goodfellow, I., Bengio, Y., & Courville, A., Deep Learning, MIT Press, 2016.