IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.133144

EFFICIENT DATA ENTRY AND DOCUMENT CLASSIFICATION USING AI FOR BUSINESS ABSTRACT

Ms. Devibala Subramanian[1], Vigneshwaran. J[2]

Assistant Professor, PG & Research Department of Computer Science, Sri Ramakrishna College of Arts and Science[1]

UG-Student Final year PG & Research Department of Computer Science,

Sri Ramakrishna College of Arts and Science[2]

Abstract: This project introduces a comprehensive document processing application that encompasses text extraction and file organization functionalities. The application offers a user-friendly interface for uploading documents and provides options for both text extraction and file sorting. Utilizing libraries for PDF parsing and Optical Character Recognition (OCR), the application extracts text from PDF and JPG files. Additionally, it employs file extension-based sorting to categorize files into separate folders according to their types. The backend logic efficiently handles file manipulation tasks, including folder creation and file movement, with robust error handling mechanisms in place. Moreover, the application ensures data security and user privacy during file processing and storage. Through thorough testing and validation, the application guarantees reliability and accuracy in document processing. This project aims to streamline document management processes, enhance user productivity, and provide a seamless experience for users dealing with diverse document formats.

Keywords: DocumentProX

I. INTRODUCTION

In today's digital landscape, efficient document management is essential for individuals and organizations alike. However, handling diverse file formats and extracting relevant information from documents pose significant challenges. Existing solutions often lack comprehensive functionalities, necessitating the use of multiple tools and increasing complexity. To address these issues, we present a document processing application that integrates text extraction and file organization functionalities seamlessly. This application offers a user-friendly interface for uploading documents and employs advanced parsing and Optical Character Recognition (OCR) techniques to extract text accurately from PDF and JPG files. Furthermore, it automatically sorts documents into separate folders based on their types, streamlining the organization process.

II. RELATED WORKS

The project "Hybrid content analysis" by Baden, C., Kligler-Vilenchik, N., and Yarchi, M. in 2020 aimed to combine theoretical frameworks with computer-assisted classification techniques to analyze large text corpora. Their hybrid approach integrates traditional content analysis methods with automated tools to address challenges in analyzing vast textual data. Through this work, they sought to enhance the efficiency and accuracy of content analysis in communication research. The project "Text classification using label names only: A language model self-training approach" by Meng, Y. et al. in 2020 introduces a novel method for text classification that relies solely on label names, bypassing the need for labeled training data. Their approach utilizes self-training techniques with language models, leveraging the semantic information contained within label names to improve classification accuracy. This innovative approach has implications for various natural language processing tasks, offering a potential solution to the challenge of data scarcity in text classification.

The project titled "Capsule network algorithm for performance optimization of text classification" by Manoharan, J.S. in 2021 introduces a novel approach using capsule network algorithms to optimize the performance of text classification tasks. Capsule networks are a type of neural network architecture that aims to overcome limitations of traditional convolutional neural networks (CNNs) in tasks involving hierarchical relationships. By applying capsule networks to text classification, the research seeks to enhance accuracy and efficiency in categorizing textual data. This work contributes to advancing the capabilities of machine learning techniques in natural language processing applications.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

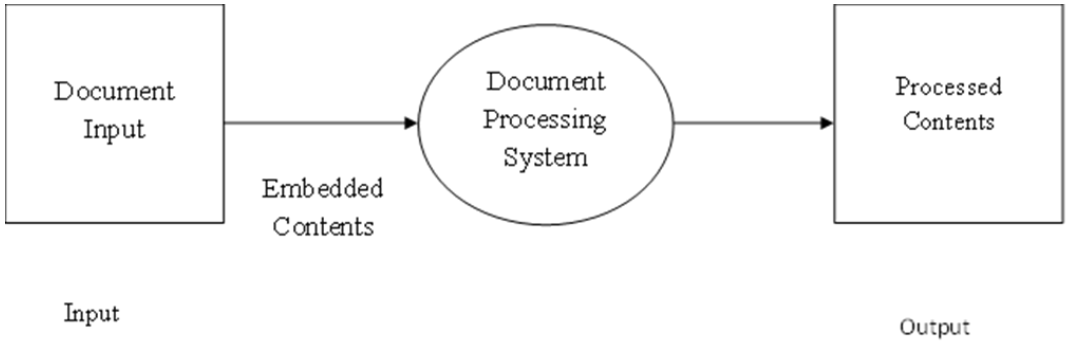
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

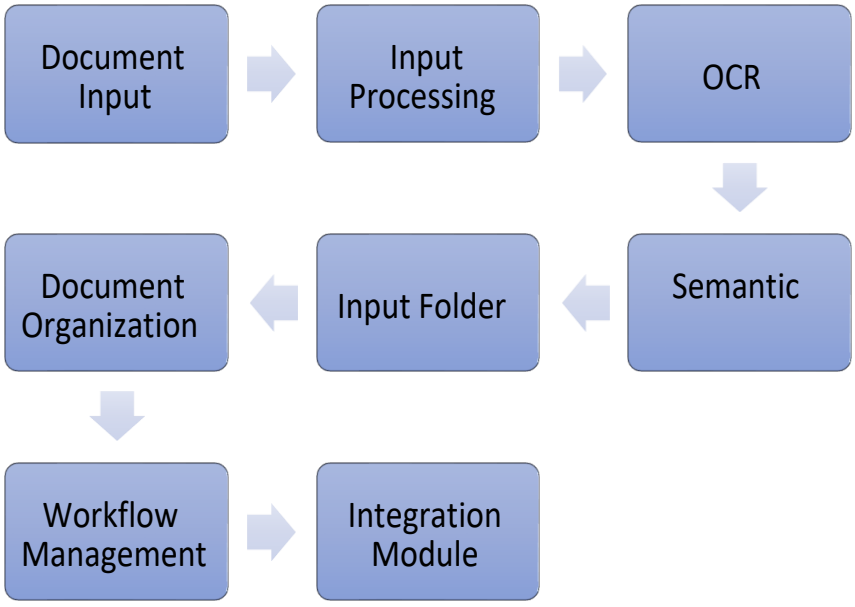
DOI: 10.17148/IJARCCE.2024.133144

III. PROPOSED METHODOLOGY

A customized document processing platform for data entry and classification aims to strengthen the integration of text extraction and intelligent organization functions. Users can upload files in various formats (such as PDF, JPG and extendable DOCX, XLSX and TIFF formats). Using the latest technology such as natural language processing (NLP) and optical character recognition (OCR), the system extracts textual content and converts scanned images into editable texts. It improves data extraction and classification using Named Entity Recognition (NER) to identify specific entities and Speech Recognition (POS) for contextual understanding. Documents are automatically categorized based on extracted items and keywords, and user-defined tags are included for easy organization. One of the system's key features will be its ability to automatically organize and categorize documents based on various criteria. Using document clustering techniques, the system will group similar documents together based on their content or characteristics, enabling efficient organization and retrieval. Moreover, metadata extraction will capture additional information about the documents, such as author, date created, and keywords, facilitating indexing and search. Users will also have the flexibility to define custom document processing workflows tailored to their specific requirements, enabling fine-grained control over document processing pipelines and allowing for quick setup and configuration.



Context Diagram



Data F low diagram

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.133144

IV. RESULTS AND DISCUSSION

- The system should categorize documents accurately based on their content, assigning them to the appropriate predefined categories.
- The system should accurately classify the sentiment of the text and correctly identify named entities present in the document.
- The extracted text from the PDF document should match the original text, indicating accurate text extraction
- The selected PDF document should be successfully uploaded to the system and displayed in the document list.

V. CONCLUSION

In conclusion, the development of the document processing system represents a significant achievement in automating and enhancing document management and analysis tasks. Through the integration of advanced technologies such as text extraction, semantic analysis, and workflow management, the system offers users a powerful platform for efficiently processing and extracting insights from various document types. The comprehensive testing efforts, including unit testing, integration testing, validation testing, and performance testing, have ensured the system's reliability, functionality, and performance under diverse scenarios and workloads. By systematically validating each component and feature, we have established confidence in the system's ability to meet user requirements and expectations effectively. Moving forward, continuous monitoring, maintenance, and enhancement of the system will be essential to address evolving user needs, technological advancements, and emerging challenges in document processing. Overall, the document processing system stands as a testament to our commitment to innovation, quality, and user-centric design, empowering users to streamline document workflows, extract valuable insights, and make informed decisions efficiently.

REFERENCES

[1]. Manoharan, J.S., 2021. Capsule network algorithm for performance optimization of text classification. Journal of Soft Computing Paradigm (JSCP), 3(01), pp.1-9.

[2]. Lurie, F., Passman, M., Meisner, M., Dalsing, M., Masuda, E., Welch, H., Bush, R.L., Blebea, J., Carpentier, P.H., De Maeseneer, M. and Gasparis, A., 2020. The 2020 update of the CEAP classification system and reporting standards. Journal of Vascular Surgery: Venous and Lymphatic Disorders, 8(3), pp.342-352.

[3]. Turner, H., 2020. Cataloguing culture: Legacies of colonialism in museum documentation. UBC Press.

[4]. Watanabe, K. and Zhou, Y., 2022. Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. Social Science Computer Review, 40(2), pp.346-366.

[5]. Li, I., Pan, J., Goldwasser, J., Verma, N., Wong, W.P., Nuzumlalı, M.Y., Rosand, B., Li, Y., Zhang, M., Chang, D. and Taylor, R.A., 2022. Neural natural language processing for unstructured data in electronic health records: a review. Computer Science Review, 46, p.100511.