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.2024.13347

# TEXT EXTRACTOR: OCR-NER FORM FILLING AUTOMATION

**Prajwal U[1], Shodhan Kumar Shetty[2], Sujan J Acharya[3], Swapnil Shetty[4], Maryjo M George[5]**

Student, Dept. of Artificial Intelligence & Machine Learning, Mangalore Institute of Technology & Engineering, Moodabidri, India[1-4]

Assistant Professor, Dept. of Artificial Intelligence & Machine Learning, Mangalore Institute of Technology & Engineering, Moodabidri, India[5]

**Abstract**: "Text Extractor OCR-NER Form Filling Automation" is an innovative software solution designed to streamline the process of extracting text from documents, performing Optical Character Recognition (OCR) and Named Entity Recognition (NER) tasks, and automating form filling tasks. This project aims to enhance efficiency and accuracy in data extraction and form completion processes across various industries. The system leverages OCR technology to extract text from scanned documents, images enabling users to digitize and analyze textual content effectively. Additionally, it employs Named Entity Recognition techniques to identify and categorize specific entities such as names, dates, locations, and organizations within the extracted text. Key features of the application include a user-friendly interface for uploading and processing documents, robust OCR and NER algorithms for accurate text extraction and entity recognition, and automation capabilities for filling predefined form fields with extracted information. Through this project, users can significantly reduce manual data entry efforts, minimize errors associated with manual transcription, and expedite the processing of documents and forms.

**Keywords:** Text Extractor, OCR-NER, Form Filling, Automation, Software Solution, Optical Character Recognition, Named Entity Recognition, Data Extraction, Document Management, Data Processing.

## I. INTRODUCTION

Text Extractor OCR-NER Form Filling Automation is a software solution designed to streamline the process of data extraction and form filling by making use of technologies such as Optical Character Recognition (OCR) and Named Entity Recognition (NER). The goal of our project is to provide users with a robust and efficient tool that automates the time-consuming task of manually extracting information from documents and filling forms. The information from legal documents can be either scanned or directly uploaded to the application, extracting relevant information from the image and performing Named Entity Recognition (NER) to extract the entities.

The application aims to automate the entire workflow, from data extraction to form filling. The Optical Character Recognition (OCR) technology, which enables the software to accurately convert scanned documents and images into editable text. This functionality allows users to digitize paper-based documents and extract text data from them with ease. Additionally, Named Entity Recognition (NER) technology, which identifies and categorizes specific entities within the extracted text, such as names, addresses, dates, and more. Our solution employs a custom-trained spaCy pipeline for Named Entity Recognition (NER), which intelligently identifies and categorizes specific entities within the extracted text. This includes names, addresses, dates, and more, ensuring precise and accurate data extraction.

The project offers a user-friendly interface that allows users to customize and configure the extraction and form-filling process according to their specific requirements. Overall, the Text Extractor: OCR-NER Form Filling Automation project aims to revolutionize the way organizations handle data extraction and form filling tasks. By harnessing the power of OCR and NER technologies, we empower users to automate repetitive tasks, reduce manual errors, and improve productivity, ultimately leading to significant time and cost savings.

## II. LITERATURE SURVEY

In paper [1] the study aims to enhance information extraction in legal texts by developing a legal NER system. Custom Dataset was used for training NER in the Indian legal domain and experiment with various annotation tools. The resulting dataset is used to train a Spacy pre-trained pipeline for accurate legal name entity prediction.

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.2024.13347

In paper [2] a generalized neural network model called SciNER is introduced. This model is designed to recognize scientific entities within free text. Utilizing bidirectional LSTM networks, SciNER combines word embeddings, subword embeddings, and external knowledge from DBpedia to enhance its accuracy. Notably, SciNER outperforms a leading domain-specific extraction toolkit by up to 50% in terms of F1 score, while also being adaptable to new domains.

In paper [3] the application of machine learning algorithms for recognizing named entities are discussed. The study examines literature from 2018 to 2020 and identifies three approaches: machine learning, deep learning, and a combination of both. Notably, the combination of Conditional Random Field (CRF) machine learning and Bidirectional Long Short-Term Memory (Bi-LSTM) deep learning is commonly used in this context.

In paper [4] various Named Entity Recognition (NER) approaches are discussed. The study compares popular NLP libraries, including Python's SpaCy, Apache OpenNLP, and TensorFlow, based on criteria such as training accuracy, F-score, prediction time, model size, and ease of training. Notably, Python's SpaCy emerges as the top performer, offering higher accuracy and optimal results.

## III. SCOPE AND METHODOLOGY

**Aim of the project:**

Our project aims to streamline data extraction and form filling processes through the integration of Optical Character Recognition (OCR) and Named Entity Recognition (NER) technologies. By leveraging Google's ML Kit for OCR and a custom-trained spaCy pipeline for NER, we seek to automate the tedious task of manually extracting information from documents and populating forms. Our goal is to provide organizations with a robust and efficient tool that accelerates document processing workflows, reduces manual errors, and enhances productivity.

Through seamless integration of OCR and NER, we aim to revolutionize how organizations handle document processing tasks, ultimately leading to significant time and cost savings. Our solution offers a user-friendly interface and customizable settings, allowing users to tailor the workflow to their specific requirements. By empowering users to automate repetitive tasks and streamline operations, we aim to drive efficiency and improve overall organizational performance.

**Existing System**

The Existing system includes OCR Miner. This system is specifically designed to extract relevant information from scanned document images, with a focus on invoices. The primary objective of OCR Miner is to automate the extraction of indexing metadata from (semi-)structured documents, particularly invoices. Handling invoices manually can be time-consuming and prone to errors. OCR Miner addresses this challenge by combining text analysis techniques with layout features.

It analyzes both the textual content extracted from the scanned image and the spatial arrangement of elements on the page. By doing so, OCR Miner enhances accuracy and efficiency in extracting critical information such as invoice numbers, dates, vendor details, and line items. Its applications extend to automating data entry, organizing invoices, and improving overall efficiency in handling large volumes of scanned documents.

**Proposed System**

Our project aims to develop a comprehensive solution for automating document processing tasks through the integration of Optical Character Recognition (OCR) and Named Entity Recognition (NER) technologies. Leveraging Google's ML Kit for OCR and a custom-trained spaCy pipeline for NER, our software will enable users to effortlessly extract information from scanned documents and images, while accurately identifying and categorizing specific entities such as names, addresses, and dates.

The system will offer a user-friendly interface for seamless interaction, allowing users to customize extraction settings and streamline form filling processes. With its focus on efficiency, accuracy, and usability, our proposed work seeks to transform document processing workflows and enhance productivity across industries

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.2024.13347

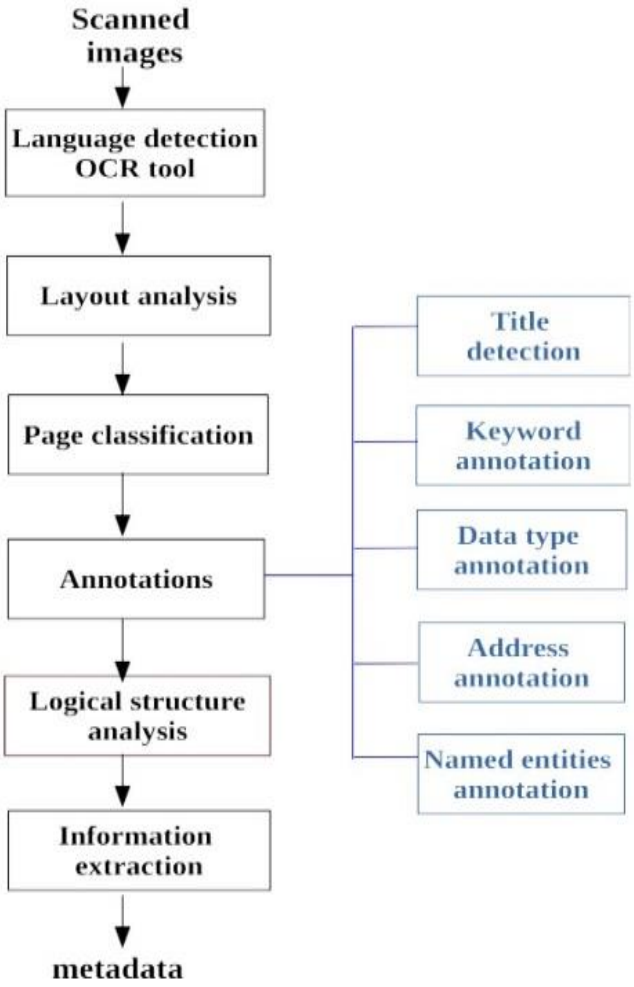


fig 1. Proposed System

**System Architecture**

The system architecture diagram illustrates the integration of Optical Character Recognition (OCR) and Named Entity Recognition (NER) technologies within our Text Extractor system. The diagram showcases the flow of data from document input to output, highlighting the interaction between the OCR engine, NER model, database, and user interface components. By visualizing the system's components and their interactions, the architecture diagram provides a clear overview of how our solution processes documents, extracts information, and populates forms, ultimately facilitating a streamlined document processing workflow.

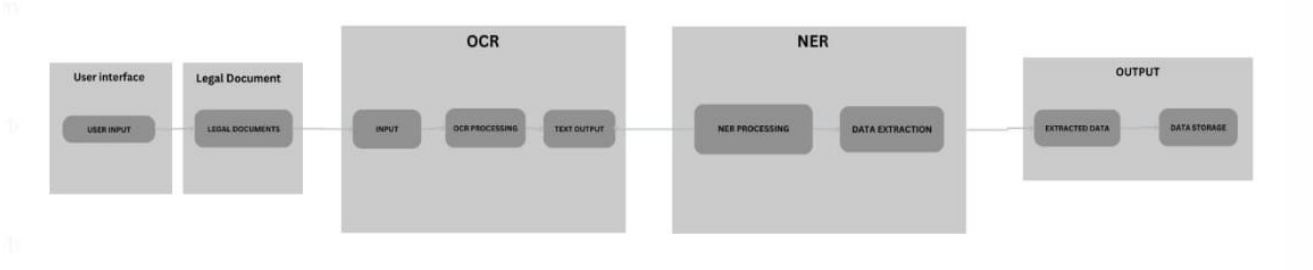


fig 2. System Architecture

  

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.2024.13347

## IV. CONCLUSIONS

This software represents a significant advancement in document processing efficiency. By seamlessly integrating OCR and NER technologies, our solution simplifies data extraction and form filling tasks, leading to improved productivity and reduced manual errors. The user-friendly interface and customizable features make it adaptable to various use cases and industries. With its ability to automate repetitive tasks and enhance accuracy, our project has the potential to revolutionize document processing workflows and drive efficiency gains across organizations.

## REFERENCES

[1]. Varsha Naik, Purvang Patel and Rajeswari Kannan, "Legal Entity Extraction: An Experimental Study of NER Approach for Legal Documents" International Journal of Advanced Computer Science and Applications(IJACSA), 14(3), 2023.
[2]. Hong Z, Tchoua R, Chard K, Foster I. SciNER: Extracting Named Entities from Scientific Literature. Computational Science – ICCS 2020. 2020 Jun 15;12138:308–21. doi: 10.1007/978-3-030-50417-5_23. PMCID: PMC7302801.
[3]. H.T. Ha, A. Horák, Information extraction from scanned invoice images using text analysis and layout features, Signal Processing: Image Communication, Volume 102,2022.
[4]. Shelar, Hemlata & Kaur, Gagandeep & Heda, Neha & Agrawal, Poorva. (2020). Named Entity Recognition Approaches and Their Comparison for Custom NER Model. Science & Technology Libraries. 39. 1-14. 10.1080/0194262X.2020.1759479.
[5]. Arora, Kawal & Bist, Ankur & Prakash, Roshan & Chaurasia, Saksham. (2020). Custom OCR for Identity Documents:OCRXNet. Aptisi Transactions On Technopreneurship (ATT). 2. 112-119. 10.34306/att.v2i2.87.
[6]. A Tool for Facilitating OCR Postediting in Historical Documents Alberto Poncelas, Mohammad Aboomar, Jan Buts, James Hadley, Andy Way, 23 Apr 2020 arXiv:2004.11471.
[7]. Haimonti Dutta, Aayushee Gupta, PNRank: Unsupervised ranking of person name entities from noisy OCR text, Decision Support Systems, Volume 152, 2022.
[8]. Naseer, Salman & Ghafoor, Muhammad & Sohaib, & Khalid Alvi, Sohaib & Kiran, Anam & Rehman, Shafique Ur & Murtaza, Ghulam & Campus, Jehlum & Jehlum, Pakistan. (2022). Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance.
[9]. Juae Kim, Yejin Kim, Sangwoo Kang,Weakly labeled data augmentation for social media named entity recognition, Expert Systems with Applications,Volume 209, 2022.
[10]. Yang, Tianjiao & He, Ying & Yang, Ning. (2022). Named Entity Recognition of Medical Text Based on the Deep Neural Network. Journal of Healthcare Engineering. 2022. 1-10. 10.1155/2022/3990563.

  