IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 7, July 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307168|

Optimizing OCR for Enhanced Image Analysis using Custom NLP-NER Models

Ananya M, Bhavani R

P.G Student, Department of MCA, The National Institute of Engineering, Mysuru, India

Assistant Professor, Department of MCA, The National Institute of Engineering, Mysuru, India

ABSTRACT: This paper presents an innovative approach to document analysis by integrating computer vision and natural language processing techniques. The proposed system enhances traditional Optical Character Recognition (OCR) methods by incorporating machine learning algorithms, particularly Named Entity Recognition (NER). This integration aims to improve pattern recognition and meaning extraction from unstructured documents across various industries. The research focuses on developing an adaptable and efficient automated system capable of processing large volumes of documents while approaching human-level comprehension in text mining. Key features include custom NER models for domain-specific entity identification, cross-industry applicability, and optimization for high-volume document processing. The potential impact of this technology spans multiple sectors, including finance, healthcare, and government administration, promising significant improvements in document processing automation and data extraction accuracy.

KEY WORDS: Computer vision, natural language processing, optical character recognition (OCR), named entity recognition (NER), machine learning, document analysis, automation, adaptability.

I. INTRODUCTION

This work centers on the main problem of how to extract information from different source documents effectively and efficiently within the factors of modern digital environment. While today's OCR solutions are capable of processing document content, some fail to sense the different layouts and formats thus making unstructured documents a challenge. This paper introduces a new solution connected to the relation between OCR and state-of-art NLP methods including Custom NER models. The principal goals are to: increase the efficiency of OCR text recognition, establish the general framework adapted to different types of documents, and upgrade the speed and quality of information extraction in the most provable spheres. This integrated system has a vision of addressing issues like the flexibility to handle different layouts of a document, better efficiency to capture relevant information from the documents and higher speed while dealing with large documents. Having an outlook in finance, healthcare and government administration, this research aims at enhancing business processes, aiding decision-making, and enhancing analysis of less explored document archives.

II. LITERATURE SURVEY

[1]This paper proposes the usage of a BERT-based Named Entity Recognition (NER) model for improved accuracy in entity recognition from unstructured documents. The type of methodology applied in the study is a systematic literature review (SLR). The studied techniques include Optical Character Recognition (OCR), Robotic Process Automation (RPA), and Named Entity Recognition (NER). The tasks proposed for implementation include collecting the dataset, OCR'ing, annotation, pre-processing, and validation with the help of services such as Google Vision API and UBIAI annotation tools. The challenge or limitation that the research faced includes problems of data diversity and quality, whereas the future enhancement includes the non-use of templates when constructing the model and enhanced pre-processing techniques.**[2]** Based on the reading activity, the text in an image is detected, particularly invoices and printed documents in the context of introducing OCR and NLP for such uses. It is planned for automation of extraction with the help of Spacy, YOLO, and PaddleOCR for various objectives. Regarding annotation it utilizes Roboflow whereas for preprocessing, the study employs OpenCV; in the current paper, preprocessing concentrates on text cleaning. Instead, it states the encouraging effectiveness regarding printed texts and outlines some issues like handwritten invoices and an intention to enlarge the dataset to advance further. **[3]** Regarding this, real-time ID card

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307168|

information searching with modern deep learning techniques of EAST and DNN are in the scope of this study. This uses functionalities such as auto-dropping, eliminating of the skewing and the separation of text and is also engaged in tests in age prediction and the identification of signatures. The review will say that the efficiency of the described method is higher than in previous methods but will mention the problems related to the time necessary for organization of data processing +/ the problem of the formats of the cards. **[4]** This paper is oriented at providing an indexation and search for images on the basis of neural networks for text extraction and recognition. Reducing HSV color, detecting Region of Interest, Singular Value Decomposition, and Multilayer Perceptron for text classification. The extracted text is binarized since it is a connected component and such a text is identifiable via Optical Character Recognition or OCR. Despite these, according to the authors, there are demerits; non-horizontal text, identification of complex backgrounds, and texts moving in complex manners. For the envisaged future improvements such are the possible continuation of the work on the Compressed Domain Processing and the improvement of the OCR accuracy. **[5]** This paper proposes an original deep learning methodology for text extraction on Mauritian identity cards for the purpose of the automation of client identification. The method uses an image segmentation model U-net and performs character recognition using a CRNN model with LSTM cells. The outcomes only have issues with the prediction of composites' names and uppercase characters even though it acquires fairly good accuracy. It shows that deep learning could be used for the automatic text extraction from the identity documents in the future with some enhancements of name handling and the boosting of the rate of automatic processing.

III. PROPOSED SYSTEM

The use of automatic identification of various text parameters is served by several components in the proposed system of document analysis and information extraction. For instance, there is an input module that focuses on managing directory choices and image uploads. A preprocessing unit follows the above by resizing and enhancing images for analysis. The core analysis module covers OCR done for text extraction, QR/barcode decoding, and NER using a trained NER model. An extraction module is used to arrange and safeguard the acquired data. There is an evaluation module that determines the degree of NER's precision. The user interface module is PCEM's safety net, enabling users to conveniently work with it through result portrayal and export capabilities. Moreover, an API may be created that will allow to interface the system.

IV. METHODOLOGY

4. 1 Input Module:

The module of image processing system that lets the input of images and data by the user. The active interface is convenient, the application can be activated as a web application or a desktop application to make it easy for the user to choose and upload image files. After uploading an image, the module applies OpenCV's cv2. imread() function will be used to import the image as a multi-dimensional array and stored in the variable NumPy. This digital representation of the image is important to the subsequent processing of the image. As for the Input Module, it can be stated that it provides the necessary pre-processing of the raw image data for input into the subsequent Modules and forms the basis of the information extraction process.

4. 2 Image Processing Module:

This one is referred to as the Image Processing Module and this is expected to prepare the uploaded images in a way that will facilitate the differentiation of texts and the codes. To enhance the image they apply some principal methods which make the quality and the resolution of the picture better. Leadingly, it performs the binarization through which the image is converted into Bi and W color format which improves the display of the texts. After that, the method for noise removal is used to eliminate all kinds of noise that may interpose in the subsequent text extraction. Hence, the module also involves the layout of the document to identify the area as advanced of the Document as the text area, the line and the word. The above images preprocessing is very crucial in enhancing the image in a way that the next process of OCR will be well enhanced and accurate performance of text. In this way, the performance of the OCR is improved so as to produce high results in extraction of text.

4. 3 Optical Character Recognition (OCR) Module:

OCR Module is the core of the application of turning text that is visible to the human eye into one that is understandable by the machine. The essential job of text recognition from the processed images is accomplished in this module using EasyOCR. The OCR process involves two main steps: pre-processing in which the image is pre-

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307168|

processed in order to extract the characters, character recognition where each character in the image is recognized and separated and post-processing which involves processing the separated characters to correct errors of the text as well as structuring the text. As a result of this conversion, the document content is made ready for further analysis and breakdown in the actual electronic platform of the Jupyter notebook.

4. 4 Named Entity Recognition (NER) Module:

Combined with the NER Module, it provides an additional layer of enhancements to the extracted text, notably, the classification of the named entities present in it. Using such a model as BERT this module identifies such components as names, dates, locations, and other components depending on the domain. It specializes for specific domains to improve its ability in identifying the entities of interest. In limiting the NER Module to pre-categorized entities, the system's overall comprehension of the text is greatly improved thus setting the stage for advanced Information Retrieval techniques and Information Analysis**.**

4. 5 Search Functionality Module:

The Search Functionality Module serves to enhance the availability of the extracted and processed information by the users. It enables the indexes of the extracted text as well as recognized entities through the use of indexing tools such as Elasticsearch or Solr. By this indexing process, fast and rapid searching is made possible in the processed documents. Towards this end, people can enter search queries in order to quickly and efficiently receive the desired information even from a mass of analyzed text. It highly improves the necessity and functionality of the system because the users find the necessary information quickly and efficiently thanks to the work of this module, which is vital for real-life applications of the document analysis system.

4. 6 User Interface Module:

This is the User Interface Module developed using Streamlit allow the user to easily work with the entire system. It includes a convenient GUI in which the users can approve the images, as well as preview the extracted text and the recognized entities, and use search options regarding the processed information. The module does a good job in giving perspectives on the complications of the backend work in a manner that could easily be understood by the end-user. The User Interface Module is relatively important since it enables users to conveniently interrelate with the available facilities in the system ranging from input to the visualization of the output, it therefore serves as the user friendly interface for the complex image processing and text analysis facilities.

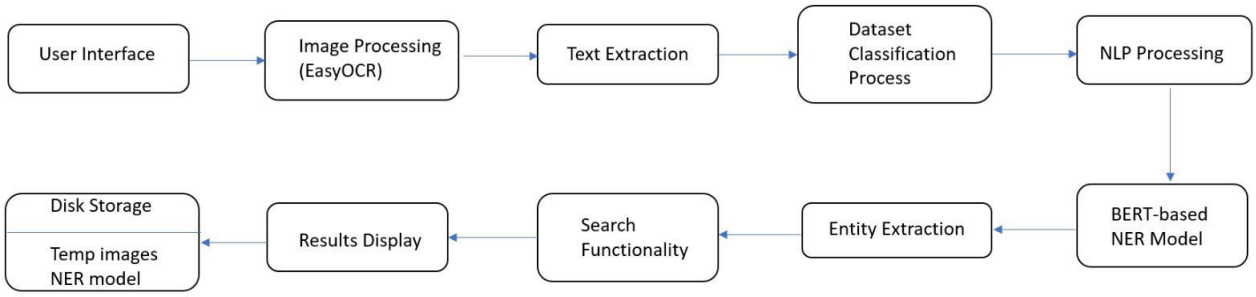

Figure 1: System Architecture

V. RESULT ANALYSIS

5.1 Text Extraction Module Analysis

The text extraction and display resume that used OCR bases presents good results from attached high resolution image of scanned documents. It demonstrates low accuracy when working with low-quality or less clear images, notifies about extracted text errors. The system helps interpret the file type and, if it is not supported, it prevents processing the file. This module shows scalability with reference to the different qualities of images while ensuring that the system is well protected through the right measures of error handling.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307168|

5.2 Named Entity Recognition(NER) Module Analysis

As it concerns the dataset type, the NER module takes processed text extract and categorizes named entities in them. In this work, it was evaluated on several datasets, proving that the proposed method achieved high classification accuracy and determining that the right NER model is used. Based on the above figures, the implemented module manages to properly fetch and present the named entities in a format that is compatible with the Streamlit application. Further, the program's search capability by key-value pairs enables users to look for particular keys and return the correct value for that key. All in all, in regard to the NER module, the performance across datasets is pretty stable, preserving the high level of accuracy in the recognition of named entities and in data search.

5.3 System Testing Results

In system testing, the following test cases were performed to ascertain the working of the OCR and NER processing system. Some of the observations made during the evaluation were successful image upload and display along with successful text extraction from well lit images and the ability to correctly handle cases where the image uploaded was not supported or was of low quality. The system correctly assigns datasets and perform text analysis using the trained NER model and shows the named entities. Besides, the regime of search works fine and provides accurate results of the key-value pairs according to the input data.

VI. CONCLUSION

By enhancing context comprehension and data extraction accuracy, OCR technology combined with specially trained NLP and NER models improves picture analysis. Text pretreatment, OCR extraction using technologies like Google Cloud Vision, and picture preprocessing for text clarity are all part of the process. The NER model then uses NLP for context to extract certain items. Lastly, post-processing verifies entities and fixes OCR problems. Accurate data extraction from complicated documents is crucial in the legal, medical, and financial sectors. This technique guarantees exact information retrieval for these businesses.

ACKNOWLEDGEMENT

I would like to thank Bhavani R, my guide, for her constant support and The National Institute of Engineering, Mysuru's dedicated staff. I am grateful to my friends and classmates for their support, and to my parents for their priceless encouragement. Lastly, I sincerely thank everyone who contributed, directly or indirectly, to the successful completion of my project.

REFERENCES

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805
2. Shagun Davessar, "Data Extraction using Optical Character Recognition and Natural Language Processing", doi.org/ JETIR2312486.
3. Niloofar Tavakolian, et. al. "Real-time information retrieval from Identity cards'', doi.org/ arXiv:2003.12103v1.
4. C. Misra, et. al. "Text Extraction and Recognition from Image using Neural Network", doi.org/ 10.5120/4927-7156.
5. Geerish Suddul, et, al. "A custom-built deep learning approach for text extraction from identity card images", doi.org/ 10.11591/ijict.v13i1.pp34-41

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details