High Level Design

Document Image Analysis

Document Name:

High Level Design for Document Image Analysis

Team Leads:

Garima Narang
Mrinal Bhardwaj
Nehaa Bansal
Shyam Mittal

Date: 17 May 2021

Contents

Introduction:

Executive Summary

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

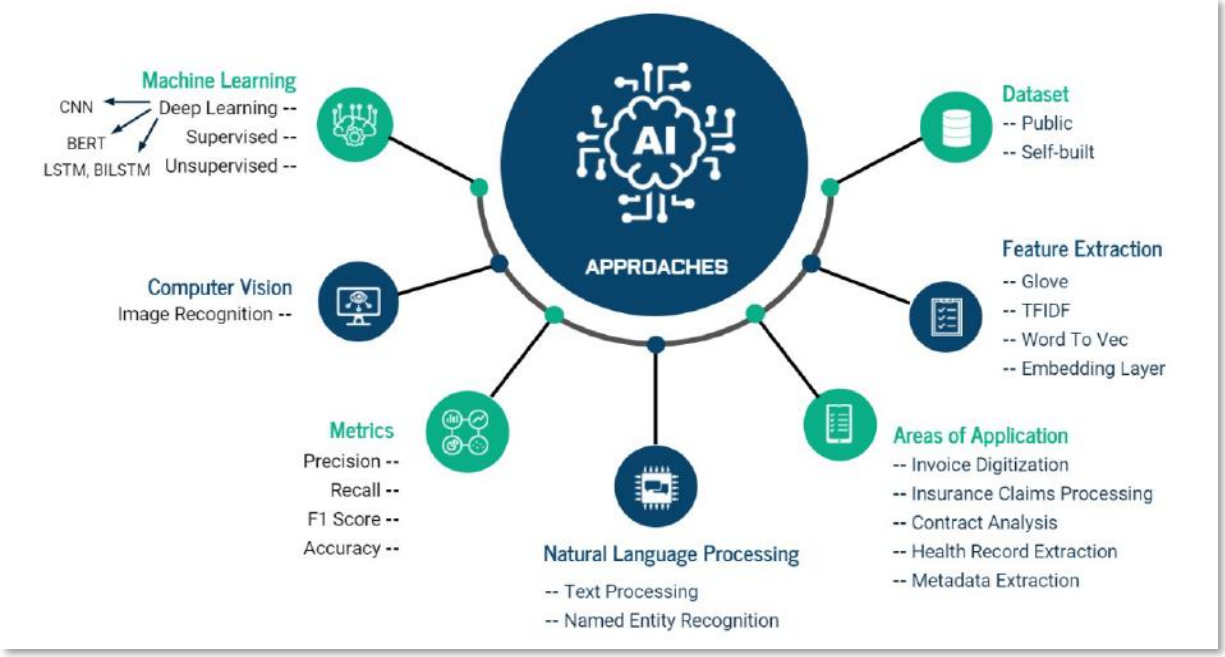
b. Problem Statement

Document Image Analysis

Problem: The unstructured data impacts 95% of the organizations and costs them millions of dollars annually. Forbes statistics states that analyzing unstructured data is an issue for 95% of business organizations, as they do not have the required expertise to deal with unstructured data. Over 150 trillion gigabytes (150 zettabytes) of unstructured data would need to be analyzed by 2025. The organizations can use data analysis tools to better understand the customer needs and forecast market variations.

Observation:

Automation helps organizations to organize and access useful information in a structured manner. Automation of the unstructured data stored in the digital format would allow the organizations to quickly gain insight into their businesses, increase their competitive edge, improve their productivity, and make innovations. The organizations thus adapt to the automation solutions by understanding the importance of Artificial Intelligence-based (AI-based) technologies such as Computer Vision (CV) and Natural Language Processing (NLP). AI technologies can understand and classify unstructured data like text, images, and scanned documents, better than traditional information extraction methods



2. Project Outline

- Significance and relevance
- Evolution Of information extraction Techniques
- Prior Research
- Research methodology, Goals.
- Quality Assessment Criteria
- Data Extraction
- Results
- Data Validation Techniques
- Background on Different AI approaches
- Comparison of OCR and AI approaches
- Discussion
- Limitations of the study
- Future Work and opportunities

3. High Level Design:

Documents are a core part of many businesses in many fields such as law, finance, and technology among others. Automatic understanding of documents such as invoices, contracts, and resumes is lucrative, opening many new avenues of business. The objective of document image analysis is to recognize the text and graphics components in images, and to extract the intended information as a human would. Document image processing and understanding has been extensively studied over the past forty years. Work in the field covers many different areas including preprocessing, physical and logical layout analysis, optical and intelligent character recognition (OCR/ICR), graphics analysis, form processing, signature verification and writer identification, and has been applied in numerous domains, including office automation, forensics, and digital libraries. Several good solutions exist for document processing and analysis, this paper tries to give a general idea for document processing and the various steps/methods used for that. Data is extracted in the end and stored in a particular format like JSPN, XML etc as per product vision.

Applications:

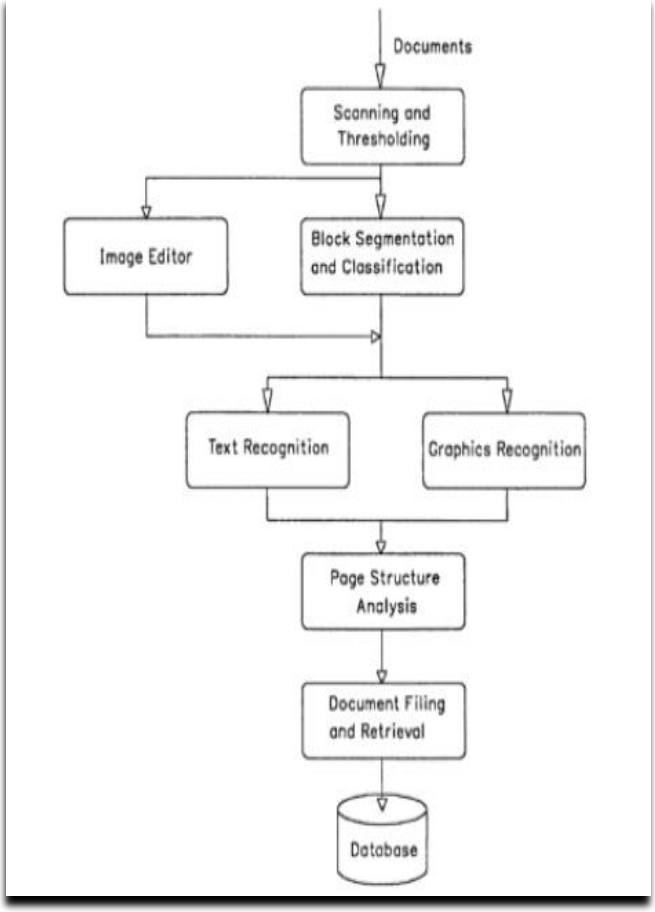
Various Application Domains for Automatic Information Extraction Techniques from Unstructured Documents. Application Areas Usage

1. Healthcare - Patient details and disease extraction from Electronic Health Record (EHRs).
2. HR - Employee details extraction and verification from their documents, Payroll process automation, Job recommendations or hiring shortlisted candidates from the collection of CVs.
3. Insurance - Automating claims processing & clearance form filling, Extraction of policy premium amount details from policy documents.

4. Travel Sector - Extracting and verifying ticket booking and traveler details.
5. Banking and Financial Services - KYC verification, Auto form-filling for cards activation, Fraud claim detection, Pattern discovery from the customer documents.
6. Government - Automating address change requests, License renewal by verifying the user details.
7. Infrastructure - Infrastructure related document processing and information extraction from documents, Invoice, and receipt digitization.
8. Legal - Contract element extraction from legal documents, identifying clauses, and involved risks.

Workflow:

The major components of a document analysis system are as follows:



- Scanning and Thresholding – Scan the document by optical means and convert the signal to digit form and then threshold the scanned image into bi-level format.
- Block Segmentation and classification – Segments an image into blocks automatically and classify each block as text, graphics, or pictures.
- Image Editor – Edit manually a scanned document into regions of text, graphics, or pictures.
- Text Recognition – Process text blocks and recognize individual characters.
- Graphical representation – Convert block segments of graphics from bitmap format to graphic symbols e.g.: vectors or high-level graphics description.
- Page Structure Analysis – Analyze spatial relationships between blocks and/or textual contents to detect page-structure components like heading, summary, paragraphs etc.
- Document filing and retrieval – Create document indices either manually or automatically by filing and provide a query interface for document retrieval from a database.
- Database - The extracted data relevant from the document are saved as per the product vision.

In this document we will discuss two methods for Document Image Analysis. One is **optical character recognition (OCR)** and the other is **LayoutParser** library.

1) Optical Character Recognition (OCR)

OCR involves two main stages. The first stage is text detection/localization in which the textual part within the image is located. This text localization within the image is essential for the second stage of OCR, text recognition, in which the text is extracted from the image. We will now discuss the common steps in OCR as shown in Figure 15

a: Text detection or localization techniques

Text detection methods are necessary to identify or locate the text within the complete image and draw a bounding box over the portion or area of the image, consisting of textual contents. The text detection techniques are classified into conventional text detection methods and text detection using Deep Learning methods. Pre-processing is an essential step in text detection. Typical pre-processing includes the following stages:

- Binarization: In binarization, the grayscale images are transformed into binary images. OpenCV offers binarization via adaptive thresholding, simple thresholding, and Otsu's binarization.
- Noise removal: Scanned documents frequently consist of noise caused by the printer, scanner, and print quality. OpenCVPython can be used to get rid of such noise, such as salt & pepper noise and Gaussian noise.
- Skew angle detection and correction: When a document is scanned, either automatically or by a person scanning a document, a slight tilt (skew) to a document is obvious.
- Line-level, word-level, and character-level segmentation: Segmentation divides the entire image into subimages to process them further. The most popular techniques used for image segmentation are: X-Y-tree decomposition, connected component labeling, Hough transforms , and histogram projection techniques.
- Thinning: Thinning aims to decrease the image parameters to its minimum necessary information, to simplify further processing or analysis and image recognition.

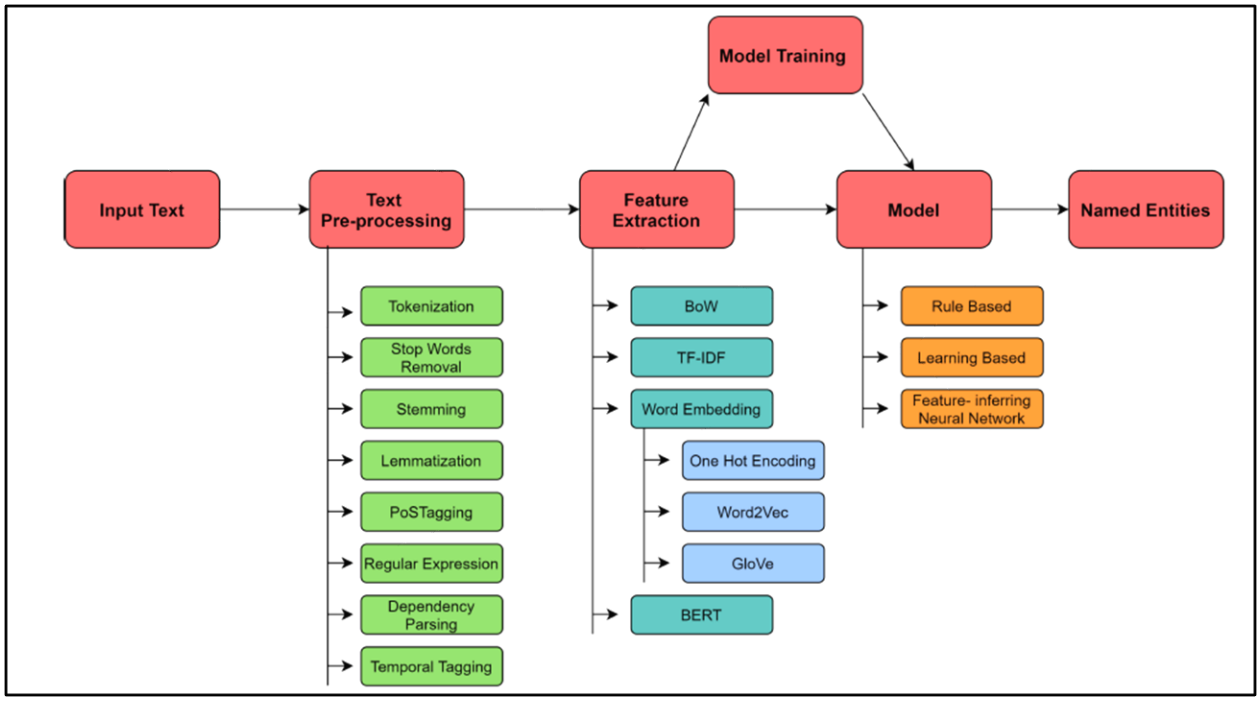
b: Text Recognition The next step is text recognition. In text recognition, the text characters are converted into various character encoding formats such as ASCII or Unicode. It can be performed in two steps (a) feature extraction and selection (b) classification

In most OCR techniques, an algorithm learns to categorize or classify the character set and numerals accurately, as it is trained on a known dataset. The most popular techniques used for the classification in OCR are mentioned below:

- K Nearest Neighbour: It classifies the objects with a similar feature in proximity
- Naïve Bayes Classifier:It is a probabilistic classification method. It uses the Bayes theorem of probability to calculate the class of new or unknown data.
- Neural Network: It has shown strong abilities to automate text detection and data extraction by recognizing the underlying relationships of characters/words. Region Based Convolutional Neural Networks (R-CNN) is used for the object detection.

- Support Vector Machine: It is the commonly used classification algorithm in OCR. It performs better than any other classification method. It is not based on any assumptions of independence as in the Naïve Bayes method. It is used for the text categorization or recognition.

Post-processing : It detects and corrects the misspelling in the output text after an image is processed using the OCR technique. Once we obtain the output text, we perform NLP(Natural Language Processing) on it. It analyses the grammatical structure at the sentence level and then creates grammatical rules to obtain the useful information about the sentence structure. Among all the techniques, NER techniques serve the most basic and essential techniques in NLP. NLP uses the sentence-level syntactic rules such as assigning grammar rules and patterns at the word or token levels, such as the regular expressions for information extraction from the given text, using NER. It automatically scans the unstructured text to locate the "named entities" like a name (first name, last name), location (such as countries, cities), organization, date, and invoice numbers in the text. Find below the NER workflow.



2) Layout Parser

A Unified Toolkit for Deep Learning Based Document Image Analysis. With the help of state-of-the-art deep learning models, Layout Parser enables extracting complicated document structures using only several lines of code. This method is also more robust and generalizable as no sophisticated rules are involved in this process.

There are basically two functionality that this parser provides:

1. **Table OCR and Results Parsing**: layoutparser can be used for conveniently OCR documents and convert the output into structured data.

2. **Deep Layout Parsing Example**: With the help of Deep Learning, layoutparser supports the analysis of very complex documents and processing of the hierarchical structure in the layouts.

The goal is to extract the text with the help of LayoutParser which uses detectron2 model at the backend.

Dataset Used For Training :

ICDAR 2019 Scanned Receipts OCR and Information Extraction(SROIE) Dataset
This dataset consists of 1000 annotated images in total for training and testing. The training and validation datasets for the SROIE competition tasks have 600 annotated images and can be downloaded from the following link:
https://drive.google.com/open?id=1ShItNWXyiY1tFDM5W02bceHuJjyeeJl2

Technical Info:

There are a lot of optical character recognition software available:

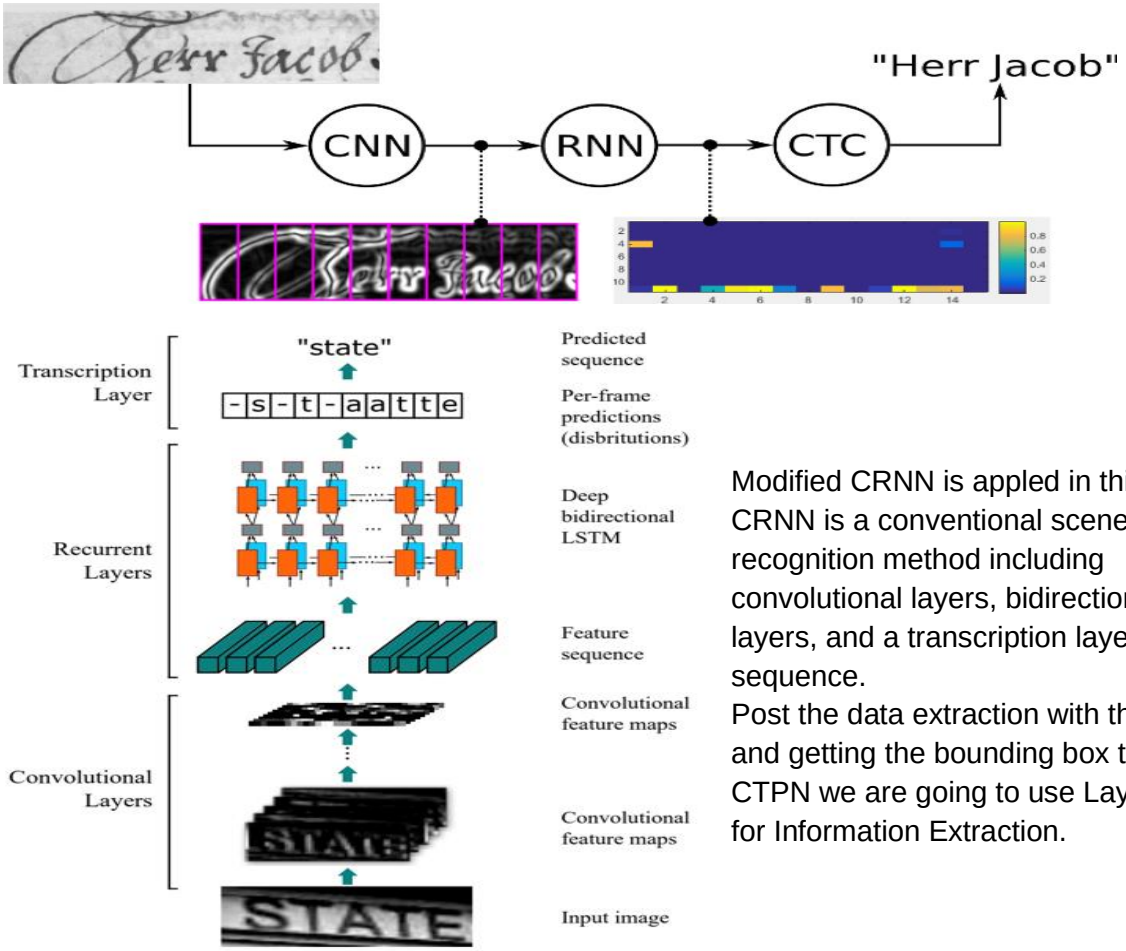
- Tesseract - an open-source OCR engine that has gained popularity among OCR developers. In 2005 HP released Tesseract as an open-source software. Since 2006 it is developed by Google.
- OCRopus - OCRopus is an open-source OCR system allowing easy evaluation and reuse of the OCR components by both researchers and companies. A collection of document analysis programs, not a turn-key OCR system. To apply it to your documents, you may need to do some image preprocessing, and possibly also train new models.
- Ocular - Ocular works best on documents printed using a hand press, including those written in multiple languages. It operates using the command line. It is a state-of-the-art historical OCR system.
- SwiftOCR - This OCR engine written in Swift and there is huge development being made into advancing the use of the Swift as the development programming language used for deep learning.
- Google vision OCR - Google vision is the OCR API developed by Google Cloud. This OCR uses a very powerful and pre-trained machine learning technology. Thanks to Google Vision, it is possible to assign labels to images, to read both printed and handwritten text. You can also detect, and extract objects and faces while obtaining other information about them such as position on the image
- AWS Textract - Amazon Textract is the OCR software that automatically extracts data from scanned documents and converts it into text that can be modified. However, AWS Textract goes beyond simple OCR. Beyond reading and transcribing, it does more by identifying the content in forms and information stored in your tables.

- Azure API vision - Azure API vision is an OCR API developed by the Microsoft group. This OCR focuses mainly on images. It will convert a document from PNG or JPEG format to an editable one. Thus, you can find a classification card of your image with categories such as object, keywords, description, format, colors, etc. In addition, this OCR will allow you to identify and tag the content. For example, you can use the object-detection tool to locate an object in an image.

Connectionist Text Proposal Network (CTPN)

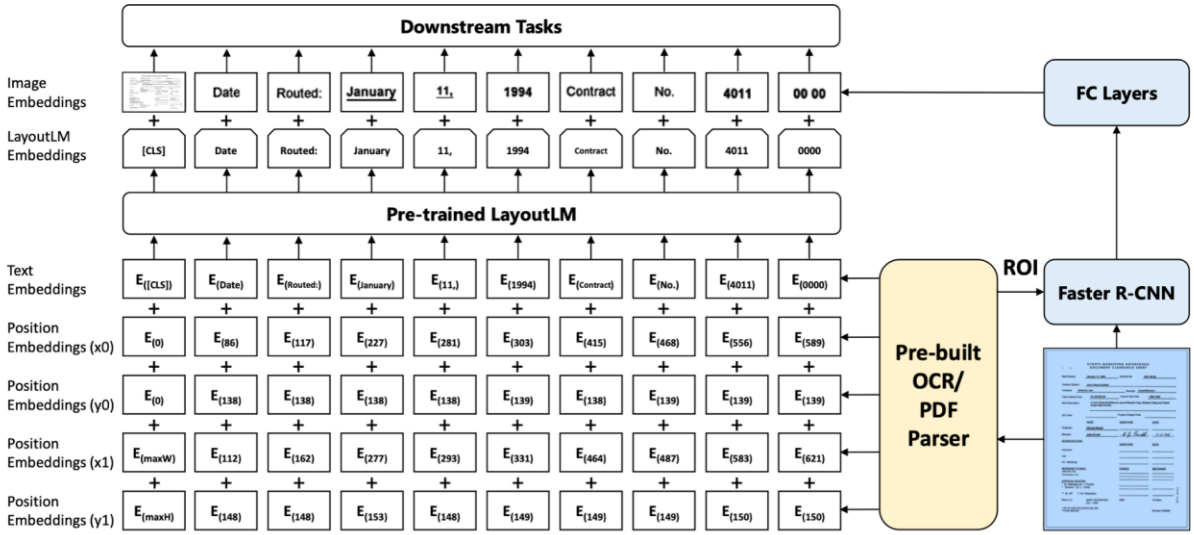
CTPN accurately localizes text lines in a natural image. The CTPN detects a text line in a sequence of fine-scale text proposals directly in convolutional feature maps. We develop a vertical anchor mechanism that jointly predicts location and text/non-text score of each fixed-width proposal, considerably improving localization accuracy. The sequential proposals are naturally connected by a recurrent neural network, which is seamlessly incorporated into the convolutional network, resulting in an end-to-end trainable model. This allows the CTPN to explore rich context information of the image, making it powerful to detect extremely ambiguous text.

Scanned Receipt OCR by Convolutional-Recurrent Neural Network



Modified CRNN is appled in this task. CRNN is a conventional scene text recognition method including convolutional layers, bidirectional LSTM layers, and a transcription layer in sequence.
Post the data extraction with the RCNN and getting the bounding box through CTPN we are going to use LayoutLM for Information Extraction.

LayoutLM



We will use this architecture to extract the relevant fields for our task.

1. Position embeddings can be attained through bounding boxes.
2. Text Embeddings can be attained with the text received via OCR.
3. Image Embeddings can be attained through Faster RCNN.

Model predicts the corresponding values using the sequence labeling method.

Related work:

If we search document image analysis on github , we see a lot of repos to develop a clear and detailed understanding of the existing automatic information extraction techniques for documents. Some other image processing techniques can be seen but none that use an extensive collection of deep learning techniques.

[1] Curtis Wigington, Chris Tensmeyer, Brian Davis, William Barrett, Brian Price, and Scott Cohen. Start, follow, read: End-to-end full-page handwriting recognition. In European Conference on Computer Vision (ECCV), pages 367–383. Springer, 2018.
[2] https://edwardbenson.com/automating-paperwork
[3] H. Arai and K. Odaka. Form reading based on background region analysis. In Proceedings of the 4th International Conference on Document Analysis and Recognition. Ulm, Germany, 1997
[4] Dr. Mehraj-Ud-Din Dar “Document image classification: A Cognition Based Approach”, J&K Science Congress University of Kashmir, 25-27 July, 2006.

Conclusion:

We presented a summary of basic building blocks that comprise a document analysis system. The document aims to explore the various extraction techniques for document. Guidelines proposed by **Ineuron team and our mentor, Prasanna Venkatesh** were adhered to conduct the literature search for this document. Our review highlights the opportunities for research in OCR, and AI-based techniques used for automatic information extraction from unstructured documents. The proposed framework aims to build the high-quality unstructured document datasets with varied and complex layouts from multiple sources, such as invoices from different suppliers, that will be publicly available to enhance future research in this domain. It helps to validate the quality of data before model training with different statistical techniques, resulting in better model performance. This project has several practical/industry implications for automatic information extraction adoption in the finance and legal sectors. The benefits of automatic information extraction adoption are clear.