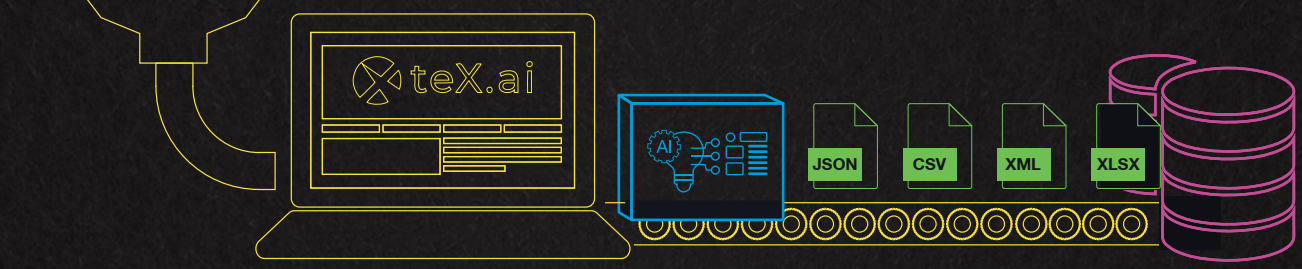
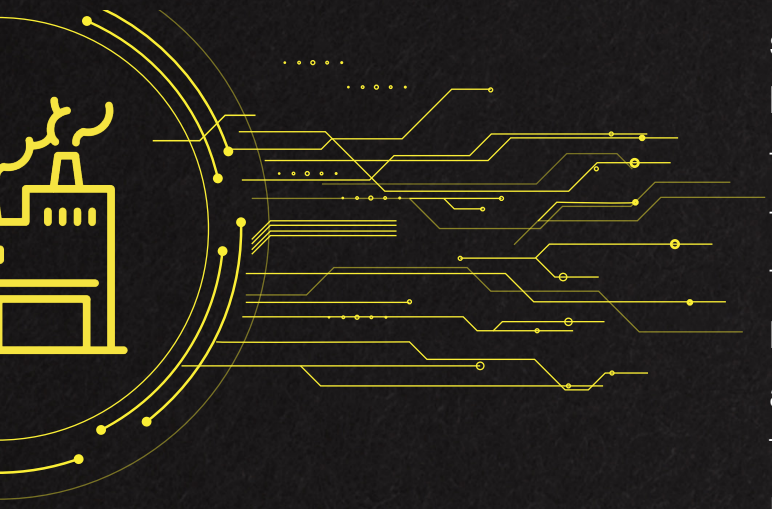
Multi-Language Text Extraction using teX.ai in the Manufacturing Industry



- 60% reduction in human intervention during the process and associated reduction in costs
- Ability to speed up the process without compromising on the accuracy
- The model exhibited an accuracy of over 80%.

teX.ai™



Customer Background

The client is a multibillion dollars firm who first started their business at German, and they have been expanding their business for more than a century. Client's products and technologies are available worldwide and they have their offices in 120 nations and locations around the world. Their products are manufactured in 3 categories: Adhesive Technologies, Beauty Care and Laundry & Home Care.

Business

teX.ai, Text extraction

Domain

Manufacturing

Technology

Python Flask, teX.ai, Java Spring Boot, Angular

Key Highlights

- 60% reduction in human intervention during the process and associated reduction in costs.
- Ability to speed up the process without compromising on the accuracy.
- The model exhibited an accuracy of over 80%.
- The front-end web portal provided an option for manual intervention in the process, if needed.

teX.ai ™

Business Requirement

- The products are manufactured via a very careful process involving numerous chemical components with meticulous calculation of its quantity.
- Each of the chemical components being used in the process and arriving at the manufacturing plants had a specific Certificate of Analysis (CoA) document, which needed to be verified for standards.
- These documents were being stored in PDF format, and given the size of the business, thousands of CoA documents were to be handled and verified. Thus, the current manual process of verification by the team of expert chemical engineers was very painstaking and not scalable. Volume of files was 1000 per day.
- Another complexity was that the documents were not only in English, but also in German, Mandarin and Thai.

Challenges

- The documents contained multiple languages apart from English.
- Presence of non-readable headers and footers in the documents.
- Redundancy of data in the PDFs where text extraction had to be done.

teX.ai™

Objective

- Text data was to be extracted from PDF documents which were for both scanned images as well as for digital PDF documents.
- A single extraction process processed multiple languages with different scripts. In other words, a single pdf document contains English and Thai or German and Mandarin etc.,
- Both Tabular and Peripheral data were required to be extracted.
- Client wanted to have a front-end web application where they themselves can upload/ edit documents.

Solution Overview

Text Conversion: To start the text extraction from the PDF documents leveraging teX.ai, it is important that the document be in text format. Tesseract OCR was used to convert all the images into text format.

Object Detection: Object Detection Neural Network algorithms were used to draw the bounding boxes around the required objects from the pdf documents.

Non-English Languages: Languages such as German, Mandarin, or Thai in the table were successfully extracted using Tesseract, Tabula and Camelot.

Frontend: Indium's Product Development team created an interactive front-end application which enabled automatic reading of PDFs from a dedicated email account which was used to send all the inputs files on a daily basis. OCR confidence were measured and given in the output.

Output: Outputs were provided both in CSV and Xml format to the client, with options to View and Download.

Editing: Clients with the Admin access were given rights to Edit and Save the output in the respective format.

teX.ai™

Input

测 试 项 目 ： 1786102　　订 单 号：4591754169

地址：上海市奉贤区南奉公路 3528 号 4 号楼　　客　户：汉高(丹阳)

电话：021-51097761 传真：021-51097762　　生产批号：WH191113

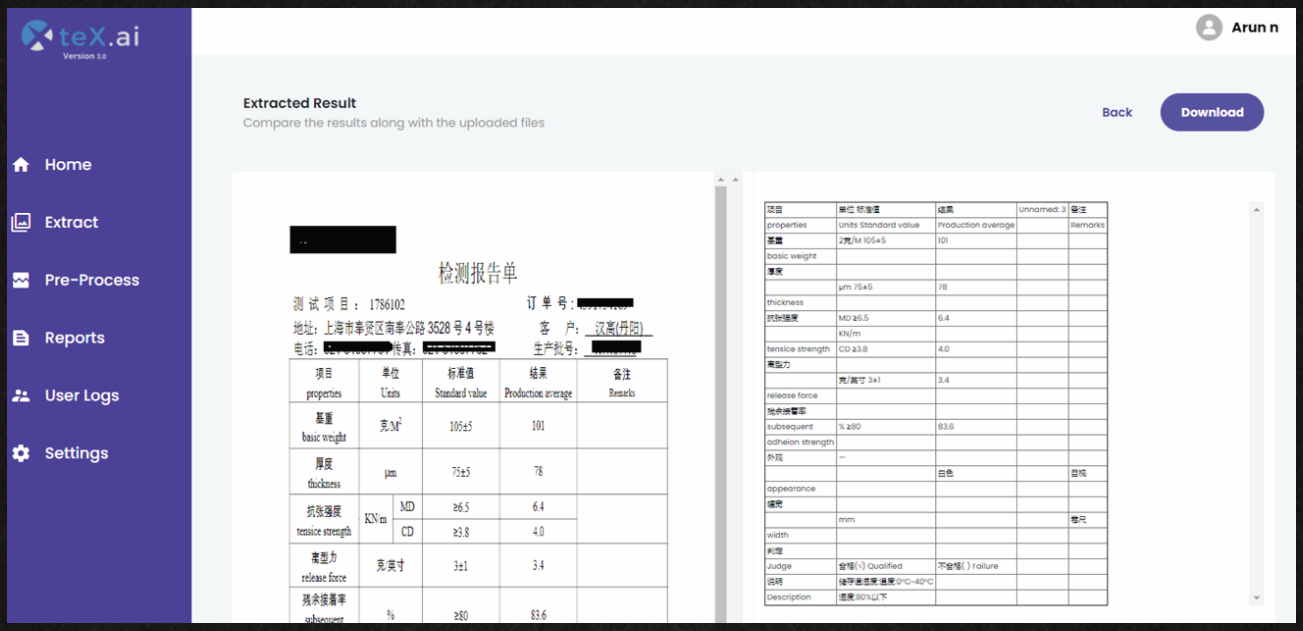
项目 properties	单位 Units		标准值 Standard value	结果 Production average	备注 Remarks
基重 basic weight	克/M^2		105±5	101	
厚度 thickness	μm		75±5	78	
抗张强度 tensice strength	KN/m	MD	≥6.5	6.4	
		CD	≥3.8	4.0	
离型力 release force	克/英寸		3±1	3.4	
残余接着率 subsequent adheion strength	%		≥80	83.6	
外观 appearance	—			白色	目视
幅宽 width	mm				卷尺
判定 Judge	合格(√) Qualified　　不合格(　) Failure				
说明 Description	储存温湿度：温度：0℃~40℃ 湿度：80%以下				

Output:

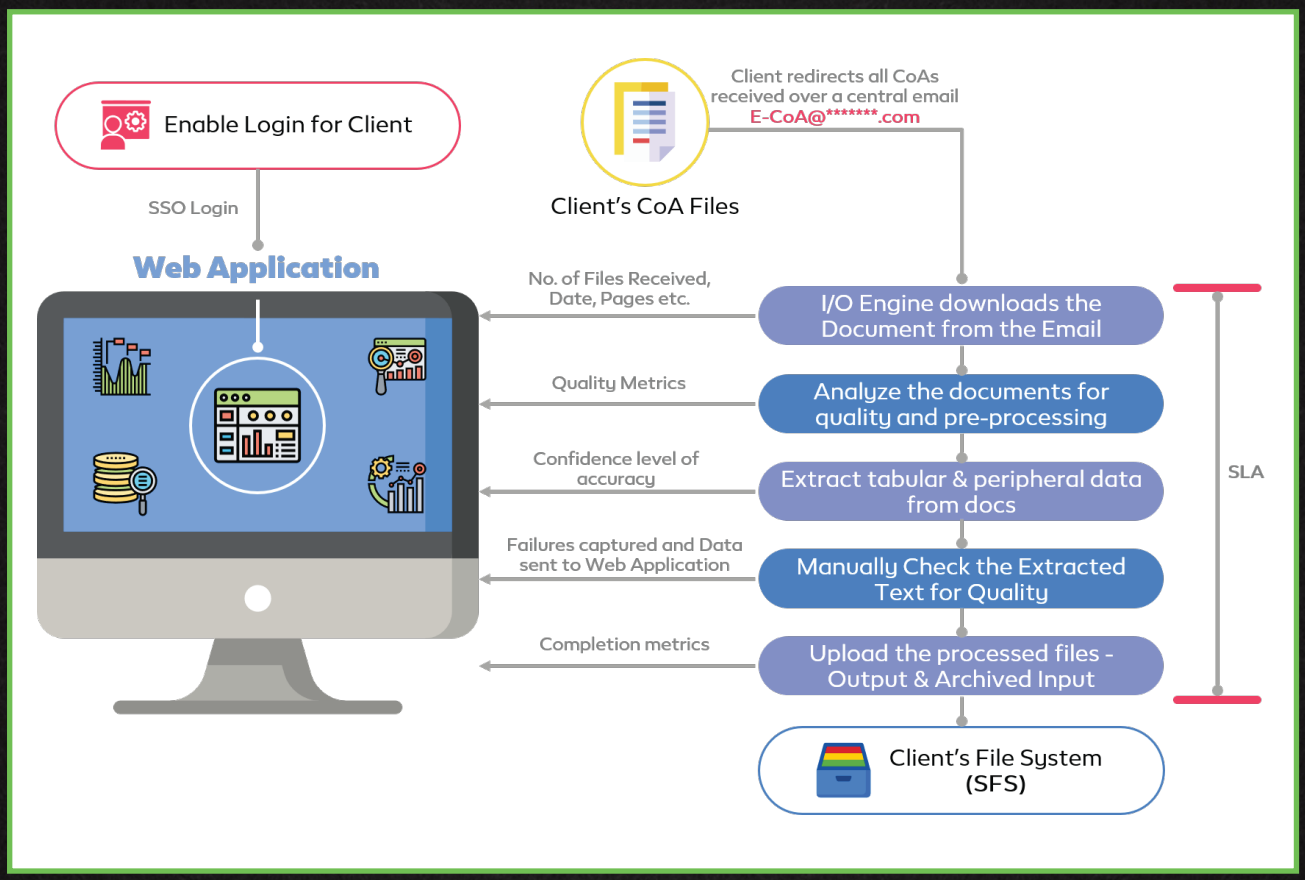
	A	B	C	D	E	F
1	col0	col1	col2	col3	col4	col5
2	项目 properties	单位 Units		标准值 Standard value	结果 Production average	备注 Remarks
3	基重 basic weight	克/M2		105±5	101	
4	厚度 thickness	μm		75±5	78	
5	抗张强度 tensice strength	KN/m	MD	≥6.5	6.4	
6			CD	≥3.8	4	
7	离型力 release force	克/英寸		3±1	3.4	
8	残余接着率 subsequent adheion strength	%		≥80	83.6	
9	外观 appearance	—			白色	目视
10	幅宽 width	mm				卷尺
11	判定 Judge	合格(√) Qualified　　不合格(　) Failure				
12	说明 Description	储存温湿度：温度：0℃~40℃ 湿度：80%以下				
13						
14						
15						
16						

teX.ai™

Output:



Process Workflow:



teX.ai ™

Business Impact

Time: teX.ai reduced the time taken to extract the data from the PDF files by 75%.

Accuracy & Validation: High level of accuracy was achieved for the text extraction process compared to their legacy method. Clients were able to efficiently validate the converted outputs by easily skimming through the front-end application as they could do a side-by-side comparison of the input and output.

Training: As the data keeps flowing the model trains itself and the AI model self-learns based on the improved edits made by the expert Chemical Engineers. This means that the accuracy and quality of the extraction would increase with use.

Get in touch to see how teX.ai can help you!

info@tex-ai.com

+1 (888) 207 5969 (Toll-free)

www.tex-ai.com

Suite 100, 19925 Stevens Creek Blvd,
Cupertino, CA — 95014

teX.ai™