SUCCESS STORY

MULTI-LANGUAGE TEXT EXTRACTION USING TEX.AI IN THE MANUFACTURING INDUSTRY

PROJECT OVERVIEW

The products undergo a meticulous manufacturing process, involving precise calculations of numerous chemical components. Each component arriving at the manufacturing plants is accompanied by a specific Certificate of Analysis (CoA) document that requires verification against established standards. Due to the large scale of the business, thousands of CoA documents, stored in PDF format, need to be handled and verified. However, the current manual verification process, performed by a team of expert chemical engineers, is time-consuming and not easily scalable. With a volume of 1000 files per day, this poses a significant challenge. Adding to the complexity, the documents are not only in English but also in German, Mandarin, and Thai. Thus, an automated solution is sought to streamline and enhance the verification process, ensuring accuracy, efficiency, and scalability.

SOLUTION DELIVERED

teX.ai

CLIENT DOMAIN

Manufacturing

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

ABOUT CLIENT

The customer, a multinational company with a rich history spanning over a century, originated in Germany and has expanded its customer base globally. With offices in 120 countries and regions, their technologies and products are widely accessible. They specialize in providing goods for the Adhesive Technologies, Beauty Care, and Laundry & Home Care industries.

BUSINESS CHALLENGES

- To check if the products undergo a meticulous manufacturing process that involves precise calculations and the use of numerous chemical components.
- Each chemical component used in the process is accompanied by a specific Certificate of Analysis (CoA) document, which must be verified against established standards.
- These CoA documents, stored in PDF format, present a significant challenge due to the large volume of files to be handled and verified. The current manual verification process by a team of expert chemical engineers is both time-consuming and not scalable. The daily volume of files reaches 1000.
- Complicating matters further, the documents are not only in English but also in German, Mandarin, and Thai.

BUSINESS REQUIREMENTS

- Text data extraction: The system needed to extract text data from PDF documents, including both scanned images and digital PDFs.
- Multilingual support: The extraction process had to handle multiple languages with different scripts. This means that a single PDF document could contain text in English, Thai, German, Mandarin, and other languages.
- Tabular and peripheral data extraction: In addition to extracting regular text data, the system was also required to extract tabular data and peripheral information from the documents.
- Front-end web application: The client wanted a user-friendly web application where they could upload and edit documents themselves.

SOLUTION HIGHLIGHTS

- The document contained multiple languages besides English. Additionally, there were non-readable headers and footers present in the documents, and there was redundancy of data in the PDFs that required text extraction, which was addressed.
- Initiated the text extraction process using teX.ai, the document needed to be in text format. Tesseract OCR was utilized to convert all the images into text format.
- For object detection, neural network algorithms were employed to draw bounding boxes around the required objects in the PDF documents.
- Non-English languages such as German, Mandarin, or Thai in the tables were successfully extracted using Tesseract, Tabula, and Camelot.
- Indium's Product Development team created an interactive frontend application that facilitated the automatic reading of PDFs from a dedicated email account. OCR confidence levels were measured and provided in the output.
- The outputs were delivered to the client in both CSV and XML formats, and gave options to view and download.
- Clients with admin access were granted the rights to edit and save the output in the respective format.

BUSINESS IMPACT

- Time: teX.ai reduced data extraction time from PDF files by 75%.
- Accuracy & Validation: The text extraction process achieved a high level of accuracy, surpassing the client's legacy method. Clients could efficiently validate converted outputs using the frontend application, allowing for easy side-by-side comparison of input and output.
- Training: The AI model self-learns and improves over time as it receives improved edits from expert Chemical Engineers, resulting in increased accuracy and extraction quality.

TECH STACK

ABOUT INDIUM

Indium Software is a fast-growing Digital Engineering company, focused on building modern solutions across Applications, Data, and Gaming for its clients. With deep expertise in next-gen offerings combining data and applications, Indium offers a wide range of services including Product Engineering, Low-Code development, Data Engineering, Ai/ML, Digital Assurance, and end-to-end Gaming services.

indium
Make Technology Work

USA	INDIA	UK	SINGAPORE
Cupertino | Princeton Toll-free: +1-888-207-5969	Chennai | Bengaluru | Mumbai | Hyderabad Toll-free: 1800-123-1191	London Ph: +44 1420 300014	Singapore Ph: +65 6812 7888

www.indiumsoftware.com

For Sales Inquiries
sales@indiumsoftware.com

For General Inquiries
info@indiumsoftware.com