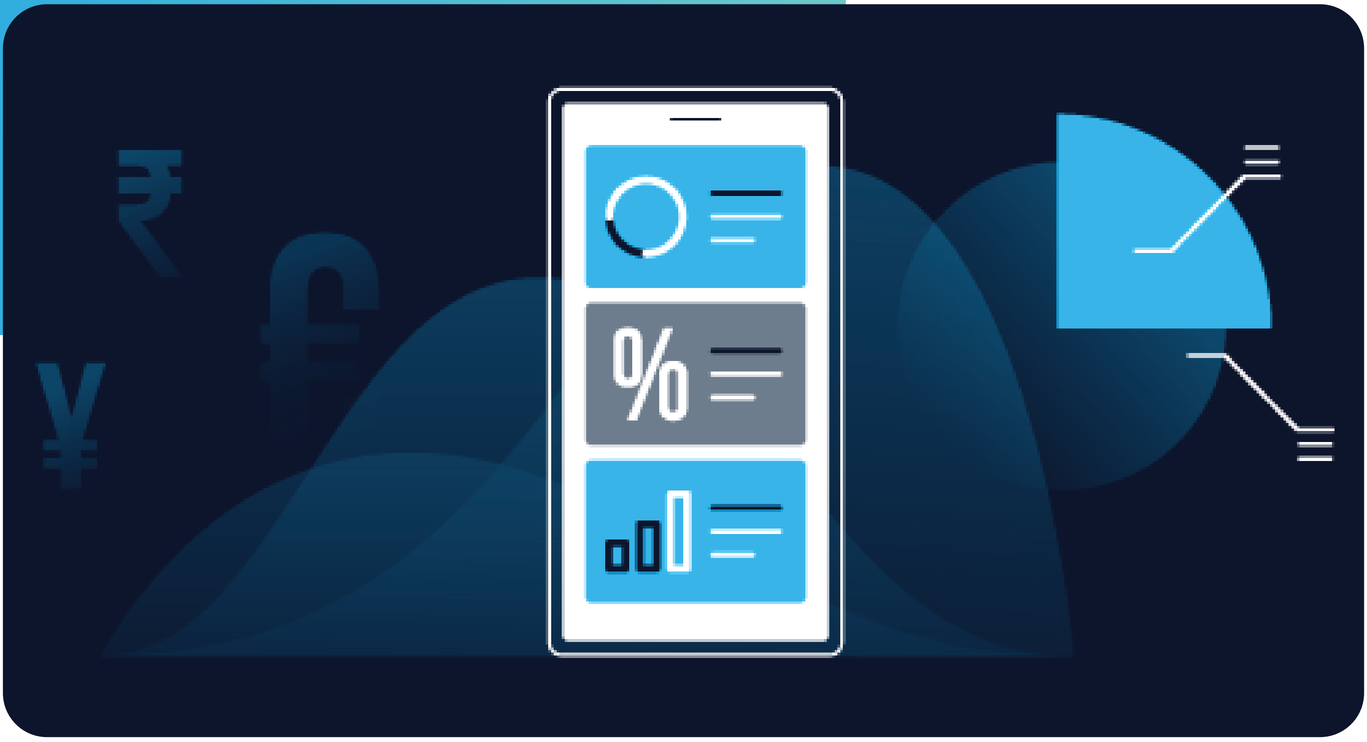
SUCCESS STORY

indium
Make Technology Work



STREAMLINING KYC PROCESSING: TEX.AI REVOLUTIONIZES TEXT EXTRACTION FOR A LEADING FINANCIAL SERVICES FIRM

PROJECT OVERVIEW

To build a text extraction model for bank statements employing the teX.ai product and build a pipeline for easy orchestration of future bank statements for text extraction.

SOLUTION DELIVERED

teX.ai, Text Extraction

CLIENT DOMAIN

Financial Services

KEY HIGHLIGHTS

- Challenges were overcome by using a combined RNN model (LSTM-CRF) to extract text.
- Despite the inherent difficulties of working with PDF documents, the project was completed in 3-4 months.
- Processing time for a single file was reduced to less than a minute, resulting in an 87% increase in daily application processing.
- The client's longstanding challenge of extracting accurate information from bank statements was permanently resolved.

ABOUT CLIENT

The client is a rapidly growing firm in India that operates in the financial domain and has embraced digital transformation to gain a competitive edge. They specialize in providing quick and efficient credit score ratings to their customers and offer assistance to banks in assessing their customers' creditworthiness. By leveraging technology and digital platforms, the client aims to revolutionize the credit delivery process in India. As part of their customer validation process, the client needs to process thousands of scanned bank statements to meet the KYC (Know Your Customer) requirements for the applicants.

BUSINESS REQUIREMENTS

- The documents to be extracted consisted of two types: scanned images and digital PDFs. The extraction process aimed to capture five key fields, which were located in both the table section (tabular data) and outside the tables (peripheral data).
- These fields included the account holder name, date, name of the bank, transaction details, and address.
- A training corpus of 2000 bank statements was initially used to train the required extraction models.
- To accommodate the daily influx of documents, the system needed to be scalable.

BUSINESS CHALLENGES

The client faced several challenges in their existing text extraction process:

- Accuracy of Text Extraction: The client was dissatisfied with the accuracy level of the text extraction output generated by their existing tools. They sought a solution that could provide higher accuracy in extracting data from bank statements.
- Varied Field Names: The field names to be extracted from the bank statements differed from bank to bank. For example, the "Credit amount" field could be represented as "Cr", "Credited", "Amount Credited", etc. This variability added complexity to the extraction process.
- Varied Text Locations: The location of the text to be extracted also varied across different banks. Some text resided within tables, while others were located outside tables. This variation required a flexible approach to accurately locate and extract the required information.

SOLUTION HIGHLIGHTS

Bank Statement Classification:
The bank statements were systematically classified based on different bank types and quality.

Table Detection and Text Extraction:
To locate tables and extract text, we utilized the bounding box method supported by a Deep Learning Network. The teX.ai platform facilitated the extraction process.

Extraction of Text Outside Tables:
For text present outside the tables, a combined RNN model was employed to accurately extract the required information.

APPROACH & IMPLEMENTATION

Bank Statement Types:

- The initial analysis revealed the presence of 200 different types of bank statements among the provided 2000 files.

Bank Statement Quality:

- Out of the total files, 830 were classified as good quality, and this subset was predominantly utilized for creating the initial NLP model.

Table Data Extraction:

- For scanned bank statements, a deep learning network was employed to detect tables by drawing bounding boxes around them. An object detection neural network was utilized for this stage of the process.

Table Contents:

- For the extraction of text from detected tables in both digital and scanned PDFs, Tabula and Camelot, embedded in teX.ai, were utilized.
- These tools enabled the accurate extraction of text from tables.

Extraction of Data Outside Tables (Peripheral Data):

- To extract the required peripheral data from both digital and scanned PDFs, a combined Recurrent Neural Network (RNN) model called LSTM-CRF was employed.
- This model effectively extracted the peripheral data regardless of the PDF document type.

Output:
The extracted data was generated as a JSON file, which the client could use in their in-house product for further processing and analysis.

Set-up:
teX.ai was set up on-premises in the client's data center, providing them with full control and security over their data. Training was provided to the client's in-house team to effectively utilize the system. Flask and Requests were leveraged to build the pipeline, facilitating the smooth flow of data extraction and processing.

BUSINESS IMPACT

The implementation of the text extraction solution had the following positive impacts on the client's business:

Time:
The processing time for a single file was reduced to less than a minute, significantly improving efficiency and turnaround time. This allowed the client to process a larger number of applications in a single day, resulting in an 80% increase in throughput.

Accuracy:
The accuracy of the text extraction was close to 90%, providing reliable and accurate data for further processing and analysis. With a larger dataset, the system was designed to further increase accuracy, ensuring high-quality output.

TECH STACK

Table Detection and Extraction:

TensorFlow CAMELOT SOFTWARE PLANNING

Tabula PDFplumber

Anago CLEANING SYSTEMS RegEx

Pycrfsuite CRF (conditional random fields)
REGEX

Pre-processing and Post Processing Tools:

OpenCV pdf Poppler

pandas {JSON}

Application Deployment and Access:

Flask web development, one drop at a time

Requests http for humans

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
Cupertino | Princeton
Toll-free: +1-888-207-5969

INDIA
Chennai | Bengaluru | Mumbai | Hyderabad
Toll-free: 1800-123-1191

UK
London
Ph: +44 1420 300014

SINGAPORE
Singapore
Ph: +65 6812 7888

www.indiumsoftware.com

For Sales Inquiries
sales@indiumsoftware.com

For General Inquiries
info@indiumsoftware.com