EASILY EXTRACT COMPLEX DATA FROM DOCUMENTS USING

ARTIFICIAL INTELLIGENCE

SureXtract

IMPACTSURE TECHNOLOGIES PRIVATE LIMITED

A 207, Eastern Business District,
LBS Road, Bhandup W,
Mumbai, Maharashtra 400078,
INDIA.

Phone: +91 22 4961 0978 | **Email**: sureask@impactsure.com | **Web**: www.impactsure.com

FIRST THINGS FIRST- EXTRACT DATA INTELLIGENTLY!

The key part of any intelligent document processing project is to extract relevant data from its source documents.

This becomes even more complex when the input documents themselves come in different sizes, layouts, fonts, numerical and text formats, languages, line variations, images, graphs, text, stamps, seals and handwritten text.

Of course, data extraction tools commercially available today are highly advanced and do a great job of extracting data from structured and unstructured documents. Based on our experience of extracting data for various banks and corporates, we have seen that while all tools extract data with varying degrees of success and accuracy, there is ample scope to improve the quality and relevance of the data that is extracted *contextually*. This is because the data that is extracted is the primary input based on which further document processing depends.

This paper discusses some of the challenges involved in extracting data from documents and how innovative approaches can be used to solve them.

CHALLENGES IN EXTRACTING DATA

Automated data extraction is a boon to reduce error prone manual data entry operations and processes. However, the accuracy of data extraction depends on the context and content of the documents and their structure.

Documents could be structured, semi-structured, or unstructured. The layout structure for semi- structured documents could vary for the same type of documents where the locations of logical objects like names, dates, descriptions and amounts could vary.

Extracting tabular and unstructured data

There are readily available tools that can extract tabular and unstructured data from PDF documents. However, it must be noted that most of these tools have the singular capability of extracting data from similar documents only in which the text headers are similar or if the documents are in the same format.

Various reports that such as employee performance, inventory management, marketing dashboards have the names of headers and textual descriptions of items that are not standardized like employee designation, performance rating, inventory types, website analytics and campaign metrics.

Similarly, financial documents like credit and analytical reports, P & L statements and Annual reports have the holdings, assets and liabilities, credit ratings et al.

Therefore, it is important to understand the context of the business application and then train the AI system to pick out the relevant information from the source documents.

When extracting data from financial statements of corporates or financial institutions, some of the key challenges that arise are:

1) The template, table structure, product descriptions, heading formats, sub-heading formats, subtotals and totals are not standardized.

2) The numerical formats and decimal separators used in different countries (For ex. Europe uses a comma as a decimal separator)
3) Names of number formats (For ex. Comma Vs Decimal separator, Lakh Vs Million, etc.)
4) Extraction of data from portfolios with multiple currencies displayed in one table
5) Identification and interpretation of currencies quoted in large units like Indonesian Rupiah, Vietnamese Dong (these are quoted in thousands and millions). These should be later converted into the home currency or converted from the home currency to currencies like USD, EUR, CHF, GBP, etc.
6) Non-standardization of names of securities that are not named as per ISIN standards
7) There are also values that need to be derived from multiple data elements which need to be identified and extracted contextually.

Extracting data from graphs and images

With Visually Rich Documents (VRDs) and unstructured documents becoming the norm in today's business world, understanding the data to be extracted is the key requirement. It is only after this stage that any other intelligent processing can be done as an effective long-term solution.

Some of the challenges that are encountered in VRDs are listed below:

1) Extraction of numbers and text from graphs is complex because all the graphs are not the same. For instance, a document may have a mixture of bar graphs, pie charts, scatter plot diagrams, radar charts and other pictorial representations of data
2) The X and Y axes may or may not be marked with standardized units of measurement and may be abbreviated
3) The legends for a graph may be represented by colours and abbreviated names
4) The contrast between colours used and the formation of fonts and numbers may not be easily distinguishable
5) Scanned or photocopied documents may have misaligned text which are not contextually understood by the extraction tool
6) Handwritten text and signatures may not be placed correctly in the spaces provided in the document.

The challenge for an off-the-shelf AI solution is to identify, understand, classify and interpret data contextually.

SUREXTRACT- EXTRACTING DATA CONFIDENTLY AND ACCURATELY

SureExtract is proven to extract data from Visually Rich Documents (VRDs) and unstructured documents. It understands the whole document contextually by analysing the layout and visual representation of information. SureXtract uses statistical methods, neural networks, decision trees, and rule-based learning techniques to intelligently capture relevant data irrespective of the position in the source document.

It can be further trained to achieve higher degrees of accuracy by enabling human-in-the loop capabilities. SureXtract can be introduced to any workflow process that requires data to be extracted accurately so that post-processing can be done efficiently.

To overcome the common challenges mentioned in the sections above, SureXtract extracts data and assigns a level of confidence to each extracted field. This configuration is done initially. The human operator

can track the level of accuracy and recheck the fields based on the confidence level displayed. The accuracy of extraction improves based on the frequency and variety of documents used over time.

A success story at a leading Financial Institution in Singapore

A leading financial institution in Singapore uses SureXtract to extract data from financial statements and annual reports daily. The extracted data is monetized by offering it as a part of the data feed to their customers.

More than 2,500 annual reports and financial reports containing at least 75 pages each have been analysed in the last 6 months. These documents contain visually rich and unstructured data that must be extracted accurately for reporting and analytical purposes.

While the accuracy and relevance of automated data extraction is more than 95%, there are still a few cases where human intervention is needed to cross verify the results. These cases are few and far between.

CONCLUSION

From our practical experience in extracting complex data from a wide variety of documents, we can confidently say that adopting any data extraction solution into your workflow will take some time.

There are too many factors that influence the variability of input documents. We suggest that organisations have a clear understanding of what their requirements for document processing are and the acceptable levels of extraction accuracy that they are comfortable with. Based on these basic parameters, the best combination of automation with human assistance can be implemented.

AI is best used to augment the capabilities of humans and not replace them. A well-trained AI solution will drive consistent, auditable, and accurate decisions and help bring in higher operational efficiencies into the business, while at the same time increasing revenues and profitability.

IMPACTSURE TECHNOLOGIES PRIVATE LIMITED

A 207, Eastern Business District,

Lal Bahadur Shastri Road,

Bhandup, Mumbai,

Maharashtra 400078, INDIA.

Phone: +91 22 4961 0978 | **Email:** sureask@impactsure.com | **Web:** www.impactsure.com

---
