Data, Policy & Innovation Centre

The Government of Odisha receives hundreds of thousands of citizen grievances each year, many submitted with supporting documents. These materials contain essential information for understanding complaints, yet their unstructured and multilingual nature slowed review and limited the government's ability to extract insights. The team worked with the Data, Policy & Innovation Centre to design a pipeline that transformed these documents into structured, privacy-preserving data suitable for analysis.

The project focused on three components. First, the team developed a document classifier trained on manually labeled pages to predict properties such as language, handwritten status, and scan quality. These predictions improved downstream processing by routing pages to appropriate text-extraction methods. Second, the team built an OCR module that converted each page into machine-readable text, enabling large-scale analysis of both English and Odia documents. Finally, the team trained a PII-tagging model to identify and redact sensitive personal information before the documents were uploaded or used for analysis.

Together, these components created a unified system that prepares grievance documents for future summarization, categorization, and dashboarding tools that can support more responsive public services in Odisha. Spanning the scope of the PII model.

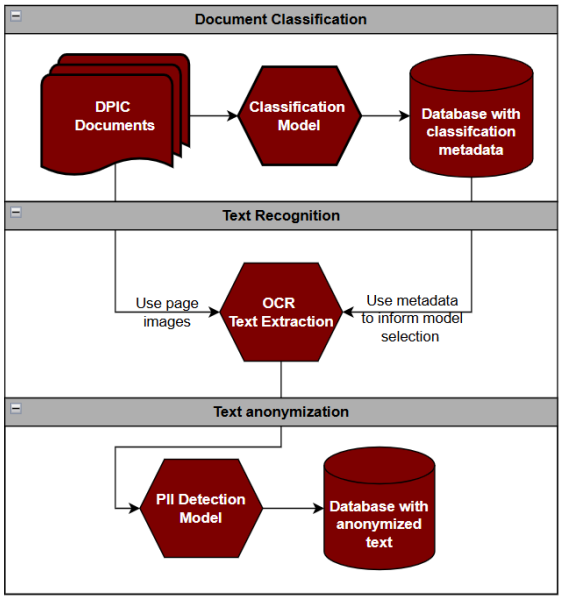

Figure: End-to-end pipeline showing document classification, text extraction, and PII anonymization used to prepare grievance documents for analysis