AI Driven Document Analysis System

Project ID 3

Group 10

23/07/2025

GANGADARI M.D.S. 220178X
GUNARATHNA H.R.A. 220186U
GUNASEKARA S.L. 220191F

Introduction

Modern organizations and individuals handle vast amounts of documents daily, including invoices, receipts, contracts, tax forms, medical records, and research papers. Manual processing, sorting, and information extraction from these documents is time-intensive, error-prone, and inefficient. Current solutions often require specialized knowledge or expensive enterprise software, creating barriers for widespread adoption across various user types and organizational needs. Additionally, most existing solutions are designed exclusively for enterprise use, leaving individual users without accessible options.

The need for automated document processing has grown significantly with the increase in digital document workflows. Organizations spend considerable resources on manual document classification, data extraction, and information retrieval. This creates bottlenecks in business processes and limits productivity. There's a clear market demand for an intelligent, user-friendly system that can automatically analyze, categorize, and make documents searchable while providing intelligent question-answering capabilities.

Proposed Solution

We plan to develop a complete AI-driven document analysis system that uses machine learning and natural language processing technologies. The system automatically processes documents for information extraction, classifies them by type, summarizes content and provides intelligent search capabilities through a user friendly dashboard along with a chatbot for document based question answering.

The solution includes the following components:

OCR and Layout Analysis Module: Extract text while preserving document structure, identifying headers, paragraphs, tables, and images for contextual understanding. Users can choose from multiple model options optimized for different priorities - high accuracy models for critical documents, speed-optimized models for bulk processing, or balanced models for general use.

Document Classification Module: Automatically categorize documents by type (invoices, financial reports, research papers, legal documents, medical records, etc.) using advanced deep learning models, eliminating manual sorting requirements.

Intelligent Summarization System: Generate concise summaries of document content, enabling quick overview and understanding without reading entire documents. The system provides multiple summarization model options allowing users to select based on their specific needs - detailed analytical summaries, brief executive summaries or domain-specific summarization approaches.

RAG-powered Search and Chatbot: Enable natural language querying and intelligent document-based question answering through Retrieval Augmented Generation (RAG) technology integrated with vector databases and advanced language models.

The system will be deployed as a SaaS web application with subscription based access, targeting both individual users and enterprise clients including financial institutions, law firms, government offices, hospitals, and research centers. A key differentiator is our flexible model selection approach - users can choose from multiple AI model options based on their specific requirements, whether they prioritize processing speed for high-volume operations, maximum accuracy for critical documents, or a balanced approach for general use cases.

Datasets

Our system will use the following publicly available datasets covering multiple document types.

RVL-CDIP Dataset: Document images across 16 categories including invoices, letters, forms, and reports.

Medical Lab Reports: Collection of medical report images for healthcare document classification and OCR tasks.

PubLayNet, **DocBank**: Datasets containing academic papers for scholarly document processing.

Financial Documents: Personal financial dataset covering various financial document types.

ChartQA Dataset: Chart and graph images for visual document element extraction.

ExpressExpense Dataset: Receipt images for OCR and information extraction.

ICDAR-2019-SROIE: Scanned receipt OCR and information extraction dataset.

FUNSD: A Dataset of scanned forms.

Similar Projects

PaperMind: PaperMind is an AI-powered document assistant designed to enhance user productivity through a range of sophisticated features. It supports multiple languages, including German, English, Dari, Pashto, and French, ensuring broad accessibility for users.

DeepDoctection: A project designed for document layout analysis, OCR and information extraction, supporting both scanned images and PDFs. Supports RAG workflows by enabling extraction and indexing of document data for retrieval tasks

docTR: DocTR supports key features like layout analysis, OCR (text detection + recognition), and key information extraction.