KLE Society’s

KLE Technological University, Hubballi



Department of Computer Applications

BCA Vᵗʰ Semester

Synopsis Report

On NLP with Gen AI Project

**PolyDoc Multi-lingual Document Understanding and System**

Submitted by

1. Darshan Janganure (01fe23bca071)
2. Pooja Madiwalar (01fe23bca074)
3. Mohammad Yaseen(01fe23bca118)
4. Sayeda Iram (01fe23bca111)

Under the Guidance of

Prof. Deepa Mulimani

KLE Technological University

Vidyanagar, Hubballi – 580031

2025-2026

**1. Introduction**

In today’s digital world, documents come in a variety of formats such as PDFs, DOCX, PPTX, and scanned images. Extracting meaningful information while preserving the original layout and formatting is a challenging task, especially for multi-lingual and complex documents. PolyDoc is designed to address this challenge by leveraging state-of-the-art AI and machine learning techniques to automate document understanding, translation, summarization, and layout preservation.

This project was chosen due to the increasing need for automated document processing in businesses, education, and government sectors, where manual processing is time-consuming and error-prone. With globalization, handling multi-lingual documents is crucial, making PolyDoc relevant for real-world applications in document management systems, digital archiving, and knowledge extraction.

**2. Objectives**

* + To develop an AI-powered system that extracts text and preserves the layout from various document formats such as PDF, DOCX, PPTX, and images.
  + To implement multi-lingual language detection and automatic translation features.
  + To summarize extracted content for quick information retrieval.
  + To maintain the document's original visual and structural layout in the output.
  + To provide structured outputs in formats like JSON and Markdown for easy integration.

**3. Problem Statement**

Digitization/ Text Extraction for multi-lingual, noisy documents while preserving document Layout including hand written documents.

**PS Description:**

In today’s global and digitized world, documents are far more than just sequences of text. They are structured, visually rich, and multilingual—ranging from legal contracts and academic papers to business reports, government forms, and presentation decks. These documents exist in diverse formats such as Word documents (DOCX), PDFs, PowerPoint slides (PPT), and scanned images including handwritten documents, often containing mixed scripts (e.g., English-Arabic or Hindi- English). Modern AI systems need to read, understand, and generate structured outputs from these documents.

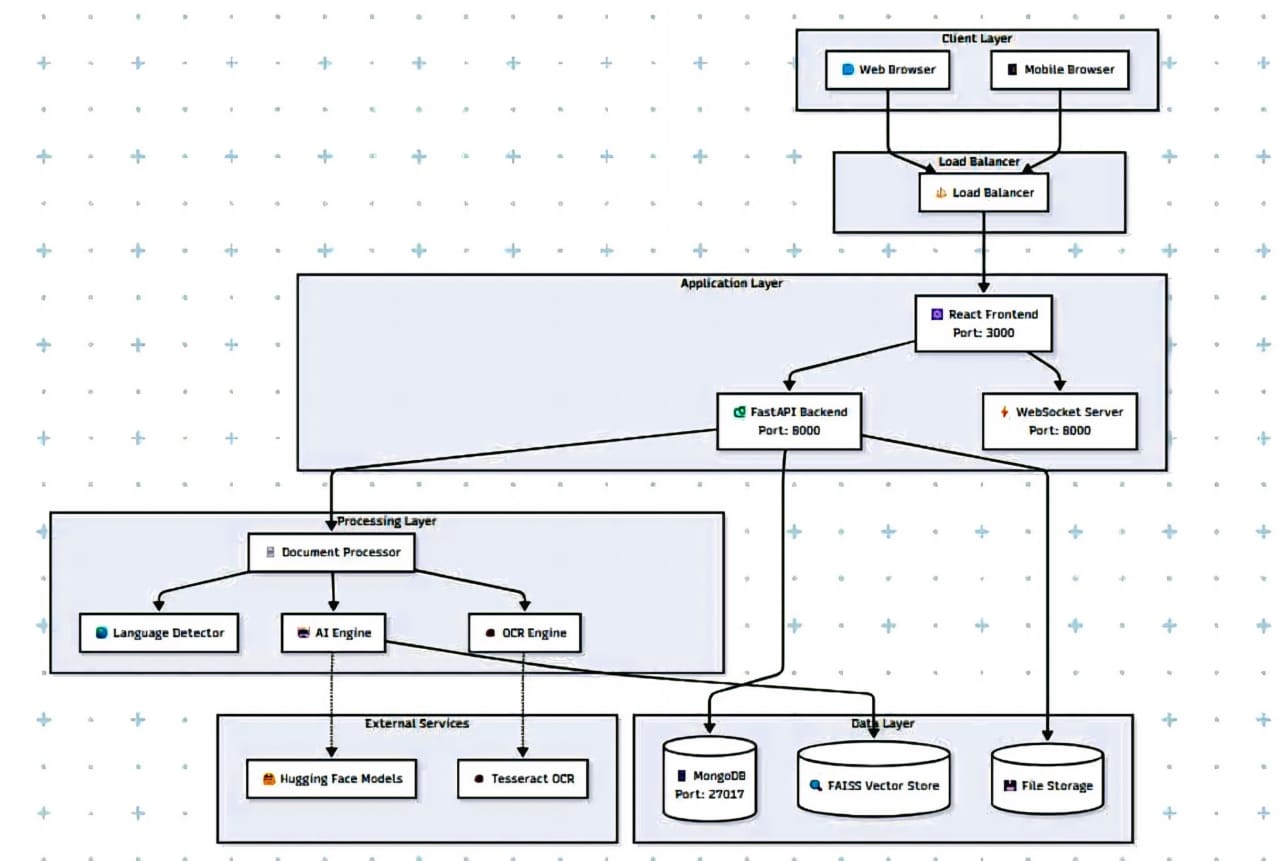
**4. Scope of the Project**

* + Text extraction from multiple document types including scanned images.
  + Multi-lingual language detection and translation support.
  + Summarization of extracted content.
  + Layout preservation using AI-based layout analysis.
  + Output formatting in JSON and Markdown.

**5. Technologies Used**

* **Programming Language:** Python , JavaScript, HTML, CSS
* **Web Frameworks & Libraries**: FastAPI, React , Uvicorn, Vite
* **Core AI Framework**: PyTorch
* **Embeddings**: HuggingFace Transformers
* **OCR & Computer Vision**: Tesseract OCR , EasyOCR

**6. System Architecture**



**7. Expected Outcome**

* + A fully functional AI-based system capable of extracting, translating, summarizing, and preserving the layout of multi-lingual documents.
  + Structured outputs that facilitate integration with other systems.
  + A user-friendly web interface for easy document upload and result visualization.
  + Improved accuracy and usability over existing solutions, saving manual processing time.

**8. Conclusion**

PolyDoc is a comprehensive, free, and open-source multi-lingual document understanding system that combines advanced AI capabilities with modern web technologies to provide intelligent document processing and real-time chat functionality. This project aims to automate and streamline document workflows, making information extraction efficient and accessible. The system’s adaptability to various languages and formats makes it highly relevant in today’s globalized digital ecosystem.