Intelligent Pdf Content Extractor and Question Answering Using Open AI

Vijaya Kumar M[1], Dr. A. Vinoth[2]
[1]*Junior Researcher, Department of Information Technology,*
Sri Krishna Adithya College of Arts and Science.
[2]*Assistant professor, Department of Information Technology,*
Sri Krishna Adithya College of Arts and Science.

***Abstract*— As digital documents grow quickly; we need smart systems that can quickly pull out and understand information. People often use Portable Document Format (PDF) files to store both structured and unstructured data, but it is still hard to get useful information from them. This paper describes an OpenAI-powered Intelligent PDF Content Extractor and Question Answering System. The system uses Natural Language Processing (NLP) techniques and OpenAI's language models to give context-aware answers to user questions by pulling text data from PDF files. The proposed system makes documents easier to access, cuts down on the amount of work that needs to be done by hand, and lets people interact with the content of documents. The system gives accurate and quick answers, as shown by experiments. This makes it useful for use in business analytics, education, and research.**

***Index Terms*—PDF Extraction, Question Answering System, OpenAI, Natural Language Processing, Information Retrieval, Artificial Intelligence**

I. INTRODUCTION

As more and more people use digital documents, it has become very hard to get useful information from big PDF files. Keyword-based search is what traditional methods use, but this doesn't always get the full meaning of the content.

Recent improvements in Artificial Intelligence and Natural Language Processing (NLP) have made it possible to make smart systems that can understand and work with human language. This study suggests a system that combines AI-powered question answering with PDF content extraction using OpenAI models. The system allows users to interact with PDF documents by asking questions and receiving accurate answers based on the document content.

II. RESEARCH OBJECTIVE

The objective of this research is to develop an intelligent system for extracting and understanding content from PDF documents and providing accurate, context-aware answers using Natural Language Processing and OpenAI models.

Specifically, the study aims to:

- Extract and preprocess textual data from PDF documents
- Enable semantic understanding of document content
- Develop a natural language-based question answering system
- Improve information retrieval using embedding and similarity techniques
- Evaluate system performance in terms of accuracy and efficiency.

III.LITERATURE REVIEW

Researchers in Natural Language Processing and Information Retrieval have looked into how to get information out of PDF files and analyze it. Traditional methods mostly used keyword-based search methods, which can't capture the meaning of the context and often give less accurate results.

Recent progress in deep learning and transformer-based models has made it much easier to understand and create text. Generative Pre-trained Transformers (GPT) and other models have made semantic analysis and generating responses that sound like a person much better. Also, retrieval-based methods like vector embeddings and similarity search techniques have made question-answering systems more accurate.

Several studies have examined Retrieval-Augmented Generation (RAG), which integrates information retrieval with language models to enhance response relevance. Even with these improvements, it is still hard to combine fast PDF content extraction with real-time, context-aware question answering.

This study seeks to overcome these constraints by creating an intelligent system that integrates PDF extraction with OpenAI-driven question answering to enhance precision and user-friendliness.

IV.PROBLEM STATEMENT

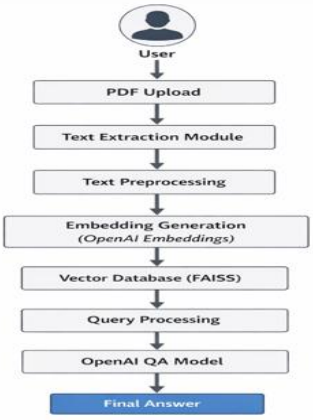
The rapid growth of digital documents, particularly in PDF format, has made information retrieval a challenging task. Existing systems primarily rely on manual reading or keyword-based search, which are inefficient and fail to capture the contextual meaning of the content. As a result, users often struggle to locate relevant information quickly and accurately.

Furthermore, traditional PDF processing tools lack the capability to support interactive and intelligent querying. There is a need for a system that can not only extract textual data from PDF documents but also understand the content semantically and provide precise, context-aware answers.

This research addresses these challenges by proposing an intelligent PDF content extraction and question answering system using advanced Natural Language Processing and OpenAI models.

V.SYSTEM ARCHITECTURE

The proposed system architecture is designed to quickly and accurately extract, process, and analyze PDF content so that it can support intelligent question answering. The most important parts of it are:



1. Module for Uploading PDFs:
- Let's people upload PDF files so they can be processed.

2. Module for Text Extraction:
- Uses parsing libraries to get text from PDF files.

3. Module for Preprocessing:
- Removes noise and breaks up the extracted text into meaningful parts to clean and organize it.

4. Module for embedding and storage:
- It turns text into vector embeddings and saves them in a vector database so they can be found quickly.

5. Module for Processing Queries:
- It takes user questions in natural language and finds the right document content using similarity search methods.

6. Module for making answers:
- Uses OpenAI models to create accurate, context-aware answers based on the information it finds.

The overall design makes sure that all the parts work together smoothly, which makes it easy to get information from PDF files and answer questions intelligently.

VI.SYSTEM FLOW

1. The system flow shows how to get information from PDF files and answer user questions step by step.
2. The user uploads a PDF file using the system's interface.
3. The text extraction module processes the uploaded document to get the raw text.
4. Cleaning, segmenting, and normalizing are all parts of preprocessing the extracted text.
5. We break the processed text into smaller pieces and turn them into vector embeddings.
6. A vector database stores these embeddings so that they can be easily found by looking for similar ones.
7. The user types in a question in plain English.
8. Using similarity search, the system processes the query and finds the text segments that are most relevant.

VII.METHODS

The system does these things:
Step 1: Get the data Libraries like PyPDF2 and pdfplumber are used to get content from PDFs.
Step 2: Preparing the Data Get rid of the noise Tokenization Normalizing text.
Step 3: Making the Embedding OpenAI embedding models turn text into embeddings.
Step 4: Search for Similarities Vector similarity is used to get relevant text chunks.
Step 5: Making Answers the OpenAI model makes answers based on the context it finds.

VIII.IMPLEMENTATION DETAILS

Language for programming: Python
Framework: Flask / Streamlit
Library:
PyPDF2
LangChain
FAISS
API for OpenAI

IX. RESULTS AND DISCUSSION

We tested the system with a number of PDF files, such as reports and academic papers. The results show:
Very accurate answers to questions based on context
Less time spent getting information
Better experience for users
The system works well with both small and medium-sized documents. But how well it works depends on how good the extracted text is.

X. ADVANTAGES

- Intelligent document interaction
- Time-efficient
- Context-aware responses
- Scalable solution

XI.APPLICATIONS

- Academic research
- E-learning platforms
- Business document analysis
- Legal document review

XII. LIMITATIONS

- Dependent on text quality in PDF
- Limited support for scanned PDFs (without OCR)
- API usage cost

XIII. FUTURE WORK

Future enhancements may include:
- OCR for scanned documents
- Multi-language support
- Voice-based query system
- Integration with cloud platforms

XVI. CONCLUSION

The paper proposed an Intelligent PDF Content Extractor and Question Answering System based on Open AI. The system can extract and process PDF content efficiently and provide accurate answers to user queries. The proposed system utilizes NLP and AI technologies to improve document usability and accessibility. It can be adopted in different domains to achieve efficient information retrieval.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the faculty and management of the institution for their continuous support and guidance in this research work. The authors would like to thank the project guide for their valuable suggestions and guidance. The authors would like to thank the peers and all individuals who have supported this work in any manner.

REFERANCES

[1] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention Is All You Need,” Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), pp. 5998–6008, 2017.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances

in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[3] OpenAI, "GPT Models and API Documentation," Accessed: 2025.

[4] H. T. Nguyen, T. D. Nguyen, and M. T. Tran, "A Survey on Question Answering Systems," International Journal of Computer Applications, vol. 179, no. 42, pp. 1–7, 2018.

[5] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009.

[6] LangChain, "LangChain Documentation," Accessed: 2025.

[7] PyPDF2 Developers, "PyPDF2 Documentation," Accessed: 2025.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, pp. 4171–4186, 2019.

[9] Facebook AI Research, "FAISS: A Library for Efficient Similarity Search and Clustering of Dense Vectors," Accessed: 2025.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Proceedings of ICLR, 2013.

[11] D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[12] Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Pearson, 2020.