© Volume 2, Issue 10, October 2026 | JATIR
e-ISSN: 2839-5786, p-ISSN: 3139-8464

Document Analyzer An AI-Based System for Intelligent Document Processing and Analysis

[1]Vivek Tiwari, [2]Jay Chand
[1]*Student Kamla Nehru Institute of Physical & Social Science Engineering Institute, Faridipur Sultanpur, Uttar Pradesh 228119, India*
[2]*Assistant Professor, Student Kamla Nehru Institute of Physical & Social Science Engineering Institute, Faridipur, Sultanpur, Uttar Pradesh 228119, India*
doi.org/10.64643/jatir.141099

***Abstract*—Documents are one of the most important sources of information in every organization. The growing volume of digital documents makes manual reading and analysis time-consuming and error-prone. This paper introduces the Document Analyzer, an AI-based system that integrates document parsing, Natural Language Processing (NLP) and Machine Learning techniques to automatically extract, summarize, classify and answer questions about the content of uploaded documents. The proposed system accepts PDF, DOCX and TXT files, extracts and pre-processes their text and uses NLP and AI techniques to generate keywords, summaries, document categories and answers to user queries. Document Analyzer is different from many current tools which are specialized for a single task like extraction or summarization. The project is discussed thoroughly in terms of system architecture, methodology, algorithm, implementation approach, advantages, limitations, and future scope.**

***Index Terms*—Document Analyzer Natural Language Processing Artificial Intelligence Text Summarization Keyword Extraction Document Classification Optical Character Recognition Question Answering**

I. INTRODUCTION

In any organization documents are one of the most important sources of information. Every day, educational institutions, companies, banks, hospitals, government departments and research organizations generate and store a huge number of documents. These documents can be hundreds of pages long and it is difficult for users to quickly locate the information they need.
A Document Analyzer is an intelligent software application that processes digital documents and extracts useful information from them. The system uses Document parsing, Natural Processing, Machine Learning and text-analysis techniques to understand the content of a document.

The proposed project offers a simple interface to upload a document. The system reads through the uploaded file and extracts the text. Then, different analysis operations can be applied such as text extraction, keyword extraction, document summarization, document classification, important information identification, question answering and document comparison. The main objective of the project is to make the analysis of documents faster, simpler and more efficient.

II. PROBLEM STATEMENT

The increasing volume of digital information has generated a big problem for users who need to analyze large documents. Reading a long document manually can take hours, days even. The main problems of traditional document analysis are: large documents are time consuming to read; important information is difficult to locate; manual summarization is time consuming; human analysis is error prone; comparison of multiple documents is difficult; unstructured text is difficult to process automatically; and users may need technical knowledge to extract information from documents.

Hence, there is a need for an intelligent system for automatically processing documents and providing useful information within a short period of time. The proposed Document Analyzer deals with these problems by integrating document processing with AI and NLP techniques

III. OBJECTIVES OF THE PROJECT

The main objectives of the Document Analyzer project are described below.

A. Automatic Text Extraction

The system should automatically extract text from uploaded documents such as PDF and DOCX files.

B. Document Summarization

The system should generate a short and meaningful summary from a lengthy document so that users can understand its main points quickly.

C. Keyword Extraction

The system should identify important keywords and terms from the document.

D. Intelligent Document Analysis

The system should analyze the content using Natural Language Processing techniques and provide meaningful results.

E. Reduce Manual Effort

The system should reduce the amount of time required to manually read and analyze documents.

F. Easy-to-Use Interface

The application should provide a simple interface through which users can upload documents and view analysis results.

G. Future AI Integration

The system should provide a foundation for advanced features such as question answering, semantic search, document comparison, and AI-based recommendations.

IV. LITERATURE REVIEW

Document analysis has become an important area of Artificial Intelligence and Natural Language Processing. Modern artificial intelligence systems can understand the semantic meaning of text, whereas traditional document processing systems are mostly designed for extracting text from files.

Optical Character Recognition (OCR) technology is a popular technique for extracting text from scanned documents and images. OCR converts characters in an image into text that can be used by a computer.

Natural Language Processing offers techniques for processing human language. The usual NLP techniques are tokenization, stop-word removal, stemming, lemmatization, named entity recognition and text classification.

Another important area of document analysis is text summarization, which can be divided into two major categories. Extractive summarization directly selects important sentences from the original document while abstractive summarization generates a new summary by the meaning of the original document.

Recent advances in transformer-based AI models have greatly improved the ability to understand and summarize text. These technologies can be used in document-analysis applications to provide smarter results.

However, many existing systems are designed to perform only a single task, such as extracting text or summarizing. The proposed Document Analyzer is a single application that integrates several document-analysis functions.

V. PROPOSED SYSTEM

The proposed Document Analyzer is an AI-based web application that accepts a digital document as input and produces useful analytical information as output. The overall flow of the system is: User → Upload Document → Document Processing → Text Extraction → NLP Processing → AI Analysis → Results. The system consists of several major modules, described below.

A. User Interface Module

The user interface allows users to upload documents, select analysis options, view extracted text, view the summary, view keywords, search information, and ask questions about the document. The interface should be simple and responsive.

B. Document Upload Module

This module accepts supported document formats such as PDF, DOCX, and TXT. The uploaded file is validated before processing.

C. Text Extraction Module

After uploading the document, the system extracts its textual content. For PDF documents, PDF parsing libraries can be used. For DOCX files, document-processing libraries can extract the paragraphs and text. If the document contains scanned images instead of selectable text, OCR technology can be integrated.

D. Text Preprocessing Module

The extracted text might contain unnecessary spaces, symbols or formatting; hence preprocessing is performed. Typical preprocessing operations are to remove extra spaces, normalize the text into a standard format, sentence segmentation, tokenization, remove extra stop words, lemmatization or stemming. The cleaned text is then fed to the NLP module.

E. NLP Analysis Module

This module analyzes the processed text and identifies useful information. It can perform keyword extraction, sentence analysis, named entity recognition, topic identification, and text classification.

F. Summarization Module

The summarization module generates a shorter version of the document. For example, if a document contains 20 pages, the system may produce a summary containing the most important information in a few paragraphs.

G. Question Answering Module

In a more advanced version of the system, users can ask questions about the uploaded document. For example, for the question “What is the main objective of the project?”, the system may respond with “The main objective is to automate document processing and extract useful information using AI and NLP. This feature enhances the application interaction.

VI. SYSTEM ARCHITECTURE

The proposed system can be divided into the following layers.

Layer 1 – User Interface: The user interacts with the application through a web interface.

Layer 2 – Backend Server: The backend receives the uploaded document and manages processing requests.
Layer 3 – Document Processing: The document is parsed and text is extracted.
Layer 4 – NLP/AI Engine: The extracted text is processed using NLP and AI algorithms.
Layer 5 – Database/Storage: User information, document metadata, and analysis results can be stored securely.
Layer 6 – Output: The final results such as summary, keywords, extracted information, and answers are displayed to the user.
System Flow :
Start → User Uploads Document → File Validation → Text Extraction → Text Preprocessing → NLP/AI Analysis → Keyword Extraction + Summarization + Classification → Generate Results → Display Results → End.

VII. METHODOLOGY

The methodology of the proposed project is divided into the following steps.
1) Document Upload – The user uploads a supported document through the web interface.
2) File Validation – The system checks the file type and size to ensure that it is a valid document.
3) Text Extraction – The system extracts readable text from the document. For scanned documents, OCR can be used to convert images into text.
4) Text Cleaning – The extracted text is cleaned by removing unnecessary characters, spaces, and formatting.
5) NLP Processing – The cleaned text is converted into a format suitable for AI analysis. Tokenization and other NLP operations are performed.
6) Keyword Extraction – Important words and phrases are identified from the document, for example Artificial Intelligence, Machine Learning, NLP, Database, Web Application, and Document Processing.
7) Summarization – The system identifies the most important information and generates a concise summary.
8) Classification – The document can be classified into categories such as Research Paper, Resume, Academic Document, Business Report, Legal Document, or General Document.
9) Result Generation – The final analysis is presented to the user in an understandable format.

VIII. TECHNOLOGIES USED

Table I lists the primary technologies used in the proposed system and their purpose.
TABLE I
TECHNOLOGIES USED IN THE PROPOSED SYSTEM

Technology	Purpose
Python	AI/NLP and backend processing

HTML	Web page structure
CSS	Interface design
JavaScript	Frontend interaction
Flask / FastAPI	Backend API
NLP Libraries	Text processing
PyPDF / PDF Parser	PDF text extraction
python-docx	DOCX processing
OCR	Scanned document processing
SQL / MongoDB	Data storage
Machine Learning	Classification and analysis

Python is particularly suitable because it provides many libraries for Artificial Intelligence, Natural Language Processing, document processing, and machine learning.

IX. ALGORITHM AND PROCESSING

Algorithm: Document Analyzer. Input: Digital document. Output: Analysis report.
1)Start the application.
2) Accept the document from the user.
3) Validate the document format.
4) Extract text from the document.
5) If text is unavailable, apply OCR.
6) Clean and preprocess the extracted text.
7) Divide the text into sentences and tokens.
8) Identify important keywords.
9) Analyze document content.
10) Generate document summary.
11) Classify the document if required.
12) Store or display the analysis results.
13) Allow the user to search or ask questions.
14) Stop.

X. IMPLEMENTATION

The implementation can be divided into frontend and backend components.

A. Frontend

The frontend is the interface the user sees. The application can be created using HTML and CSS and can be made interactive using JavaScript. The main interface might have a document upload button, a file info section, an extracted text section, a summary section, a keywords section, a search/question box, and an analysis result section.

B. Backend

The backend manages document processing and communication between the user interface and AI modules. Python with Flask or FastAPI can be used to develop the backend. The backend receives the document, extracts text, sends the text to NLP modules, and returns the analysis results.

C. Database

A database can be used to store user accounts, document information, upload history, analysis results, and generated summaries. SQL or MongoDB can be used according to project requirements.

XI. EXPECTED RESULTS

After successful implementation, the system should provide the following results. Input: a user uploads a PDF, DOCX, or TXT document. Processing: the system extracts and analyzes the document content. Output: extracted text, a short summary, important keywords, the document category, important entities, and answers to document-related questions.

The system is expected to significantly reduce the time required to understand lengthy documents.

For example, a 50-page report that requires considerable manual reading can be analyzed automatically, allowing the user to quickly understand its main topics and important information.

XII. ADVANTAGES

Saves Time: Large documents can be analyzed quickly.

Reduces Manual Work: Users do not need to read every page manually.

Easy to Use: A simple interface makes the system accessible.

Multiple Features: Text extraction, summarization, keyword extraction, and classification can be provided in one application.

Scalable: Additional AI features can be added later.

Useful in Education: Students and researchers can quickly understand research papers and reports.

Useful in Business: Organizations can analyze reports and business documents efficiently.

Improved Information Retrieval: Important information can be found quickly.

XIII. LIMITATIONS

Poor quality scanned documents may reduce theaccuracy of OCR.

Complicated tables and images can be difficult to comprehend.

Sometimes AI-generated summaries will miss important context.

Very large documents may require a lot of computational resources.
The quality of the NLP/AI model determines the results.
Sensitive documents need appropriate security and privacy mechanisms.
These limitations are addressable by improved AI models, better OCR, document-security mechanisms and improved processing techniques.

XIV. APPLICATIONS

Education: Students can analyze research papers, notes, assignments, and academic reports.
Business: Companies can analyze business reports, proposals, meeting documents, and project reports.
Research: Researchers can extract important information and summaries from research papers.
Legal Sector: Legal documents can be analyzed to identify important clauses and entities.
Human Resources: The system can analyze resumes and extract skills, education, and experience.
Government Offices: Large administrative documents can be processed and searched efficiently.

XV. FUTURE SCOPE

A. Multilingual Document Analysis
The system can be extended to support Hindi, English, and other languages.

B. Advanced Question Answering
Users can ask natural-language questions about documents and receive AI-generated answers.

C. Document Comparison
Two or more documents can be compared to identify similarities and differences.

D. Voice Interaction
Voice input can be added so users can ask questions using speech.

E. Advanced OCR
Improved OCR can help analyze scanned documents and handwritten text.

F. AI-Based Semantic Search
Instead of searching only for exact keywords, users can search based on meaning.

G. Cloud Deployment
The system can be deployed on cloud platforms so that users can access it from anywhere.

H. Security

Encryption, authentication, access control, and secure document deletion can be implemented for confidential documents.

XVI. CONCLUSION

The Document Analyzer is an AI-powered system that makes it easier to read, understand and analyze digital documents. The proposed system combines document processing, Natural Language Processing and Artificial Intelligence to extract text and automatically generate useful information.

The system is capable of performing important operations such as text extraction, keyword extraction, summarization, classification and question answering. The project automates these tasks to reduced manual efforts and save valuable time.

The suggested application can be used in educational institutions, research organizations, business, legal departments, and government offices. Its modular architecture allows the addition of advanced AI features in the future.

The Document Analyzer shows how Artificial Intelligence can be applied to solve real-world information processing problems, and offers a practical solution for transforming large volumes of unstructured document data into usable and accessible information.

REFERENCES

[1]. D. Jurafsky and J. H. Martin, Speech and Language Processing. [Natural Language Processing reference material].
[2]. C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge University Press.
[3]. Research literature related to Natural Language Processing and Automatic Text Summarization.
[4]. Python documentation and libraries for document processing and NLP.
[5]. Research papers related to Artificial Intelligence-based document analysis.
[6]. Documentation of PDF, DOCX, OCR, and machine-learning technologies.