STM JOURNALS

Current Trends in
Information Technology

https://journals.stmjournals.com/ctit

ISSN: 2249-4707 (Online)
ISSN: 2348-7895 (Print)
Volume 14, Issue 2, 2024
May–August
DOI (Journal): 10.37591/CTIT

Review CTIT

# Revolutionizing Document Analysis Using AI Based Text Extraction Methods

Preeti Singh[1,*], Yashraj Singh[2], Aditya Singh[2], Adnan Ahmad[2]

***Abstract***

*The AI Document Analyzer is an innovative application created to revolutionize how we engage with and comprehend textual data. By integrating a range of advanced technologies, including Next.js, Drizzle ORM, OpenAI, Stripe, TypeScript, and Tailwind, this tool offers a comprehensive solution for document comprehension. Central to its functionality is a chat-based interface powered by the ChatGPT API, allowing users to engage in natural language conversations with their uploaded documents. This interface not only simplifies access to information but also improves the overall user experience. A standout feature of the AI Document Analyzer is its capability to generate summaries and visualizations of complex documents. By leveraging the natural language processing capabilities of OpenAI, the application can extract and present critical insights in a concise and understandable manner. This is especially beneficial for users who need to quickly understand the main points of lengthy documents without having to go through all the content. The integration of Next.js ensures efficient server-side rendering and a smooth user interface, while Drizzle ORM handles the database interactions seamlessly. This combination guarantees that the application is both swift and dependable. Stripe integration facilitates secure and straightforward transactions, enabling users to access premium features or services with ease. TypeScript is employed for its robust type-checking capabilities, which enhance the reliability and maintainability of the codebase. Tailwind CSS is used to create a modern, responsive, and user-friendly design. Overall, the AI Document Analyzer represents a significant advancement in document analysis and information management. By combining powerful technologies and focusing on user-centric design, it offers an innovative solution for extracting and understanding insights from textual data. Whether for business, academic, or personal use, this tool redefines how we interact with documents, making information more accessible and manageable.*

**Keywords:** AI document analyzer, ChatGPT, natural language processing, text extraction, chatbot, artificial intelligence

## INTRODUCTION

In this research study, we delve into advancements in document comprehension technology, focusing on the development of the AI Document Analyzer. Crafted with modern technologies, this transformative SaaS solution revolutionizes how textual content is extracted and understood from PDF documents. Through a chat-based interface, users can pose targeted questions and receive precise answers powered by OpenAI's advanced algorithms.

Additionally, we explore the evolution of search engines, from traditional cataloging systems to digital giants like Google and Yahoo. Our comparative study extends to contemporary AI-driven conversational agents, analyzing ChatGPT by OpenAI and Google's BARD. This research

***Author for Correspondence**
Preeti Singh
E-mail: preetis88@gmail.com

[1]Assistant Professor, Department of Computer Science and Engineering, Babu Banarasi Das Engineering College, Lucknow, Uttar Pradesh, India
[2]Student, Department of Computer Science and Engineering, Babu Banarasi Das Engineering College, Lucknow, Uttar Pradesh, India

Received Date: May 19, 2024
Accepted Date: June 10, 2024
Published Date: July 05, 2024

**Citation:** Preeti Singh, Yashraj Singh, Aditya Singh, Adnan Ahmad. Revolutionizing Document Analysis Using AI Based Text Extraction Methods. Current Trends in Information Technology. 2024; 14(2): 40–46p.

 

serves as a comprehensive exploration into document analysis, search engine evolution, and the innovative technologies shaping information management.

## WHAT ARE SEARCHING TOOLS?

Searching: These tools enable the retrieval of information from extensive datasets. Traditional tools like library catalogs transitioned into digital platforms like Google and Yahoo, which use complex algorithms to index and rank web pages. This evolution revolutionizes access to information and advances document analysis.

## HISTORY OF SEARCHING TOOLS

The history of searching tools is a fascinating journey that spans centuries, reflecting the evolution of information retrieval from traditional methods to the digital age [1, 2].

### Pre-Digital Era

Before the digital revolution, searching tools were intricately tied to physical repositories like libraries and archives. Libraries employed meticulous cataloging systems, where librarians manually indexed books and documents. Users relied on card catalogs and manual indices to navigate these expansive collections.

### Early Digital Search Tools

The transition to digital searching began with early online tools such as Archie (Archive without the V) and Gopher in the late 20th century. Archie, created in 1990, facilitated file searches in FTP archives, while Gopher organized information hierarchically, paving the way for more structured access to digital content.

### 1990s: Rise of Web Search Engines

The 1990s witnessed a paradigm shift with the rise of web search engines. Yahoo, initially a directory service, evolved into a search engine, categorizing web content. In 1998, Google was founded, introducing the revolutionary PageRank algorithm. This link analysis-based algorithm substantially improved the relevance and quality of search results.

### 2000s: Advancements in Algorithmic Search

The 2000s marked the era of algorithmic search engines dominating the online landscape. Google continued to refine its algorithms, introducing innovations such as RankBrain, a machine learning algorithm enhancing query interpretation, and BERT (Bidirectional Encoder Representations from Transformers), which improved natural language understanding, capturing the context of words in search queries.

### Present: Diversification of Searching Tools

In the contemporary landscape, search tools have diversified beyond traditional search engines. Specialized tools, such as document management systems, content databases, and advanced natural language processing models like ChatGPT, have been developed. These tools cater to specific information retrieval needs, reflecting a nuanced and sophisticated approach to document analysis and comprehension.

### Future Trends

Looking ahead, the future of searching tools is likely to involve even more advanced AI-driven capabilities, leveraging deep learning and neural network models. As the volume and complexity of digital content continues to grow, the evolution of searching tools will play a pivotal role in shaping how individuals and organizations navigate and extract insights from the vast digital landscape.

 

## GOOGLE SEARCH ENGINE

Google, founded in 1998 by Larry Page and Sergey Brin, revolutionized online search with its powerful and efficient search engine. The introduction of the PageRank algorithm, which analyzed links between web pages, set Google apart in providing more relevant and authoritative results [3, 4]. Over the years, innovations like RankBrain and BERT enhanced the understanding of user queries, making Google the go-to search engine globally. Its user-centric design, speed, and continuous advancements have solidified Google's position as a leading force in the digital information landscape. Today, Google processes billions of searches daily, shaping how individuals access and navigate information on the internet.

### Working of Google Search Engine

Google's search engine functions through a complex, multi-step process that includes crawling, indexing, and ranking. Here is an explanation of each step:

1. *Crawling:* Google employs automated programs known as crawlers or spiders to traverse the vast internet and locate web pages. These crawlers follow links from one page to the next, building a comprehensive index of interlinked content.
2. *Indexing:* Information from crawled pages, including text, images, and metadata, is analyzed and stored in Google's massive database.
3. *Ranking:* Google's algorithm evaluates pages' relevance and quality using factors like PageRank, Rank Brain, and BERT to understand language and context.
4. *Results Display:* Pages are ranked based on relevance to user queries, and displayed on the search engine results page (SERP) to provide accurate information.
5. *Continuous Updates:* Google regularly updates its algorithm to improve search result quality, addressing spam and incorporating AI advancements for better understanding of queries.

In essence, Google's search engine works by systematically crawling and indexing the vast web, then employing a complex ranking algorithm to deliver results that align with the user's search intent. The combination of automated crawling, sophisticated indexing, and advanced algorithms allows Google to provide fast, accurate, and contextually relevant search results to users worldwide (Table 1).

## DOCUMENT TEXT EXTRACTION

Document text extraction refers to the process of systematically retrieving and isolating text content from various types of documents, such as PDFs, images, or scanned documents [5]. The primary goal of text extraction is to convert non-editable or non-searchable documents into machine-readable formats, making the textual information within these documents accessible for analysis, search, and manipulation.

**Table 1.** Various Natural Language Processing Tools and their key features.

| Category | Tools/Formats | Key Features | Use Cases |
|---|---|---|---|
| OCR Tools | Tesseract | Text extraction from images and scanned documents | Digitizing printed documents, content analysis |
| | ABBYY FineReader | Extraction of text from PDF data tables and web content | Indexing PDFs for search engines |
| Key Features | Text extraction from images | Extracting actionable intelligence from documents | Content analysis, web scraping, data mining |
| | Scanned documents processing | Document organization | Intelligence business analytics, content discovery |
| | OCR capabilities | Processing natural language metadata | Document retrieval, information automation |
| | Text extraction | Extraction classification-based tasks | Workflow automation, enterprise content management system |

 

The extraction process involves the use of Optical Character Recognition (OCR) technology in the case of scanned documents or images. OCR software analyzes the visual representation of characters and transforms them into machine-encoded text. For born-digital documents like PDFs, text extraction can be a straightforward process, as the text is already encoded in a digital format, and tools can directly access and extract the textual content.

**Text Extraction is a Crucial Component in Various Applications, Including**

1. *Search Engines:* Enabling the indexing and searchability of document content on the web or within databases.
2. *Information Retrieval:* Extracting specific details from documents to answer user queries or retrieve relevant information.
3. *Data Analysis:* Converting textual data from documents into a structured format for analysis and insights.
4. *Document Management:* Enabling efficient organization and categorization of documents based on their textual content.
5. *Automation:* Supporting automated workflows and processes that require the extraction of specific information from documents.

The advancements in Natural Language Processing (NLP) and machine learning have further enhanced the accuracy and efficiency of document text extraction [6], allowing for more nuanced understanding of language, context, and structure within documents.

## WHAT ARE AI CHATBOTS?

AI chatbots are advanced computer programs designed to simulate human-like conversations. They employ technologies such as natural language processing (NLP) and machine learning (ML) to comprehend and address user queries [7]. By comprehending language nuances, chatbots interact naturally with users via speech or text. They consistently learn from interactions, adjusting to comprehend intent and offer pertinent responses. Context awareness enables chatbots to remember past interactions, enhancing personalization and efficiency. They serve various functions, from answering FAQs to executing tasks, finding applications in customer support and business automation. Chatbots are adaptable across platforms, including websites and messaging apps, with some offering voice recognition capabilities. Continuous learning ensures chatbots stay updated with user behaviors, enhancing their effectiveness in improving customer service and streamlining tasks.

### Different AI Chatbots Present

In 2023, notable AI chatbots showcased advancements in conversational AI. ChatGPT, part of the GPT family, excelled in generating contextually relevant responses from diverse internet text. Google's Bard aimed for natural, context-aware interactions, reflecting progress in understanding nuanced user inputs. These chatbots, including ChatGPT and Bard, mark significant milestones in conversational AI evolution, demonstrating prowess in generating human-like language (Table 2).

**Table 2.** Comparison of different chatbots.

| Chatbot Name | Developer | Features | Natural Language Understanding | Context Management | Integration Platforms | Notable Advancements |
|---|---|---|---|---|---|---|
| ChatGPT | Open AI | Language generation, Context awareness | Yes | Yes | Varies | Versatility in open-ended conversations |
| Google's Bard | Google | Natural language understanding, Context awareness | Yes | Yes | Google platforms | Sophisticated language understanding |
| Other AI Chatbots of Interest | Various | Varied features and capabilities | Varies | Varies | Varies | Differentiated by specific advancements |

 

**ChatGPT**

ChatGPT, part of OpenAI's GPT series, is an advanced AI language model trained on diverse internet text [8]. It is designed for generating human-like responses in conversations, showcasing context awareness and versatility. Users engage by providing prompts, and ChatGPT generates coherent, contextually relevant responses. This model marks a notable progression in conversational AI, though updates beyond January 2023 should be taken into account (Figure 1).

**Working of ChatGPT**

The working of ChatGPT involves pre-training, fine-tuning, and a decoding mechanism for generating responses [9]. Initially pre-trained on a diverse internet dataset, ChatGPT learns grammar, context, and language structure. Constructed on a transformer architecture, it demonstrates proficiency in capturing extensive dependencies within sequences. Fine-tuning on a narrower dataset shapes its behavior according to ethical and safety standards. Using a decoding mechanism, ChatGPT generates responses based on user inputs, maintaining context for coherent conversations. While it can produce impressive responses, it may have limitations such as sensitivity to input phrasing and verbosity. OpenAI implements safety measures, including fine-tuning and the Moderation API, to address concerns.

**AI DOCUMENT ANALYZER CHATBOT**

An AI document analyzer chatbot utilizes artificial intelligence and natural language processing to extract information from documents through a conversational interface [10]. Users can upload documents and engage in natural language conversations to query and retrieve specific details.

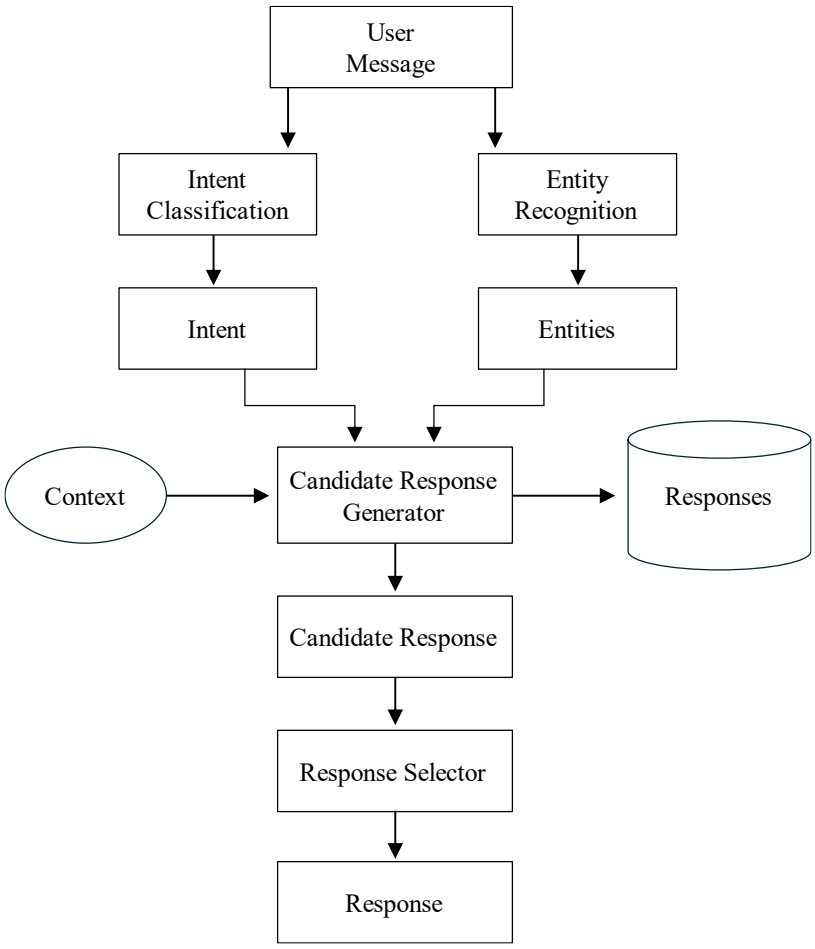


**Figure 1.** ChatGPT working methodology.

 

The chatbot demonstrates exceptional performance in extracting text, understanding user inquiries, and delivering pertinent responses while retaining context. With its user-friendly interface and AI integration, it streamlines document analysis for various applications, including business processes and education.

## OUR PROPOSED SYSTEM

The envisioned AI Document Analyzer represents an ambitious and transformative venture, poised to reshape the landscape of document interaction and comprehension. At its core, the proposed application harnesses the collective power of state-of-the-art technologies, including Next.js, DrizzleORM, OpenAI, Stripe, TypeScript, and Tailwind, to create a dynamic and intelligent platform (Figure 2).

The heart of this project lies in its ability to seamlessly extract and comprehend textual information from a diverse range of document types. The integration of the ChatGPT API allows users to engage in natural language conversations with their uploaded documents, unlocking a novel way to interact with and extract valuable insights from textual content. This includes the capability to generate concise summaries for documents, providing users with a quick and informative overview. Moreover, the AI Document Analyzer extends its functionality beyond mere textual extraction. For documents containing statistical information, the application goes a step further by automatically generating visually appealing charts and graphs. This feature not only improves data representation but also enables a more profound comprehension of intricate information.

Next.js ensures a responsive and efficient rendering process, while DrizzleORM streamlines database operations for seamless data management. OpenAI powers the natural language processing capabilities, enabling intelligent and context-aware conversations. Stripe guarantees secure transactions. In essence, the proposed AI Document Analyzer is not just a technological solution, it is an immersive and transformative tool designed to elevate the user's experience with documents.

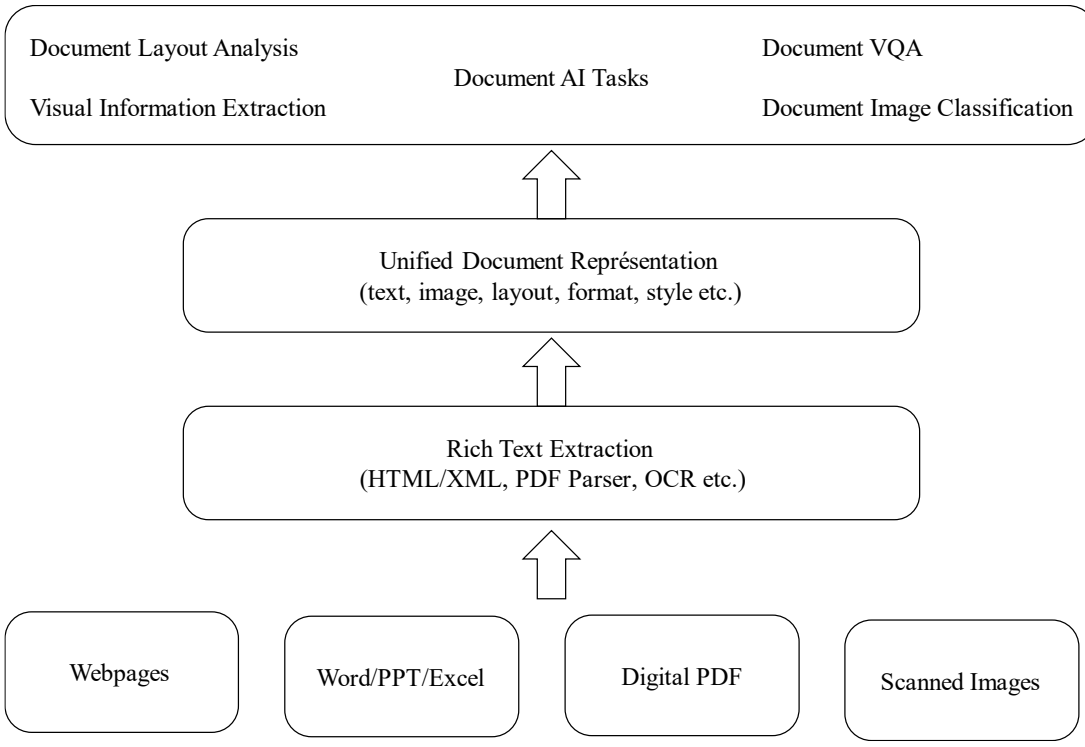


**Figure 2.** Proposed AI Document analyzer working methodology.

 

## CONCLUSION

In conclusion, the AI Document Analyzer project represents a significant advancement in document comprehension and user interaction, blending cutting-edge technologies to redefine how we extract insights from textual data. By integrating Next.js, DrizzleORM, OpenAI, Stripe, TypeScript, and Tailwind, the system introduces a chat-based interface powered by ChatGPT API, enabling natural language conversations with uploaded documents. This transformative SaaS solution not only offers functionalities like summarization and data visualization but also extends its capabilities to foster a user-centric design. Through a comparative exploration of search engine evolution, AI-driven conversational agents, and document text extraction techniques, this study contextualizes the AI Document Analyzer within the broader landscape of information management. Ultimately, the proposed system stands as a testament to the potential of innovative solutions in revolutionizing document analysis and comprehension.

## REFERENCES

1. Solanki Amrish. Advancements in Artificial Intelligence: A Comprehensive Review and Future Prospects. International Journal of Artificial Intelligence Research and Development (IJAIRD). 2024; 2(1): 53–64.
2. Khan MS, Ahmad I. Herbal medicine: current trends and future prospects. In New look to phytomedicine. Academic Press. 2019 Jan 1; 3–13.
3. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. Article No.: 159, Pages 1877-1901.
4. Seymour Tom, Frantsvog Dean, Kumar Satheesh. History Of Search Engines. International Journal of Management & Information Systems (IJMIS). 2011; 15(4): 47–58. 10.19030/ijmis.v15i4.5799.
5. Følstad Asbjørn, Brandtzaeg Petter. Chatbots and the new world of HCI. Interactions. 2017; 24(4): 38–42. 10.1145/3085558.
6. Maithili K, Raja SN, Kumar RR, Koli S. A Survey (NLP) Natural Language Processing and Transactions on (NNL) Neural Networks and learning Systems. In E3S Web of Conferences. EDP Sciences. 2023; 430: 01148.
7. Khan Muskan, Ning Chu, Chang Jung. Natural Language Processing Techniques for Enhancing Information Systems Management. 2023.
8. Dong-Min Park, Seong-Soo Jeong, Yeong-Seok Seo. Systematic Review on Chatbot Techniques and Applications. J Inf Process Syst. 2022; 18(1): 26–47. Available from: https://jips-k.org/pub-reader/707
9. Kuhail MA, Bahja M, Al-Shamaileh O, Thomas J, Alkazemi A, Negreiros J. Assessing the Impact of Chatbot-Human Personality Congruence on User Behavior: A Chatbot-based Advising System Case. IEEE Access. 2024 May 20; 12: 71761–71782.
10. Klopfenstein Lorenz, Delpriori Saverio, Malatini Silvia, Bogliolo Alessandro. The Rise of Bots: A Survey of Conversational Interfaces, Patterns, and Paradigms. DIS'17: Proceedings of the 2017 Conference on Designing Interactive Systems. 2017; 555–565. 10.1145/3064663.3064672.

 