International Conference on Advances in Modern Technology of Research in Engineering Field (AIMTREF) April, 2026
Organized By: SKN Sinhgad Institute of Technology& Science, Kusgaon (Bk), Lonavala, Pune

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

An AI-Based Document Analyzer Using Gemini API for Conversational Knowledge Retrieval

Prof. S. P. Gunjal, Kunal Jagtap, Hrishikesh Joshi, Bhavya Bhat, Anushka Gaikwad
Department, Department of Computer Engineering, SKN Sinhgad Institute of Technology and Science, Lonavala, Maharashtra

Abstract- The rapid growth of digital documents across domains such as research, business, and education has created significant challenges in efficient information extraction. Traditional methods relying on manual reading or keyword-based search lack contextual understanding and are time-consuming. This paper presents PaperSense, an AI-powered document analyzer that enables users to upload documents and interact with them using natural language queries. The system integrates the Google Gemini 2.5 Flash model for contextual understanding, summarization, and question answering. Built using Python and Streamlit, the system follows a modular architecture separating frontend and backend processing. Experimental evaluation demonstrates improved efficiency in document comprehension and faster information retrieval. The system provides a scalable and user-friendly solution for intelligent document analysis.

Keywords- Generative AI, Document Analysis, Gemini API, Natural Language Processing, Conversational AI, Streamlit

I. INTRODUCTION

The exponential growth of digital documents has led to information overload, making manual analysis inefficient and time-consuming. Researchers and professionals often spend excessive time extracting relevant insights from large documents. As the volume of information increases, traditional approaches fail to keep pace with the demand for quick and accurate understanding of content. Recent advancements in Large Language Models have significantly improved the ability of machines to understand and process textual data with contextual awareness. These models enable tasks such as summarization, semantic understanding, and question answering. In this context, PaperSense is introduced as a system that transforms static documents into interactive conversational systems.

By allowing users to query documents using natural language, the system enhances accessibility and efficiency in information retrieval.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

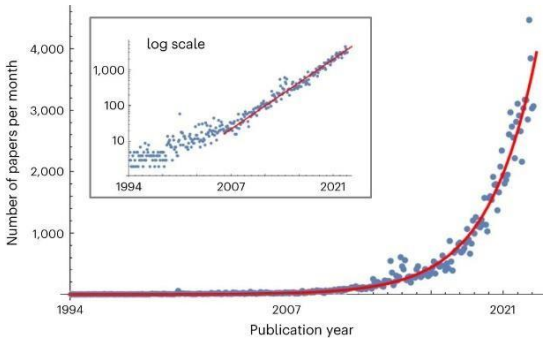

Fig. 1. Growth of digital data and inefficiency of manual analysis.

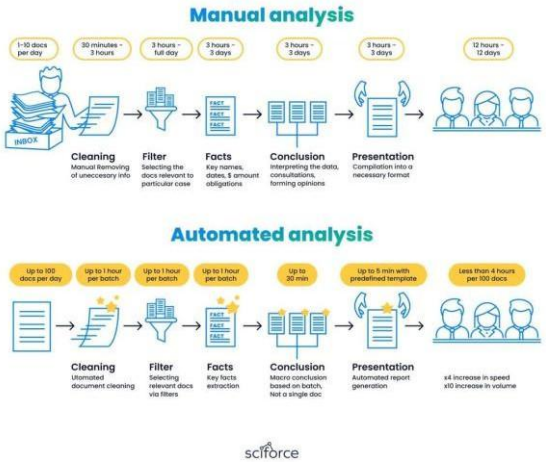

Fig. 2. Manual vs Automated Analysis Process

II. LITERATURE REVIEW

Table I. Comparison of existing NLP models.

Model	Strength	Limitation
Transformer	Parallel processing	High complexity
BERT	Bidirectional context	Expensive training
SciBERT	Scientific domain optimized	Limited generalization
GPT	Strong text generation	Limited bidirectional context

The development of modern natural language processing systems is largely influenced by foundational models such as Transformer, BERT, SciBERT, and GPT. The Transformer architecture introduced parallel processing capabilities that significantly improved training efficiency but suffered from high computational complexity. BERT improved contextual understanding by introducing bidirectional encoding; however, it required extensive computational resources for training. SciBERT extended this approach to scientific text, achieving improved performance in domain-specific tasks but limiting its general applicability. GPT introduced generative capabilities, enabling text generation and conversational interaction, but lacked full bidirectional context understanding. Although these models have advanced the field of NLP, existing systems still lack an integrated framework that supports real-time conversational interaction with documents and multi-format analysis.

III. PROBLEM STATEMENT

The increasing volume of unstructured textual data has created significant challenges in extracting meaningful insights efficiently. Traditional document analysis systems lack the ability to understand

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

context, process large documents effectively, and support natural language interaction. These limitations result in increased time consumption and reduced accuracy in information retrieval. Therefore, there is a need for a system that can intelligently analyze documents and provide context-aware responses through a conversational interface.

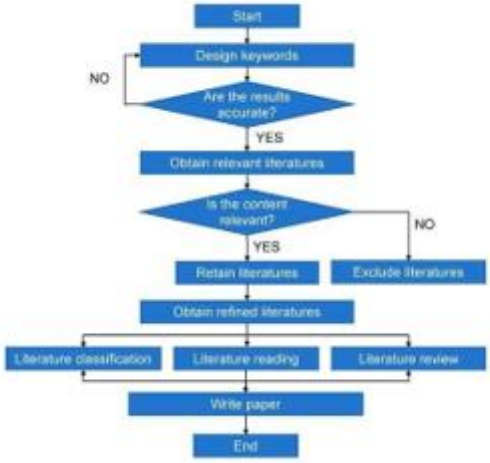

Fig. 3. Literature Review Workflow Diagram

IV. PROPOSED SYSTEM

PaperSense is designed as a web-based application that enables users to upload documents and interact with them using conversational artificial intelligence. The system processes the uploaded document and converts it into a format suitable for analysis by a large language model. Once processed, the document content is sent to the Gemini API, which performs contextual analysis and generates responses based on user queries. The workflow begins with document upload, followed by preprocessing and transmission to the AI model. The user then inputs a query, and the system generates a relevant response that is displayed through a chat interface. This approach allows users to extract information efficiently without manually reading the entire document..

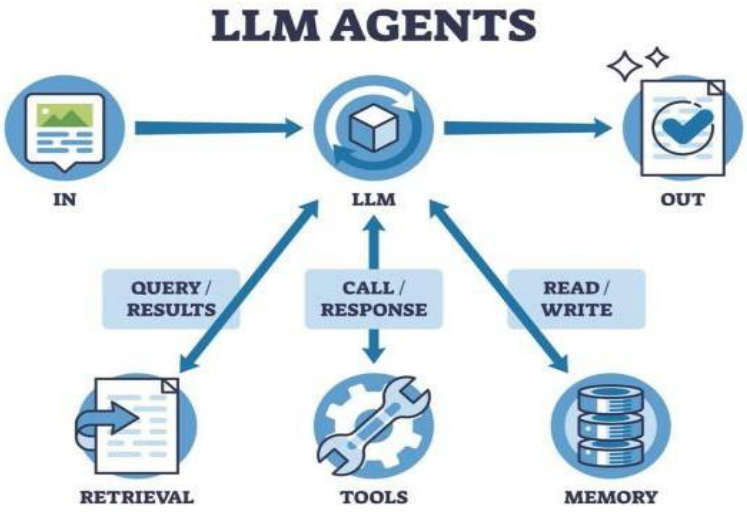

Fig. 4. System workflow of PaperSense.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

V. SYSTEM ARCHITECTURE

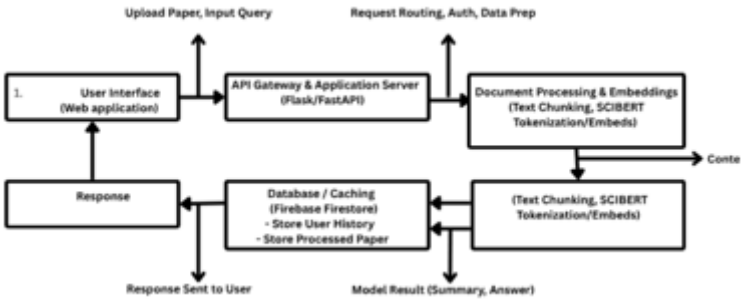

Fig. 5. Architecture of PaperSense.

The system architecture follows a layered modular design consisting of a presentation layer, an application layer, a processing layer, and an AI layer. The presentation layer is implemented using Streamlit and provides an interactive interface for file upload and chat interaction. The application layer manages the workflow and session state, ensuring smooth interaction between the user and the system. The processing layer is responsible for engage in multi-turn interactions with the document. This enhances the overall user experience by providing continuity in the conversation.

VI. ALGORITHM DESIGN

The system operates through a series of algorithms that manage client initialization, document upload, query processing, and file cleanup. The client initialization process involves retrieving and validating the API key before establishing a connection with the Gemini service. The document upload algorithm ensures that files are temporarily stored, uploaded to the API, and then removed from local storage to maintain efficiency. The query processing algorithm combines the user query with document content and sends it to the AI model for response generation. The cleanup algorithm ensures that uploaded files are deleted after use, preventing unnecessary resource consumption.

converting uploaded documents into structured text and preparing them for analysis. The AI layer integrates the Gemini API, which performs document understanding and generates responses. This modular architecture ensures scalability, maintainability, and efficient resource utilization.

VII. METHODOLOGY

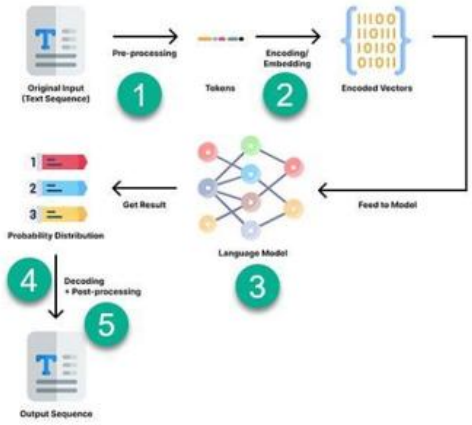

Fig. 6.. Document processing and AI pipeline.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

The methodology involves document processing, AI integration, and conversational interaction. The uploaded file is temporarily stored and converted into structured text using Python-based processing techniques. The processed content is then passed to the Gemini 2.5 Flash model, which is capable of handling large context windows and generating accurate responses.

Implementation

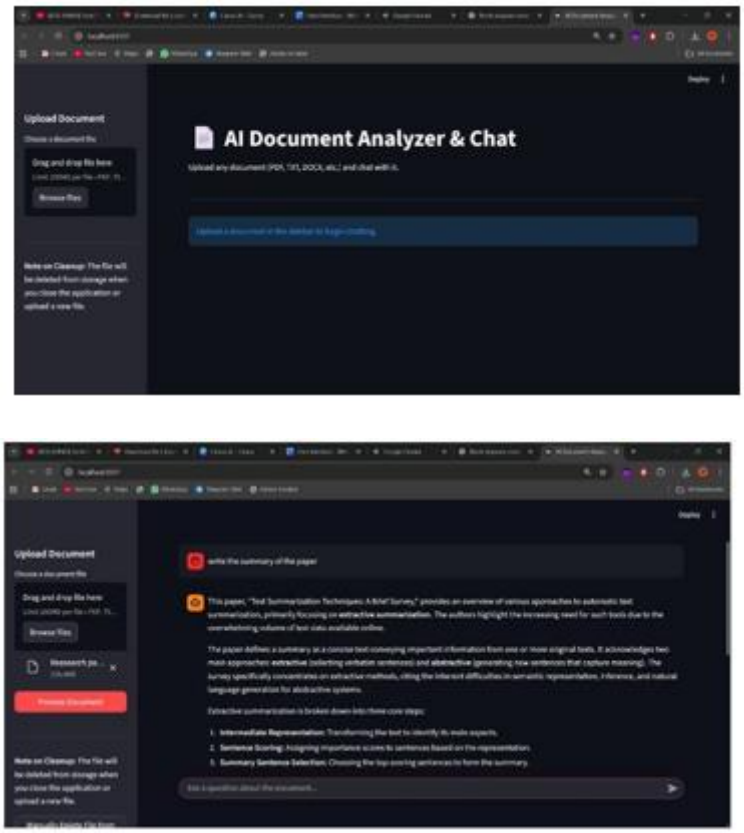

Fig. 7. User interface of the system.

The implementation of PaperSense is carried out using Python and Streamlit. The frontend is developed using Streamlit, providing an intuitive interface for file upload and chat-based interaction. The backend is implemented in a modular structure where the main application file handles user interaction and control flow, while a separate engine module manages AI processing and file operations.

The system incorporates error handling mechanisms to ensure stability and reliability during execution.

VIII. RESULTS AND DISCUSSION

The system was evaluated using various document types, including research papers and spreadsheets. The results indicate that PaperSense significantly reduces the time required for document analysis while maintaining high accuracy in responses. The conversational interface allows users to retrieve information quickly and efficiently. The system demonstrates improved performance compared to traditional manual analysis methods, particularly in terms of response time and contextual understanding.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

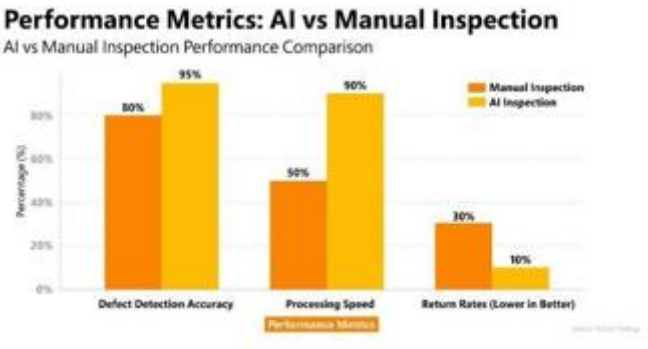

Fig. 8. Performance comparison between manual and AI analysis.

Observations

- Faster response time
- Improved accuracy
- Reduced manual effort

Advantages

- Real-time document interaction
- Context-aware responses
- Multi-format support
- User-friendly interface

Limitations

- API dependency
- Session-based memory
- Limited offline capability

Future Scope

- Multi-document analysis
- Multimodal support
- Persistent storage
- RAG integration

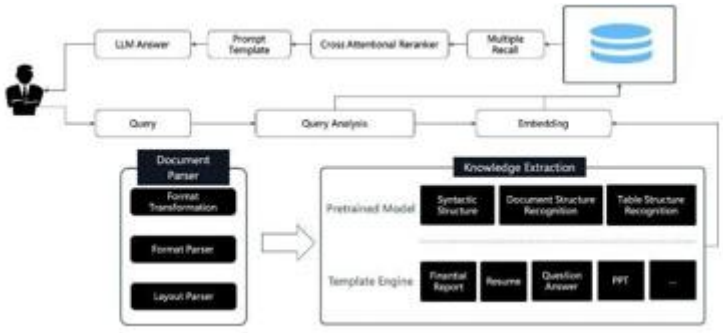

Fig. 9. Future system enhancements.

IX. CONCLUSION

PaperSense demonstrates the effective use of Generative AI in document analysis. By combining conversational AI with document processing, the system enables efficient and accurate information retrieval. The architecture ensures scalability and usability across multiple domains.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

REFERENCES

1. A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, 2017.
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL, 2019.
3. I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," EMNLP, 2019.
4. A. Radford et al., "Improving Language Understanding by Generative Pre-Training," OpenAI, 2018.
5. M. Allahyari et al., "Text Summarization Techniques: A Brief Survey," arXiv preprint arXiv:1707.02268, 2017.
6. C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," ACL Workshop, 2004.
7. T. Brown et al., "Language Models are Few-Shot Learners," NeurIPS, 2020.
8. K. Clark et al., "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," ICLR, 2020.
9. Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," EMNLP, 2019.
10. A. See, P. J. Liu, and C. D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks," ACL, 2017.
11. S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, 2009.
12. J. Guu et al., "REALM: Retrieval-Augmented Language Model Pre-Training," ICML, 2020.
13. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.
14. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT Networks," EMNLP, 2019.
15. J. K. Kenter and M. de Rijke, "Short Text Similarity with Word Embeddings," CIKM, 2015.
16. OpenAI, "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.
17. Google, "Gemini: A Family of Highly Capable Multimodal Models," Google Research, 2023.
18. R. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," ICLR, 2013.
19. T. Kudo and J. Richardson, "SentencePiece: A Simple and Language Independent Subword Tokenizer," EMNLP, 2018.
20. J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," EMNLP, 2014.