iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.60840

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue IV Apr 2024- Available at www.ijraset.com

PDF-Driven Q&A: A Research Paper

Prof. Sneha R. Sontakke[1], Shreyash V. Gondane[2], Anup S. Nandedkar[3], Sahil T. Chauhan[4], Nama Z. Choudhari[5]
[1]*Assistant Professor at P R Pote College of Engineering and Management Amravati*
[2, 3, 4, 5]*Student at P R Pote College of Engineering and Management Amravati*

Abstract: ***Traditional methods of PDF retrieval often suffer from inefficiencies and inaccuracies due to their reliance on keyword-based search algorithms. These methods usually don't understand what words really mean and have trouble with things like different words meaning the same thing, one word meaning many things, and understanding the context. This makes the search results not very good. This project proposes a novel approach to address these shortcomings by developing a comprehensive system for efficient PDF document management and context-aware question answering. The system integrates various components, including a user-friendly interface for PDF upload, Longchain techniques for breaking down lengthy PDFs into manageable chunks, and vectorization using OpenAI's language models. These vectorized chunks are stored in the Faiss Vector Database for rapid retrieval. User queries are converted into vectors using the same language model, and a semantic search algorithm matches them against the stored PDF chunks to retrieve contextually relevant answers. By presenting these answers in a user-friendly format, the system aims to enhance accessibility and usability, enabling seamless access to pertinent information within PDF documents.***

Keywords: Chatbot, PDF retrieval Information, Longchain Technique ,Faiss Vector Database, Chunks.

I. INTRODUCTION

In an era marked by the exponential growth of digital information, effective management and retrieval of knowledge from vast repositories such as PDF documents pose significant challenges. To address this need, we propose a methodology for implementing a comprehensive system that enables users to upload PDF files, automatically breaks them down into manageable "chunks," and facilitates context-aware question answering. This methodology encompasses two key components: PDF Upload and Chunking, and Context-Aware Question Answering. The former focuses on providing users with a user-friendly interface for uploading PDF files and employs advanced techniques to segment lengthy documents into smaller, more digestible units. Leveraging state-of-the-art technologies such as OpenAI's language models and the Faiss Vector Database, this component ensures efficient storage and retrieval of the vectorized PDF chunks. The latter component, Context-Aware Question Answering, addresses the challenge of retrieving relevant information from the stored PDF chunks in response to user queries. By converting queries into vectors using the same model employed for PDF chunk vectorization, the system maintains consistency in representation. A semantic search algorithm is then utilized to match these vectors against the stored chunks, considering semantic similarity to deliver contextually relevant answers to the users. Through this methodology, we aim to provide a structured approach for enhancing information accessibility and user experience, empowering individuals to efficiently navigate and extract insights from large repositories of PDF documents

II. LITERATURE SURVEY

PDF summarization software has evolved over the years in response to the increasing volume of digital information and the need for efficient information retrieval methods. The history of PDF summarization software can be traced back to the emergence of PDF (Portable Document Format) as a widely used format for sharing and distributing documents digitally.

PDF was introduced by Adobe Systems in the early 1990s as a file format that preserves the formatting of documents across different platforms and devices. Its popularity grew rapidly due to its ability to maintain document fidelity and its widespread compatibility. As the use of PDF documents became more prevalent, users encountered challenges in managing and extracting information from these files. PDF documents often contain large volumes of text and complex structures, making it difficult to locate and retrieve specific information efficiently.

Early attempts to address these challenges involved manual methods of document analysis and extraction. Users would manually sift through PDF documents to identify key information, a process that was time-consuming and prone to errors. The development of advanced NLP and ML techniques in the late 20th and early 21st centuries revolutionized the field of document summarization. These technologies enabled the creation of algorithms capable of understanding the semantic meaning of text and identifying key concepts within documents.

©IJRASET: All Rights are Reserved

[Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019] introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications.[1]

[Sakib Shahriar and Kadhim Hayawi, June 2023] introduces the advent of AI-empowered chatbots capable of constructing human-like sentences and articulating cohesive essays has captivated global interest. This paper provides a historical perspective on chatbots, focusing on the technology underpinning the Chat Generative Pre-trained Transformer, better known as ChatGPT. We underscore the potential utility of ChatGPT across a multitude of fields, including healthcare, education, and research.[2]

[Xingxing Zhang, et al., July 2019] Study neural extractive summarization models usually employ a hierarchical encoder for document encoding and they are trained using sentence-level labels, which are created heuristically using rule-based methods. we propose HIBERT (as shorthand for Hierachical Bidirectional Encoder Representations from Transformers) for document encoding and a method to pre-train it using unlabeled data. We apply the pre-trained HIBERT to our summarization model and it outperforms its randomly initialized counterpart by 1.25 ROUGE on the CNN/Dailymail dataset and by 2.0 ROUGE on a version of New York Times dataset.[3]

[Divakar Yadav, et al,. March 2022]Study one of the most pressing issues that have arisen due to the rapid growth of the Internet is known as information overloading. Simplifying the relevant information in the form of a summary will assist many people because the material on any topic is plentiful on the Internet. Manually summarising massive amounts of text is quite challenging for humans. So, it has increased the need for more complex and powerful summarizers. Researchers have been trying to improve approaches for creating summaries since the 1950s, such that the machine-generated summary matches the human-created summary. This study provides a detailed state-of-theart analysis of text summarization concepts such as summarization approaches, techniques used, standard datasets, evaluation metrics and future scopes for research.[4]

III. METHODOLOGY

A. Work-Flow of Bot

Before proceeding to the implementation phase, one must be aware of the robot's workflow. Let's examine Fig. 1, which illustrates the basic workflow of the PDF-Driven Q&A.

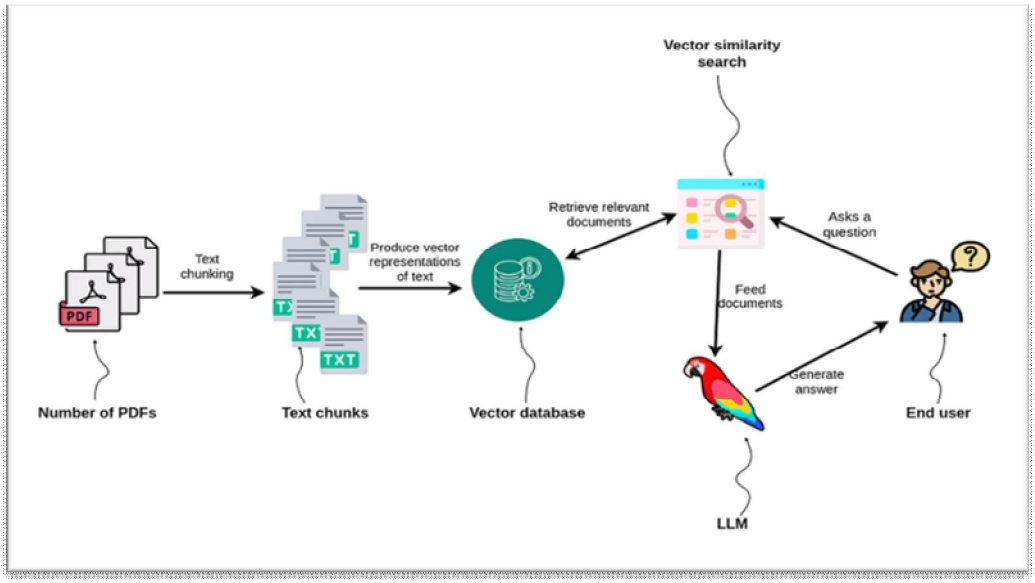


Figure 1: Work Flow

The proposed system is a Document reader designed to enhance document comprehension and user engagement. Leveraging two advanced large language models (LLMs), GPT-3.5-turbo and instructor-xl, the system offers advanced functionalities for document summarization and interactive exploration.

1) *Phase 1(data preprocessing): Document Processing and Chunking:*

Upon uploading PDF files through a user-friendly interface, the system employs Long chain techniques to break down lengthy documents into manageable "chunks" based on logical divisions such as paragraphs or sections. This process enhances the granularity of search and improves the user experience by providing focused content.

2) *Phase 2 (Model training)*

a) *Vectorization with OpenAI:* Each chunk of PDF content undergoes transformation into high-dimensional vectors using OpenAI's state-of-the-art language models. These vectors capture the semantic essence of the textual content, enabling advanced search techniques and facilitating accurate summarization.

b) *Faiss Vector Database:* The vectorized PDF chunks are efficiently stored in the Fuiss Vector Database, known for its scalability and rapid retrieval capabilities. Fuiss provides a robust foundation for organizing and accessing the vectorized content, ensuring efficient storage and retrieval operations.

3) *Phase 3(Interactive Question Answering)*

a) *Context-Aware Question Answering:* Users can pose questions or queries to the system, which converts these queries into vectors using the same OpenAI model employed for PDF chunk vectorization. Utilizing the cosine similarity metric, the system performs vector matching against the Fuiss-stored PDF chunks to retrieve contextually relevant answers. This approach ensures accurate and context-aware responses to user queries, enhancing the overall usability and effectiveness of the system.

b) *User Interface and Experience:* The system features a user-friendly interface that simplifies the process of uploading documents, posing queries, and interacting with the document content. Clear instructions and intuitive design elements guide users through the process, ensuring a seamless and engaging user experience.

Further details are provided in the snapshots given below:

➢ *User Interface Layout / Home Page*

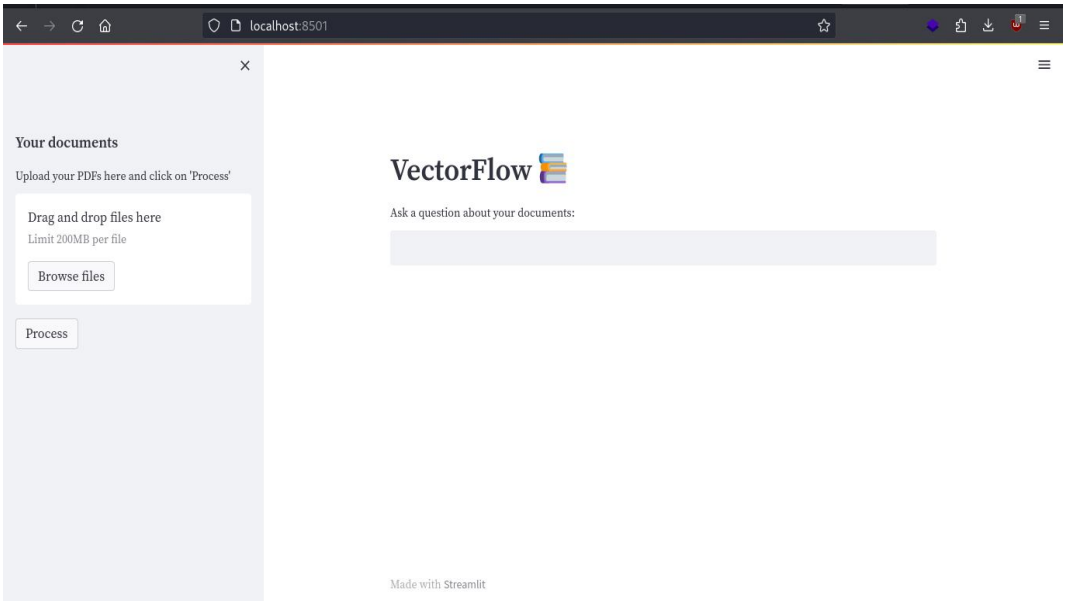


Figure 2: User Interface Layout / Home Page

➢ *Uploading Page*

Upload your PDF documents to platform in just a few clicks. Simply select them from your computer, and system will quickly process them. Once uploaded, you can immediately access advanced features for document summarization and question answering. Streamline your workflow and unlock valuable insights from your PDFs effortlessly with intuitive uploading process.

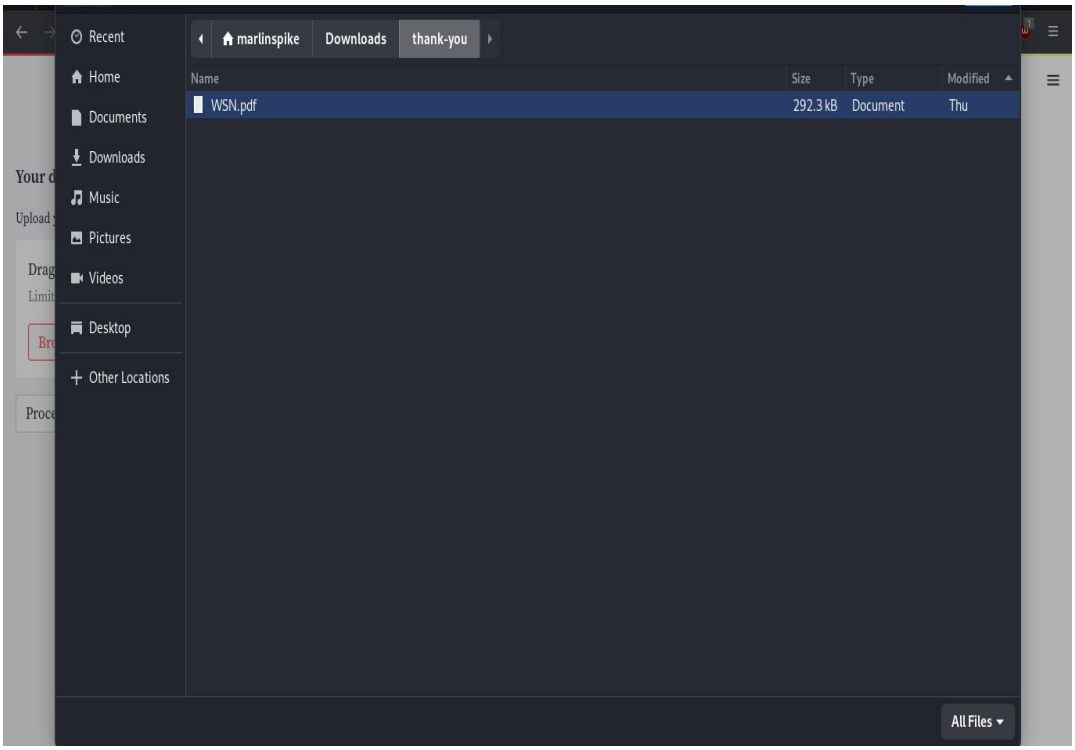


Figure 3: Uploading Page

➢ *Document Processing /Model Training*

Document processing involves extracting text and structural information from PDF files. This data is then used to train machine learning models, such as natural language processing models, to analyze and understand the content. Through iterative training, these models learn to recognize patterns, extract key information, and improve document comprehension accuracy, enhancing the overall functionality and usability of the PDF reader system.

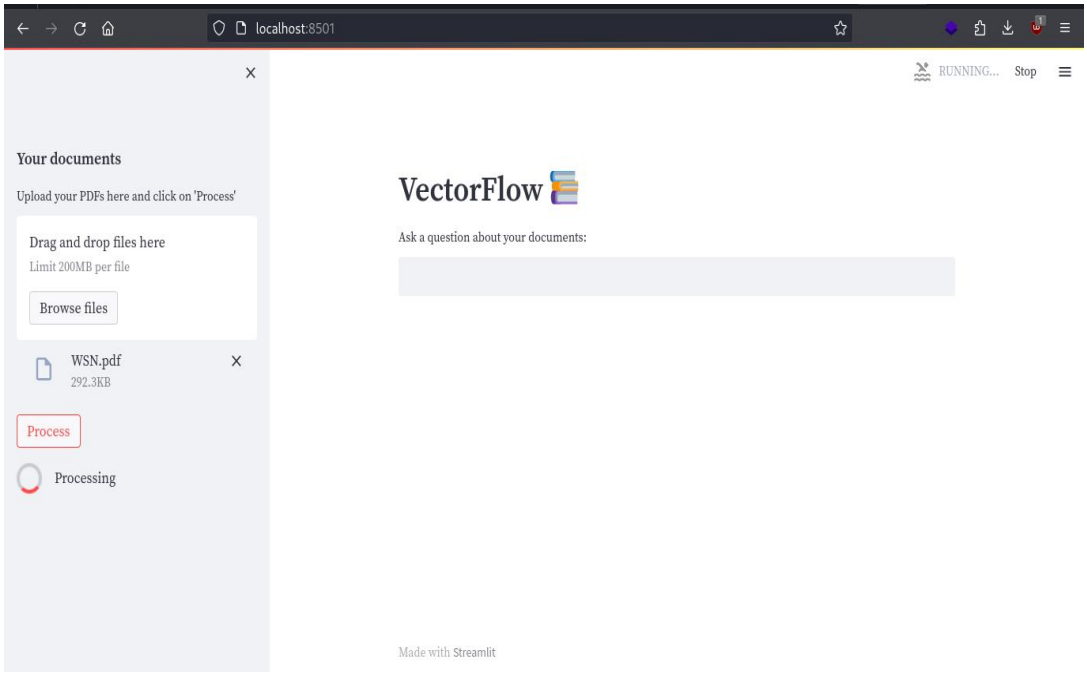


Figure 4: Document Processing /Model training

IV. RESULT

1) *User Interaction / Interactive Question Answering:* In this phase, users can harness the power of system to obtain precise answers to their queries directly from their PDF documents. Simply input your question in natural language, and our system will analyze the content of your uploaded PDFs to provide accurate responses.

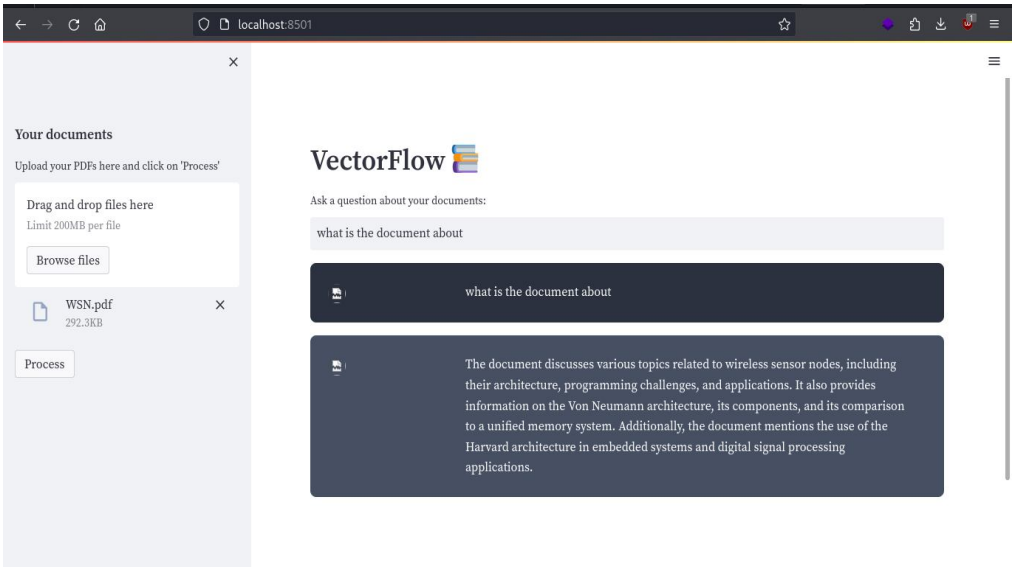


Figure 5: User Interaction / Interactive Question answering

V. LIMITATION

1) *Technical Limitations*: AI-powered chatbots may encounter technical limitations such as processing speed, memory constraints, or compatibility issues with certain document formats or languages.
2) *Handling of Non-Textual Content:* PDFs can include non-textual content such as images, graphs, and charts, which may contain valuable information relevant to the user's query. However, the chatbot may lack the capability to analyze and interpret such content, limiting its ability to provide comprehensive answers based solely on textual information extracted from the PDF
3) *Integration Complexity:* Integrating AI document readers into existing workflows and systems can be complex and time-consuming. Compatibility issues, data synchronization, and customization requirements may arise during the integration process, requiring specialized technical expertise and resources

VI. CONCLUSION

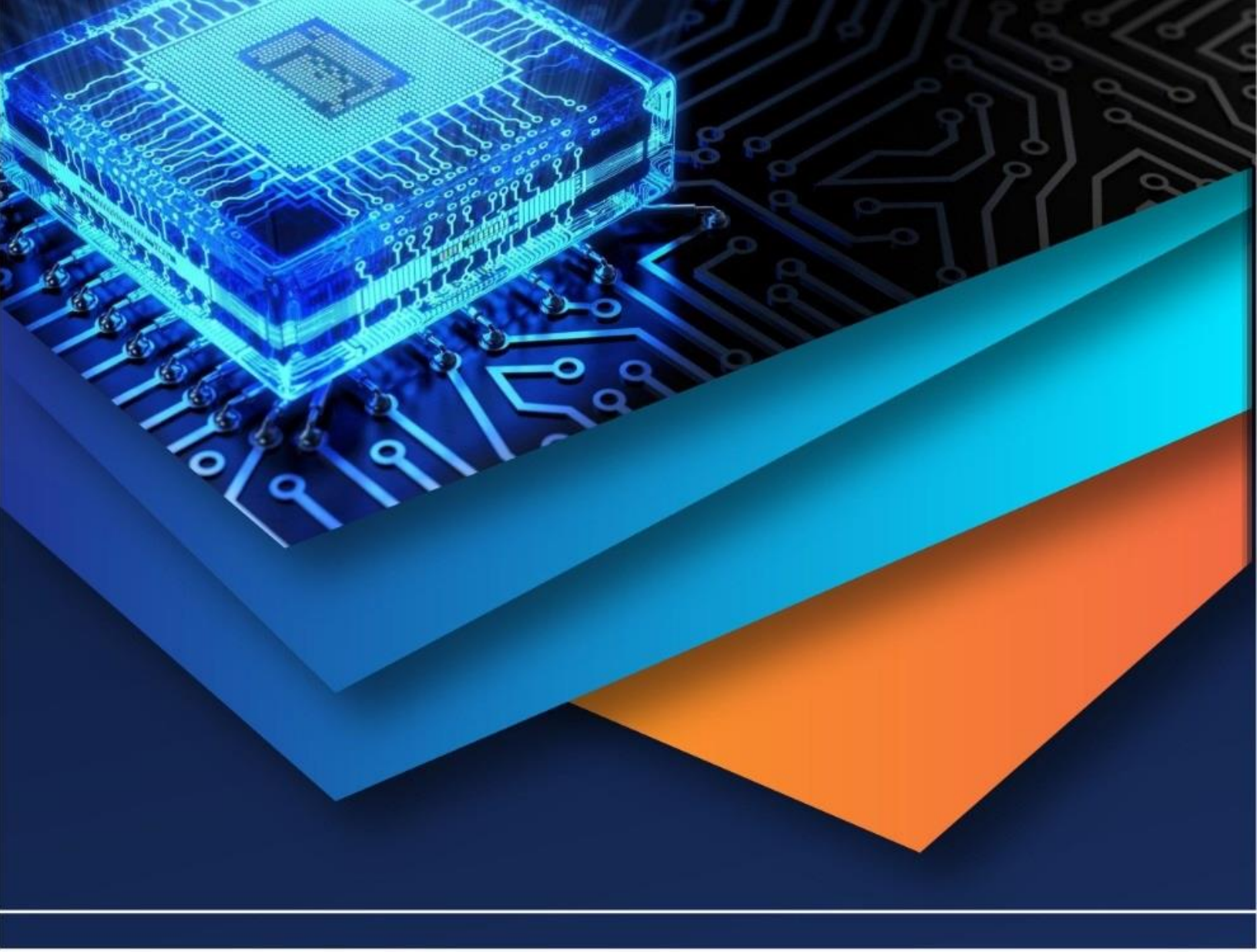
PDF-driven question-and-answer (Q&A) retrieval systems play a pivotal role in modern information management by efficiently extracting and categorizing data from documents. Leveraging advanced algorithms, these systems automate the processing of PDF files, identifying pertinent information and structuring it for easy access and analysis. However, challenges such as initial investment requirements, data dependency, and technical limitations pose obstacles to their implementation. Despite these challenges, PDF-driven Q&A retrieval systems offer invaluable solutions, streamlining information retrieval processes and enhancing efficiency within organizations. By automating tasks, reducing manual intervention, and facilitating rapid access to relevant data, these systems contribute to heightened productivity, informed decision-making, and optimized resource utilization. Continuous advancements in artificial intelligence and natural language processing further augment their capabilities, reinforcing their status as indispensable tools for modern enterprises.

REFERENCES

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). DOI: 10.18653/v1/N19-1423
[2] Shahriar, S., & Hayawi, K. (2023). Let's Have a Chat! A Conversation with ChatGPT: Technology, Applications, and Limitations. Artificial Intelligence and Applications. DOI: 10.47852/bonviewAIA3202939.
[3] Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

[4] Yadav, D., Desai, J., & Yadav, A. K. (2022). Automatic Text Summarization Methods: A Comprehensive Review. arXiv. Retrieved from https://doi.org/10.48550/arXiv.2204.01849
[5] Lin, J., Pradeep, R., Teofili, T., & Xian, J. (2023, August 29). Vector Search with OpenAI Embeddings: Lucene Is All You Need. arXiv:2308.14963v1 [cs.IR].
[6] Pandiarajan, S., Yazhmozhi, V. M., & Praveen Kumar, P. (2015). Semantic Search Engine Using Natural Language Processing. In Proceedings of the International Conference on Advanced Computing and Communication Systems (pp. 641-649). DOI: 10.1007/978-3-319-07674-4_53.
[7] Shaikh, A., More, D., Puttoo, R., Shrivastav, S., & Shinde, S. (2019). A Survey Paper on Chatbots. International Research Journal of Engineering and Technology (IRJET), 06(04), Volume: 06 Issue: 04.
[8] Shrivastava, A. (June 2023). Understanding the Fundamental Limitations of Vector-Based Retrieval for Building LLM-Powered Chatbots (Part 1/3). ThirdAI Blog. Retrieved from [Medium].
[9] D. Hingu, D. Shah, and S. S. Udmale, "Automatic Text Summarization of Wikipedia Articles," 2015 International Conference on Communication, Information & Computing Technology (ICCICT), Mumbai, India, Jan. 16-17.
[10] X. Yang, K. Yang, T. Cui, M. Chen, and L. He, "A Study of Text Vectorization Method Combining Topic Model and Transfer Learning," ISJ Theoretical & Applied SciencE, vol. 1, no. 2, pp. XX-XX, Year.
[11] K. Singh and M. Shashi, "Vectorization of Text Documents for Identifying Unifiable News Articles," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 7, 2019.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)