International Journal of Scientific Research and Engineering Development-– Volume 7 Issue 2, Mar-Apr 2024
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

AUTOMATED QUERY ANSWERING SYSTEM USING NLP TECHNIQUE

Swathi A, Venkata Seshu Kumar B, Akash Chowdary E, Vamsi Krishna D, Ram Teja A
Department of IT
KKR & KSR Institute of Technology and Sciences, Guntur.
Email : swathiann51@gmail.com,seshukumarbalem@gmail.com,20jr1a1252@gmail.com,
20jr1a1256@gmail.com,20jr1a1238@gmail.com,

----------**********----------

Abstract:

In the realm of automated query answering systems, the integration of Natural Language Processing (NLP) techniques is a pivotal endeavor. Our project presents an innovative approach to enhancing user interactions with documents through advanced NLP methodologies. By leveraging Streamlit, a user-friendly web application framework, we offer a seamless interface for users to engage with our system. Powered by the Langchain library, our system encompasses essential NLP functionalities including document loading, text segmentation, embeddings generation, and vector storage. Through the implementation of a Conversational Retrieval Chain, our system intelligently incorporates conversation history, enabling contextually relevant responses. Moreover, the integration of Long-Short Term Memory (LSTM) models ensures continual refinement of responses based on user interactions, thereby enhancing the conversational experience. Our project aims to provide users with an intuitive platform for querying information from documents, exemplifying the practical application of advanced NLP techniques in real-world scenarios.

***Keywords*—**NLP Techniques , Streamlit , Langchain , Conversational Retrieval , Memory Management , Web Interface

----------**********----------

INTRODUCTION :-

In the era of abundant digital information, efficient access to relevant knowledge is essential. Our project focuses on developing an Automated Query Answering System (AQAS) utilizing advanced Natural Language Processing (NLP) techniques. The system aims to provide users with a seamless platform for querying information from documents, emphasizing natural language interactions. Leveraging the Streamlit framework, we prioritize user-friendliness and accessibility. The Langchain library forms the backbone of our system, enabling essential NLP functionalities such as document loading, text segmentation, and embeddings generation. A key feature of our system is the incorporation of a Conversational Retrieval Chain, enhancing user experience by leveraging historical interactions. By integrating Long-Short Term Memory (LSTM) models, our system continually refines responses, ensuring accuracy and relevance. Through this project, we address the growing demand for intelligent query answering systems, empowering users to access and extract insights from textual documents efficiently.

LITERATURE REVIEW

The paper suggests creating a smart system that learns from text to answer user questions effectively. It highlights the importance of such systems as users increasingly seek direct answers. Various methods and types of question answering

systems are discussed to emphasize the need for accurate responses. Challenges like sticking to specific domains and extracting precise answers are identified, with proposed solutions including better language representation and formal verification methods. In conclusion, the paper urges further research to refine question answering systems, recognizing their significance in providing accurate information and enhancing user experiences. [1]

The paper presents an automated question answering (QA) approach to assist requirements engineers in extracting compliance-related information from regulations efficiently. Leveraging large-scale language models like BERT, the approach identifies relevant text passages within regulations. Empirical evaluation on 107 question-answer pairs from European regulations, including the GDPR, demonstrates high recall and accuracy rates. Future work includes extending the approach to handle cross-referencing within regulations and conducting user studies to assess practical usability. Overall, the proposed QA approach offers a streamlined alternative to manually parsing through lengthy regulatory documents, promising significant benefits for software compliance processes. [2]

The research paper discusses using the BERT algorithm for text summarization and question answering, providing a time-saving solution for users. Through NLP analysis, it shortens and refines text, creating a summary that is then processed by BERT for answering user queries. The system's potential applications include e-commerce, government, and customer service sectors. Future enhancements are suggested, such as handling larger text sizes, providing multiple answers, and improving summary precision. Overall, the approach shows promise in efficiently handling growing volumes of data and delivering instant insights. [3]

The abstract highlights the increasing use of the internet and the importance of question answering systems in various applications such as information retrieval and language learning. In conclusion, the QA system is envisioned as a valuable learning companion in education, capable of solving domain-specific problems and providing relevant answers. It aims to enhance the education system by enabling self-paced learning and evaluating answers akin to human performance. However, challenges such as knowledge representation, precise understanding, and evaluating complex answers remain, presenting opportunities for further exploration and development in the QA domain. Despite the complexity involved, the demand for such systems remains high, indicating significant potential for future advancements. [4]

This paper surveys recent work in stateless QA systems, focusing on RDF, Linked Data, and textual documents. It identifies main challenges, categorizes approaches, and reviews 21 systems and 23 datasets. The QA process, categorized into three main types, is discussed alongside commonly used methods. Evaluation datasets and tools are also addressed, highlighting trends and challenges in the field. [5]

This paper presents a novel approach for multilingual and KB-agnostic question answering (QA) over structured data. The approach translates natural language questions into SPARQL queries, enabling querying multiple KBs in different languages. Evaluation on five large KBs and five languages demonstrates its effectiveness. The approach is easily adaptable to new KBs and languages, ensuring portability and outperforming existing systems. Our contributions include qualitative and quantitative improvements in QA. [6]

This paper introduces an Amharic non-factoid QA system for biography, definition, and description questions, employing a hybrid approach for question classification and lexical patterns for document filtering and answer extraction. Evaluation demonstrates promising results, with minimal tool requirements. Future research aims to develop co-reference resolution tools, automate

lexical pattern generation, handle complex question types, and create a standard QA dataset. [7]

This paper surveys computational and NLP approaches to studying Hadith, a key resource in Islamic jurisprudence. It categorizes research into content-based, narration-based, and overall studies. It highlights renewed interest in Hadith among non-specialists and outlines future research directions, including sentiment and emotion mining. The survey aims to provide an overview for researchers interested in leveraging computational techniques for understanding Hadith. Identified challenges and opportunities include exploring diverse research types and addressing emerging research directions. [8]

Research in Open Domain Question Answering Systems (OD-QAS) emphasizes the need for strong external knowledge, often sourced from dynamic sources like the web. However, the web's unstructured data poses a challenge, requiring models for information extraction. This study proposes a pipeline architecture comprising preprocessing, information extraction, and text processing stages. Factoid questions serve as input, with the model generating snippets or sentences containing target answers. Three search engines, Yahoo!, Bing, and Ask, aid in retrieving relevant information, with slight variations in average precision and snippet yield observed across engines. Yahoo! demonstrates the highest snippet count, while Bing achieves the best average precision. [9]

PROPOSED METHODLOGY :-

The project implements an Automated Query Answering System using various Natural Language Processing (NLP) techniques.

3.1 Document Loading and Processing:

The system allows users to upload PDF documents through a web interface.Upon uploading, the documents are loaded using the load_documents() function.Which utilizes DirectoryLoader from Langchain to load PDF files from a specified directory.

3.2 Text Segmentation and Chunking:

After loading documents,the text is segmented into smaller chunks using the split_text_into_chunks() function.

This segmentation is crucial for processing large documents efficiently and enabling context-aware responses

3.3 Embeddings Generation:

The system generates embeddings for the text chunks using HuggingFace's Transformer-based models through the create_embeddings() function.These embeddings capture semantic representations of the text, facilitating similarity calculations between queries and document chunks.

3.4 Vector Store Creation: Using the generated embeddings, a vector store is created using the FAISS library through the create_vector_store() function.This vector store enables fast and efficient retrieval of document chunks based on their semantic similarity to user queries.

3.5 LLMS Model Initialization: An LLMS (Large Language Model for Search) model is initialized using the create_llms_model() function. This LLMS model, based on CTransformers, is configured with the help parameters suitable for conversational

3.6 Retrieval Chain Setup:

A Conversational Retrieval Chain is constructed using Langchain'sConversationalRetrievalChain class. Upon receiving a query, the system employs the Conversational Retrieval Chain to generate contextually relevant responses.

3.7 Chat History Management:

The system maintains a session state to store conversation history, past user inputs, and generated responses.This allows for seamless continuation of conversations and enhances the user experience.

4.Architectural Design:-

The architectural design for the Automated Query Answering System using NLP comprises distinct layers and components to ensure effective user interaction, document analysis, retrieval, and answer generation. At the presentation layer, users engage with the system through a web-based interface, where they input questions and upload documents via a browser. This layer, powered by Streamlit, facilitates seamless interaction and intuitive user experiences. The application layer acts as the system's core, orchestrating processes such as question analysis, document retrieval, answer generation, and NLP processing. Within this layer, modules manage the flow of data and logic, utilizing libraries like Hugging Face Transformers for NLP tasks. Document retrieval and analysis components access document repositories, employing techniques like vectorization and similarity search to identify pertinent documents based on user queries. PyPDF2 is integrated for loading document content, while answer generation modules synthesize responses based on the analyzed documents and user questions. Overall, this architectural design ensures cohesive functionality and efficient processing of user queries within the Automated Query Answering System.

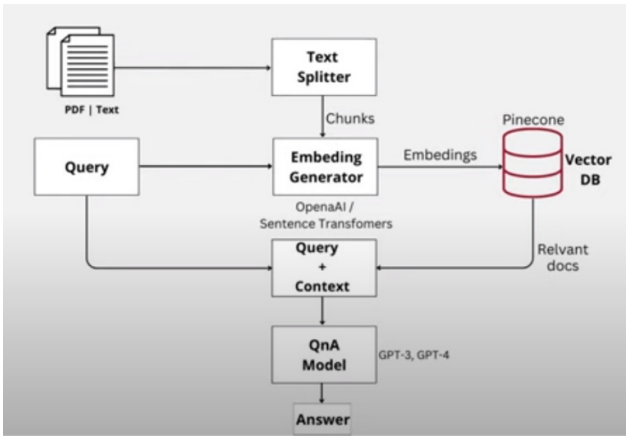


Fig 1: system diagram

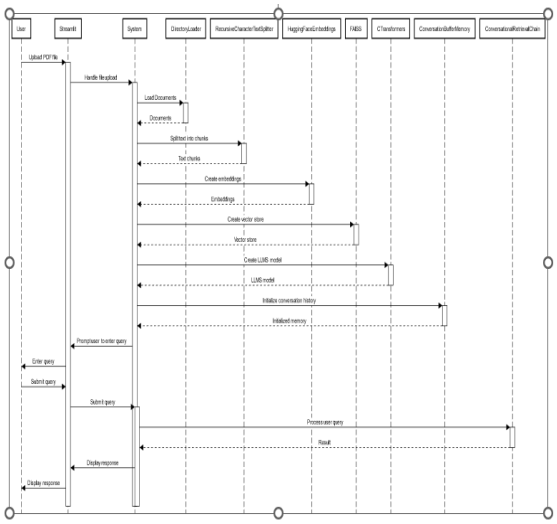


Fig 2:- work flow daigram

5.Result:

Fig-1:user interface

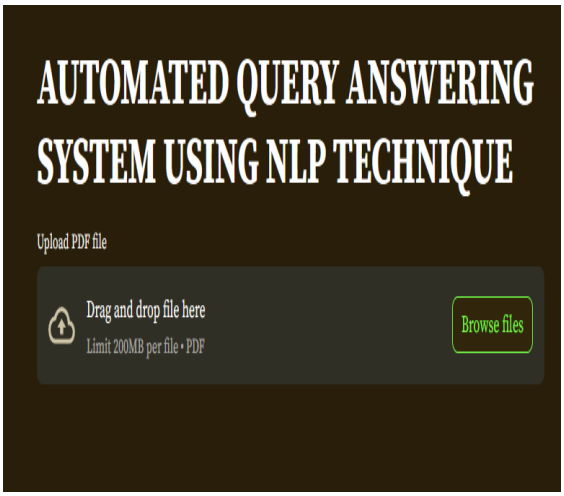


Fig-2:- upload the document

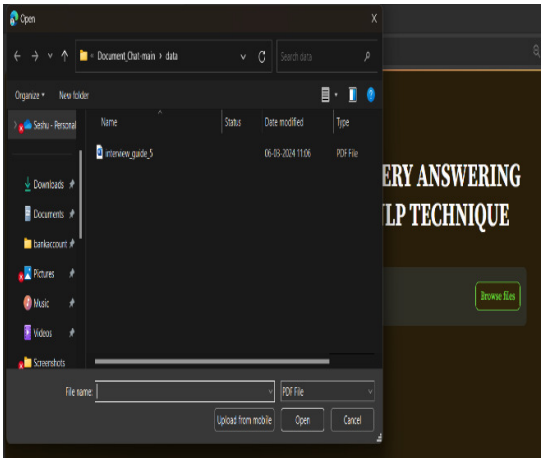


Fig-3:-processing the document

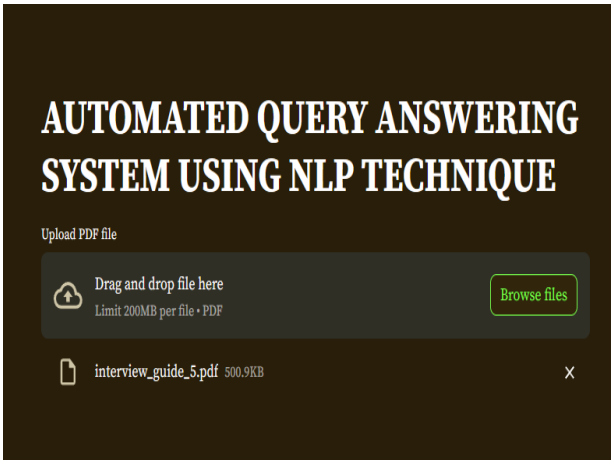


Fig-4:- user input query

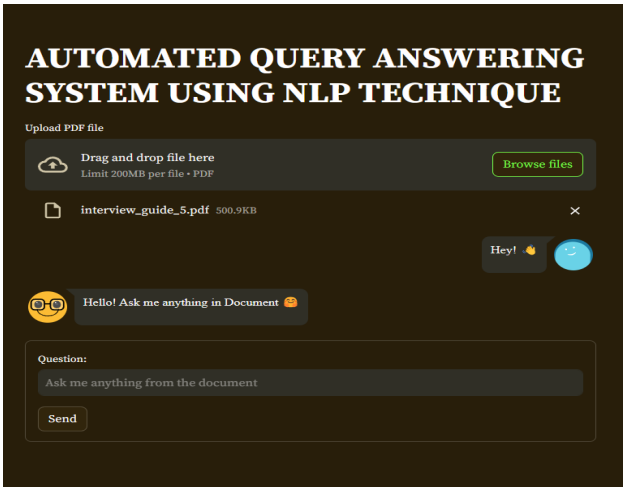


Fig-5:-Query processing

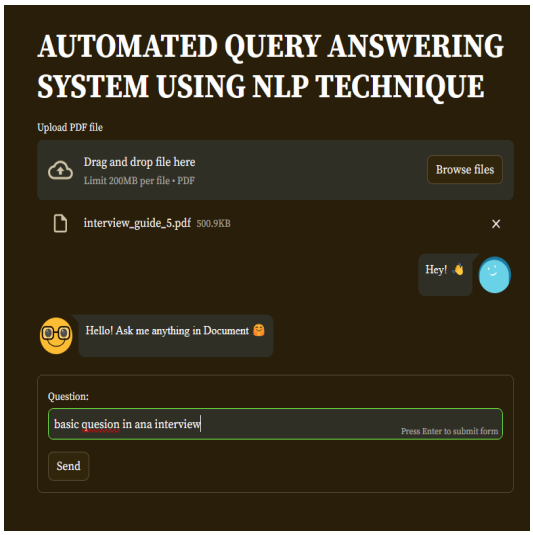
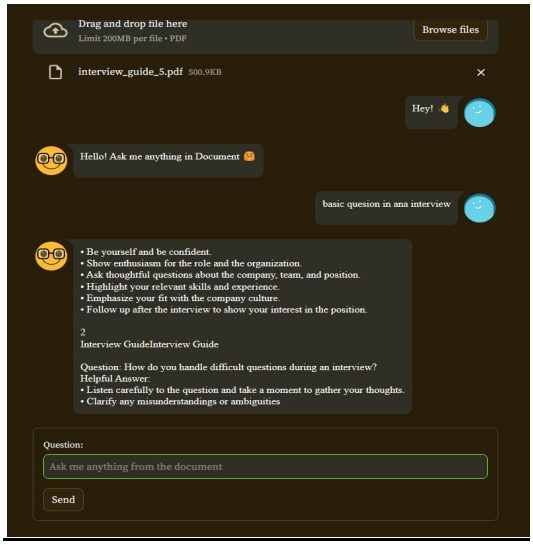


Fig-6:- answer generation



6.Acknowledgement:-

I extend my heartfelt gratitude to the trailblazers and visionaries in the fields of Natural Language Processing (NLP) and conversational AI, whose groundbreaking research laid the groundwork for automated query answering systems. Their pioneering work has reshaped our understanding of language understanding and paved the way for transformative applications in information retrieval.

I am deeply appreciative of the developers and contributors behind the open-source libraries and frameworks, including Streamlit, Langchain, HuggingFace, and FAISS. Their tireless efforts and commitment to democratizing access to advanced NLP techniques have empowered us to build this Automated Query Answering System.

Furthermore, I would like to acknowledge the support of academic and industrial institutions that have fostered an environment conducive to innovation and collaboration in the field of NLP. Their investments in research and technology have accelerated progress and enabled the development of cutting-edge solutions like ours.

I am indebted to my mentors, collaborators, and peers who have generously shared their expertise, provided invaluable feedback, and guided me throughout this project. Their insights and encouragement have been instrumental in shaping our approach and refining the capabilities of our system.

Lastly, I express my gratitude to the broader community of NLP enthusiasts, educators, and practitioners whose passion and dedication drive continuous exploration and discovery. Their contributions to the collective knowledge base inspire us to push the boundaries of what is possible in automated query answering and beyond.

7.Conclusion:-

the Automated Query Answering System using NLP represents a significant advancement in leveraging natural language processing techniques to enhance information retrieval and user interaction. Through the integration of Streamlit for intuitive web-based presentation and modules like Hugging Face Transformers for sophisticated NLP tasks, the system offers users a seamless experience for querying documents and receiving

accurate responses. The architectural design emphasizes modularity, scalability, and flexibility, allowing for easy integration of new features and updates to meet evolving requirements. By facilitating efficient processing of user queries, accurate NLP analysis, and generation of relevant responses, the system addresses the inherent challenges of document question answering systems. Moving forward, continued refinement and optimization of the system's components and algorithms will further enhance its capabilities and broaden its potential applications across various domains, ultimately advancing the state-of-the-art in automated information retrieval and knowledge extraction.

8.REFERENCES:-

[1]Shivani Singh, Nishtha Das, Rachel Michael, Poonam Tanwar "The Question Answering System Using NLP and AI" international Journal of Scientific & Engineering Research Volume 7, Issue 12, December-2016

[2]Sallam Abualhaija , Chetan Arora , Amin Sleimi , Lionel C. Briand "Automated Question Answering for Improved Understanding of Compliance Requirements: A Multi-Document Study ", School of Electrical Engineering and Computer Science, University of Ottawa, Canada,

[3] E. Dimitrakis, K. Sgontzos, Y. Tzitzikas. " Question Answering Systems over Linked Data and Documents." Journal of Intelligent Information Systems, vol. 55, no. 2, pp. 233-59, 2020. doi:10.1007/s10844-019-00584-7.

[4] Ajitkumar M. Pundge , Khillare S.A , C. Namrata Mahender , "Question Answering System, Approaches and Techniques", Dr. B.A.M.U, Aurangabad, Volume 141 – No.3, May 2016

[5] Anshul Kumar, Abhinav Panwar, Anurag, Dr. Mukesh Rawat, "Research Paper on Question Answering System using BERT", Meerut Institute of Engineering and Technology, Meerut, Uttar Pradesh, Volume 15 Issue 12 * December 2022

[6] D. Diefenbach, A. Both, K. Singh, Pierre Maret. "Towards a question answering system over the semantic web". Semantic Web, pp. 1-16, 2018. https://doi.org/10.3233/SW-190343.

[7] T. Abedissa, M. Libsie. "Amharic Question Answering for Biography, Definition, and Description Questions". Information and Communication Technology for Development for Africa, vol 1026. Springer, Cham, 2019. doi.org/10.1007/978-3-030-26630-1_26.

[8] A. Azmi, A. Omar, a. Hussain. "Computational and Natural Language Processing Based Studies of Hadith Literature: A Survey." Artificial Intelligence Review, vol. 52, no. 2, Springer, pp. 1369-414, 2019. https://doi.org/10.1007/s10462-019-09692-w.

[9] A. NS, A. UTAMI. "Information Extraction from Web as Knowledge Resources for Indonesian Question Answering System". In Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019) (pp. 419-425), May 2020. https://doi.org/10.2991/aisr.k.200424.064.

[10] G. Popek, W. Lorkiewicz. "Grounding of Modal Responses in Question Answering System Equipped with Hierarchical Categorisation." Procedia Computer Science, vol. 176, Elsevier B.V., pp. 3163-72, 2020. doi:10.1016/j.procs.2020.09.172

[11] Z. Huang et al., "Recent Trends in Deep Learning Based Open-Domain Textual Question Answering Systems," in IEEE Access, vol. 8, pp. 94341-94356, 2020. doi: 10.1109/ACCESS.2020.2988903.

[12] R. Bakis, D. Connors, P. Dube, P. Kapanipathi, R. Kumar, D. Malioutov, & C. Venkatramani. "Performance of natural language classifiers in a question-answering system" . IBM Journal of Research and Development, 61(4):14:1-14:10, 2017. https://doi.org/10.1147/JRD.2017.2711719.

[13] P. Baudiš, J. Šedivý. "Modeling of the question answering task in the yodaqasystem". In International Conference of the Cross-Language Evaluation Forum for European Languages. September 2015. https://doi.org/10.1007/978-3-319-24027-5_20.

[14 A. Bhandwaldar, W. Zadrozny. “UNCC QA: biomedical question answering system”. In Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering. November 2018. https://doi.org/10.18653/v1/W18-5308.

[15] L.Jovita, A.Hartawan, and D.Suhartono, “Using vector space model in question answering system,” Procedia Computer Science vol. 59, pp. 305 - 311, 2015. https://doi.org/10.1016/j.procs.2015.07.570.

[16] H. Xiao, X. Huang, J. Zhang, D. Li, P. Li. "Knowledge Graph Embedding Based Question Answering." WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining, no. Ccl, pp. 105-13, 2019. doi:10.1145/3289600.3290956.

[17] S. Hazrina, N. Sharef, H. Ibrahim, M. AzmiMurad, S. MohdNoah. “ Review on the advancements of disambiguation in semantic question answering system”. Information Processing & Management, 2017. https://doi.org/10.1016/j.ipm.2016.06.006.

[18] S. Hamed, M. Ab Aziz. “A Question Answering System on Holy Quran Translation Based on Question Expansion Technique and Neural Network Classification.” Journal of Computer Science, vol 12(3):169 177, January 2016. https://doi.org/10.3844/jcssp.2016.169.177.

[19] M. Espositoa, E. Damianoa, A. Minutoloa, G. De Pietroa, H. Fujitaal. "Hybrid Query Expansion Using Lexical Resources and Word Embeddings for Sentence Retrieval in Question Answering." Information Sciences, vol. 514, Elsevier, 2020, pp. 88-105. https://doi.org/10.1016/j.ins.2019.12.002.

[20] T. Dodiya, S. Jain. “Question classification for medical domain question answering system”. In 2016 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), December 2016. https://doi.org/10.1109/WIECON-ECE.2016.8009118.

