International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

VOICE ENABLED RETRIEVAL INTELLIGENCE SYSTEM USING NLP & RAG

VIJAYALAKSHMI J (ASP)

ASVITHAA K CHARULATHA V HARIKARAN K HARISH MUGUNTHAN M

BACHELOR OF TECHNOLOGY – 4th YEAR
DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

SRI SHAKTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

(AUTONOMOUS)
COIMBATORE - 641062

KEYWORDS

- Voice Recognition
- Natural Language Processing (NLP)
- Semantic Search
- Sentence Transformers
- FAISS
- Retrieval-Augmented Generation
- Intelligent Systems

ABSTRACT

Traditional information retrieval systems rely on text-based queries and keyword matching, making the process less intuitive, time-consuming, and often lacking contextual accuracy. To address these challenges, this project presents a Voice Enabled Retrieval Intelligence System that enables seamless interaction using voice commands combined with advanced Natural Language Processing techniques. The system captures user speech input and converts it into text using OpenAI Whisper, followed by semantic understanding through Sentence Transformers. By leveraging FAISS for efficient vector similarity search and embedding-based retrieval, it identifies the most relevant information from the knowledge base. The integration of Retrieval-Augmented Generation (RAG) enhances response quality by generating context-aware and accurate answers. The system provides both textual and voice-based outputs, enabling a more interactive and accessible user experience. Built using Streamlit for the interface, the system ensures real-time processing, scalability, and improved search efficiency, making it suitable for intelligent assistants and knowledge retrieval applications.

1. INTRODUCTION

In the modern digital era, the demand for efficient and intelligent information retrieval systems has increased significantly due to the rapid growth of data across various domains. Traditional information retrieval methods primarily rely on manual text input and keyword-based search techniques, which often make the process less intuitive, time-consuming, and limited in understanding the true intent of user queries. These systems frequently fail to capture contextual meaning, resulting in less accurate and sometimes irrelevant results. As a result, there is a growing need for more advanced, user-friendly, and context-aware retrieval mechanisms.

To address these limitations, the Voice Enabled Retrieval Intelligence System is proposed as a solution that enhances human-computer interaction through voice-based communication combined with Natural Language Processing (NLP) techniques. The system allows users to interact using voice commands, reducing dependency on manual text input and improving accessibility. It converts speech into text and processes it to understand user intent effectively.

The system utilizes Sentence Transformers to generate semantic embeddings, enabling better contextual understanding compared to traditional keyword-based approaches. For efficient retrieval, FAISS (Facebook AI Similarity Search) is used to perform fast vector similarity search and identify the most relevant information from the knowledge base.

Additionally, the integration of Retrieval-Augmented Generation (RAG) improves response quality by combining retrieval with generation, producing context-aware and accurate answers. The system supports both text and voice-

based outputs, providing an interactive and user-friendly experience suitable for applications such as intelligent assistants and knowledge retrieval systems.

Furthermore, this system emphasizes real-time performance and scalability, ensuring efficient handling of large volumes of data and user queries. By combining voice interaction, semantic search, and intelligent response generation, it significantly enhances the overall efficiency, accuracy, and user experience of modern information retrieval systems. Moreover, the system is designed with flexibility and adaptability in mind, allowing it to be extended across various domains such as education, healthcare, and customer support. Its ability to understand natural language queries and deliver precise, context-aware responses makes it a valuable tool for improving decision-making and knowledge accessibility in real-world applications.

2. LITERATURE REVIEW

Recent advancements in artificial intelligence have significantly improved intelligent information retrieval systems, particularly in the areas of voice-based interaction and natural language understanding. AI-driven approaches have transformed traditional systems by enabling more intuitive and efficient human-computer interaction. These systems utilize advanced machine learning and NLP techniques to process user queries more effectively, making information access faster and more user-friendly. As a result, AI-based retrieval models provide a strong foundation for the development of modern voice-enabled retrieval intelligence systems.

Traditional retrieval systems rely on keyword-based search, which often fails to capture user intent and contextual meaning. Studies highlight the need for semantic search techniques that improve accuracy and efficiency by understanding the context of user queries. Several researchers have proposed AI-based frameworks that combine semantic analysis with efficient retrieval mechanisms. Techniques such as semantic embeddings and vector similarity search enable systems to retrieve relevant information more effectively and accurately.

The integration of retrieval with generative models, such as Retrieval-Augmented Generation (RAG), further enhances system performance by generating context-aware responses. These approaches improve both the quality and relevance of retrieved information.

Additionally, optimization techniques and improved model architectures help achieve faster processing and scalability in real-time applications. Structured datasets and conversational AI systems also play a key role in improving system performance, usability, and user interaction. Overall, existing research emphasizes the transition from traditional keyword-based systems to intelligent, context-aware retrieval models, highlighting the importance of AI techniques in developing efficient and user-friendly voice-enabled systems.

3. DATA ACQUISITION

Data acquisition is an essential step in the development of the Voice Enabled Retrieval Intelligence System, as it involves collecting relevant data required for efficient information retrieval. The system gathers input from multiple sources, including user interactions and knowledge bases, to ensure accurate and meaningful responses. Proper data acquisition enables the system to handle diverse queries and improve overall performance.

3.1. User Voice Input Data User interactions through voice commands form the primary source of data for the system. These inputs include queries related to information retrieval, general questions, and knowledge-based requests. The system captures real-time voice input and converts it into text using speech recognition techniques, enabling further processing and semantic analysis.

3.2. Text Query Data In addition to voice input, the system also accepts text-based queries entered by users. These queries serve as an alternative input method and help in improving accessibility. The collected text data is used to understand user intent and ensure accurate retrieval of relevant information.

3.3. Knowledge Base Data The system relies on a structured knowledge base consisting of documents, text files, and domain-specific data. This data acts as the primary source for retrieving information and generating responses. A well-organized knowledge base ensures that the system can provide accurate, relevant, and context-aware outputs.

3.4. External Data Sources The system can integrate external sources such as online datasets, APIs, or publicly available information repositories. These sources enhance the system's knowledge by providing up-to-date and diverse information, improving the overall quality and reliability of responses.

3.5. Interaction And Usage Data User interaction data, including previous queries and system responses, is collected to analyze usage patterns and improve system performance. This data helps in understanding common user needs, refining retrieval accuracy, and enhancing the overall user experience over time.

4. DATA PREPROCESSING

Data preprocessing is a crucial step in the system pipeline, as it prepares raw input data for further processing and analysis. It ensures that the collected data is clean, structured, and suitable for semantic understanding and retrieval. Proper preprocessing improves system accuracy, efficiency, and overall performance in handling user queries.

4.1. Speech-to-Text Conversion

Following data acquisition, the system performs speech-to-text conversion to transform user voice input into textual format. This is achieved using advanced models such as OpenAI Whisper, which can handle different accents, speech speeds, and noisy environments. The converted text represents the user's query in a machine-readable form, enabling further processing through NLP techniques. High accuracy at this stage is essential, as errors in transcription can affect subsequent processing.

4.2. Text Cleaning and Normalization

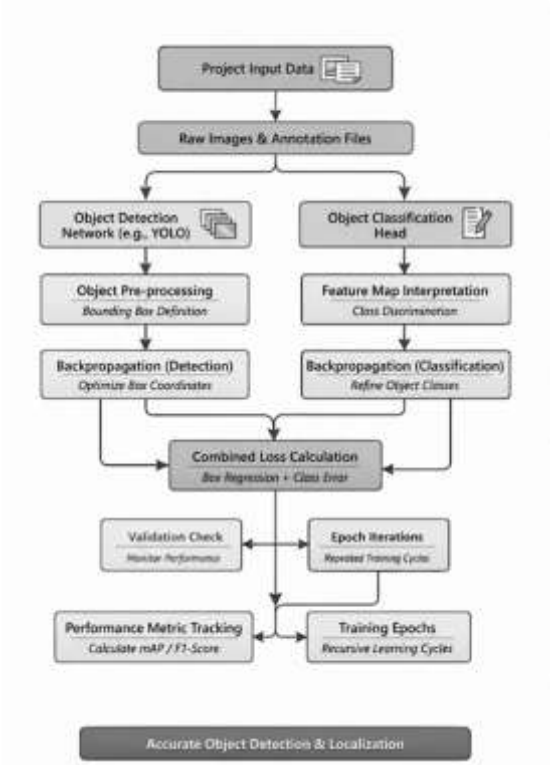
The converted text is cleaned and normalized to remove noise, unnecessary symbols, and inconsistencies. This includes processes such as lowercasing, removing special characters, and correcting formatting issues. These steps ensure that the text is standardized and ready for semantic analysis.

4.3. Tokenization and Processing

In this stage, the cleaned text is broken down into smaller units such as words or tokens. Tokenization helps the system understand the structure and meaning of the query. This processed data is then prepared for embedding generation and further semantic interpretation.

4.4. Embedding Generation and Processing Pipeline

The system uses Sentence Transformers to convert processed text into semantic embeddings that capture the contextual meaning of the query. These embeddings are then used for similarity search using FAISS, enabling efficient retrieval of relevant information. The integration of Retrieval-Augmented Generation (RAG) further enhances the pipeline by generating context-aware responses. This complete AI-driven framework ensures accurate, scalable, and real-time information retrieval.



5.REAL TIME ANALYSIS AND MONITORING

Real-time analysis and monitoring play a crucial role in ensuring the efficiency, accuracy, and responsiveness of the Voice Enabled Retrieval Intelligence System. This component enables the system to process user queries

instantly while continuously tracking system performance and output quality. By maintaining real-time monitoring, the system ensures reliable operation, faster response delivery, and an improved interactive user experience.

5.1. Real-Time Query Processing

The system is designed to handle and process user queries in real time, whether they are received as voice or text input. Once the input is captured and preprocessed, it is immediately passed through the NLP pipeline for semantic understanding and retrieval. This rapid processing minimizes delays and ensures that users receive accurate and relevant responses almost instantly, enhancing the overall system efficiency.

5.2. Semantic Analysis and Retrieval Monitoring

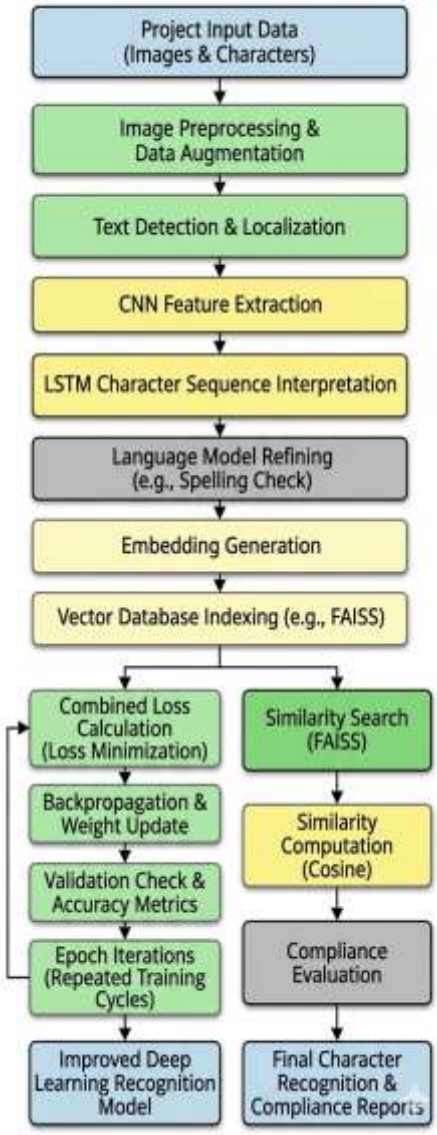
The system continuously monitors the performance of semantic analysis and retrieval processes to ensure accuracy and relevance. By using Sentence Transformers for embedding generation and FAISS for similarity search, it evaluates how effectively user queries are matched with the stored knowledge base. This monitoring helps in identifying mismatches or irrelevant results and allows for improvements in retrieval precision.

5.3. Response Generation and Feedback Handling

The Retrieval-Augmented Generation (RAG) model generates meaningful and context-aware responses in real time. The system also incorporates mechanisms to monitor response quality and collect user feedback when available. This feedback is used to refine the model, improve response accuracy, and enhance the system's ability to handle complex queries effectively.

5.4. System Performance and Efficiency Tracking

The system tracks important performance metrics such as response time, processing speed, and retrieval accuracy. Continuous monitoring of these parameters helps in identifying performance bottlenecks and optimizing system components. This ensures that the system remains scalable, reliable, and capable of delivering consistent performance in real-time environments.



6. INTELLIGENT RESPONSE GENERATION AND CONTROL

This component focuses on generating accurate, context-aware responses based on user queries and controlling the overall response flow of the system. It integrates advanced Natural Language Processing techniques with Retrieval-Augmented Generation (RAG) to ensure meaningful and relevant outputs. The system intelligently combines retrieved information with generative models to enhance response quality and user interaction.

6.1. Query Understanding and Interpretation

The system first interprets the user query by analyzing its semantic meaning using Sentence Transformers. Instead of relying on keyword matching, it understands the context and intent behind the query. This enables the system to handle complex and natural language inputs more effectively, ensuring accurate information retrieval.

6.2. Retrieval of Relevant Information

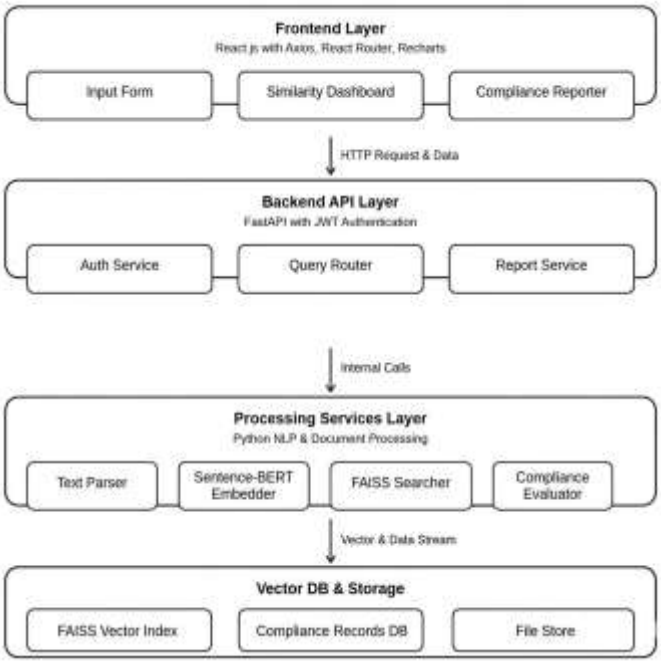
After understanding the query, the system retrieves the most relevant information from the knowledge base using FAISS-based similarity search. The use of vector embeddings allows the system to identify contextually similar data rather than exact keyword matches. This improves the accuracy and relevance of retrieved results.

6.3. Response Generation using RAG

The Retrieval-Augmented Generation (RAG) model combines retrieved data with generative AI techniques to produce meaningful and context-aware responses. This approach ensures that the output is not only accurate but also well-structured and easy to understand. It enhances the system's ability to provide intelligent and human-like responses.

6.4. Output Control and Delivery

The final response is delivered to the user in both text and voice formats. The system ensures that the output is clear, concise, and relevant to the query. Text-to-speech technology is used to convert responses into voice, enabling a more interactive and accessible user experience.



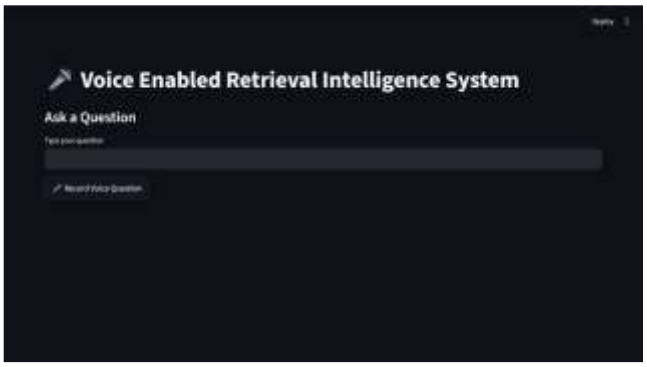
7.SYSTEM OPTIMIZATION AND PERFORMANCE ENHANCEMENT

7.1. Performance Optimization

The system achieves performance optimization by enhancing the efficiency of each stage in the processing pipeline, including speech recognition, text processing, embedding generation, and similarity search. Advanced models such as OpenAI Whisper are optimized to accurately convert speech to text even in noisy environments, while Sentence Transformers generate high-quality semantic embeddings with minimal processing time. The use of FAISS for vector similarity search significantly reduces retrieval time by enabling fast and efficient indexing of large datasets. Additionally, the system minimizes computational overhead through optimized model execution and efficient memory management techniques. These optimizations ensure that the system can process user queries in real time with reduced latency, improved accuracy, and consistent performance, even when handling multiple requests or large-scale data.

7.2. Continuous Improvement and System Enhancement

The system incorporates continuous improvement strategies to enhance its accuracy, adaptability, and overall performance over time. By analyzing user interactions, query patterns, and system responses, it identifies areas where improvements are required. Feedback mechanisms play a key role in refining the system, allowing it to learn from incorrect or less relevant responses and adjust accordingly. Regular updates to the knowledge base ensure that the information provided remains accurate and up to date. Furthermore, iterative testing and performance evaluation help in fine-tuning model parameters and improving the effectiveness of retrieval and response generation processes. These enhancements enable the system to evolve continuously, maintain high reliability, and deliver more precise and context-aware responses, making it suitable for dynamic and real-world applications.



8.REFERENCES:

[1] Albert, L. A. (2025). Artificial intelligence in systems: Integrating AI into the engineering curriculum. Available at SSRN 5240570.

[2] Ljubić, H., Tomić, Z., & Volarić, T. (2025). From vision to structure: Simplifying syllabus development for educators. In EDULEARN25 Proceedings (pp. 1576–1581). IATED.

[3] G. Adorni, D. Grosso, & D. Ponzini (2024). Building conversational AI systems: A framework for integrating AI into real-world applications. In ICERI2024 Proceedings (pp. 9216–9226). IATED.

[4] S. N. A. Mohamed Mahtar (2015). Analyzing artificial intelligence systems and evaluation methods using structured approaches (Doctoral dissertation, Universiti Teknologi MARA).

[5] Mittal, P. Parthasarathy, & S. Joshi (2025). AI systems and ethical considerations in intelligent applications. In Annual ACM India Compute Conference (pp. 70–85). Springer Nature Switzerland.

[6] K. Dutta, S. Paul, & A. Anand (2024). AI-driven architectures for intelligent and context-aware systems. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 5(1), 180–185.

[7] B. Schneider. Understanding behavioral impact of AI systems and user interaction models in intelligent environments.

[8] P. Chomphooyod, L. Jeerapradit, A. Suchato, & P. Punyabukkana (2025). Multi-agent AI systems for automated content generation and retrieval using modern frameworks. In IEEE iSTEM-Ed Conference (pp. 1–6).

[9] D. Miedema et al. (2025). Data systems and intelligent retrieval: Curriculum recommendations and industry needs. In ITiCSE Working Group Reports (pp. 95–123).

[10] Sadovnychenko, N. Pastukhova, & V. Miasoiedov (2025). Ensuring compliance and integrity in digital and AI-driven systems. Baltija Publishing.