Healthcare Mining: An AI-Powered Approach to Revolutionizing Information Retrieval in Healthcare

Sujith Ramprasad Tellakula, Ronit Patil, Vishnu Batla, Simran Panchal, Jay Mistry
Arizona State University
Tempe, AZ, USA, 85281
{stellak1, rpatil43, vbatla, spanch12, jmistry3}@asu.edu

***Abstract*—The "Healthcare Mining" project leverages artificial intelligence (AI) and machine learning (ML) to revolutionize healthcare information retrieval, addressing the challenge of extracting precise and relevant health information from vast digital data sources. By employing advanced natural language processing (NLP) techniques and a unique algorithmic framework, our system surpasses traditional search methods, offering users tailored, accurate, and timely health insights. The integration of OpenAI's Embeddings and vector similarity search techniques enables a semantic understanding of user queries, ensuring the delivery of contextually appropriate results. Our approach not only enhances the accuracy and efficiency of health information retrieval but also adapts to the evolving nature of healthcare data, promising a significant improvement in public health outcomes through informed decision-making. This project represents a pivotal step towards accessible, reliable medical knowledge, empowering individuals with the information needed to make better health-related decisions.**

***Index Terms*—Artificial Intelligence, Machine Learning, Healthcare Information Retrieval, Natural Language Processing, Semantic Search, Vector Similarity, Data Mining, Web Scrapping, Retrieval-Augmented Generation, Langchain Agents**

I. Introduction

In an era where digital information grows exponentially, the healthcare sector faces a paradox: while there is more health-related information available than ever before, finding accurate, relevant, and trustworthy data has become increasingly challenging. Traditional search engines and medical databases often yield results that are either too broad or misaligned with the user's specific needs, leading to confusion and potential misinformation. This issue is compounded by the diverse nature of medical terminology and the layperson's understanding, creating a gap between the information sought and the information retrieved.

The "Healthcare Mining" project emerges as a solution to these challenges, utilizing the latest advancements in artificial intelligence (AI) and machine learning (ML) to revolutionize how healthcare information is accessed and understood. By leveraging sophisticated natural language processing (NLP) techniques, our system aims to decode the complexities of human language, enabling a more intuitive and effective search experience that aligns closely with the user's intent and contextual needs.

Moreover, the project addresses the critical issue of information reliability and timeliness, which are paramount in healthcare decision-making. With an innovative approach to data mining and analysis, "Healthcare Mining" seeks to filter and prioritize information not just by relevance, but also by credibility and recency, ensuring that users have access to the most accurate and up-to-date health information.

This introduction not only sets the project within the broader context of digital healthcare challenges but also highlights the transformative potential of AI and ML in bridging the gap between vast data resources and the specific information needs of users. As we move forward, "Healthcare Mining" stands as a testament to the power of technology to enhance the accessibility and reliability of healthcare information, promising a future where informed health decisions are within everyone's reach.

II. Problem Definition

In the vast landscape of digital healthcare information, individuals face significant challenges in locating precise, relevant, and credible information. Traditional search methods often yield results that are either too broad, outdated, or not sufficiently reliable for making informed health decisions. The complexity of medical terminology further exacerbates this issue, as does the dynamic nature of medical knowledge, where new insights and guidelines are constantly emerging. "Healthcare Mining" addresses these critical gaps by leveraging artificial intelligence (AI) and machine learning (ML) technologies. Our aim is to create a system that not only understands the nuanced queries of users but also ensures the relevance, credibility, and timeliness of the information provided. This middle ground approach seeks to revolutionize how healthcare information is retrieved, making it accessible, accurate, and actionable for all users.

III. Data Sets

Our project leverages a comprehensive dataset compiled from various healthcare forums. The dataset, structured in JSON format for ease of processing, covers an extensive array of healthcare topics including diseases, symptoms, treatments, and medications. Key sources include:

- WebMD
- Patient.info
- MedlinePlus

Each entry in the dataset is enriched with metadata. This structured approach enables our "Healthcare Mining" system

to deliver precise, relevant, and up-to-date healthcare information.

IV. State-of-the-Art Methods & Algorithms

The landscape of healthcare information retrieval is rapidly evolving, with several advanced methods and algorithms setting the current standards. These include techniques for efficiently mining and interpreting vast amounts of unstructured data from medical literature, patient forums, and healthcare discussions.

Natural Language Processing (NLP) has become pivotal, with tools like MetaMap translating biomedical text into structured data, aligning with the Unified Medical Language System (UMLS) to enhance data interoperability and semantic understanding.

Graph Theory applications, such as symptom relation graphs (SympGraph), analyze co-occurrence and relationships between medical terms, offering insights into symptom-disease correlations and patient experiences.

Vector Space Models and **Embedding Techniques** transform textual information into vector representations, facilitating the application of machine learning algorithms for pattern recognition, trend analysis, and predictive modeling in healthcare contexts.

Retrieval and Ranking Algorithms have been refined to prioritize the relevance and credibility of information sources, incorporating user feedback and engagement metrics to adapt search results to user needs dynamically.

These methodologies underscore a shift towards more personalized, accurate, and context-aware information retrieval systems in healthcare, promising to significantly enhance patient knowledge and support clinical decision-making.

This section emphasizes the integration of diverse computational techniques to address the complexities of healthcare data, reflecting the interdisciplinary nature of current research efforts aimed at improving access to and the quality of healthcare information.

V. Research and Development Plan

Our comprehensive R&D plan is structured to ensure the development of a robust healthcare information retrieval system, leveraging the latest in AI and machine learning technologies.

A. Data Collection and Preprocessing

Our data collection process leverages Beautiful Soup, Scrapy, AsyncIO, and Playwright for web scraping, to extract detailed healthcare information from various online sources. This phase involves:

- **Identifying Data Sources:** Selecting authoritative and credible healthcare websites and forums for data extraction.
- **Web Scraping:** Utilizing the above mentioned tools to navigate the DOM of web pages, extract relevant data, and handle the intricacies of HTML and XML parsing efficiently.

B. System Design and Development

Our system architecture is streamlined, focusing on the frontend application developed with Streamlit, which directly interfaces with OpenAI's API.

- **Streamlit as the UI Framework:** Streamlit facilitates the creation of an interactive user interface, allowing users to input queries, upload files, and receive information in real-time.
- **Direct API Calls to OpenAI:** The query embedding generation, is performed through direct calls to OpenAI's API. This approach ensures efficient handling of user requests and leverages OpenAI's powerful natural language processing capabilities.

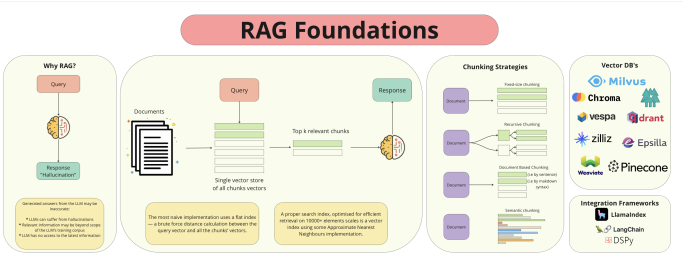

Fig. 1. Retrieval Augmented Generation - Workflow

C. Implementation of Algorithms

Our Retrieval Agent, developed using Langchain, is designed for advanced query processing and information retrieval. It employs a vector similarity search mechanism, crucial for matching user queries with the most relevant documents in our vector database. Figure 1 shows the entire workflow in the process of Retrieval Augmented Generation.

- **Vector Similarity Search:** We utilize cosine similarity to compare the user's input query embeddings with document embeddings stored in the vector database, efficiently identifying the top 5 documents that closely match the query. Cosine Similarity is defined as the cosine of the angle between these two vectors. The mathematical representation is given by the following equation:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{1}$$

- **Retrieval Agent Capabilities:** These top matches are then processed by the OpenAI API, which generates a coherent and contextually relevant response to the user's query, showcasing the agent's sophisticated ability to interpret and respond to user inputs accurately.

D. User Interface Development

The user interface, crafted with Streamlit, prioritizes simplicity and efficiency, facilitating seamless interaction for users. This interface is the gateway for users to access the advanced retrieval capabilities of our system. The basic UI for the application can be seen in the figures below. Figure

2 shows an empty chat and Figure 3 shows a to and fro conversation between a user and our system.

- **UI/UX Design Principles:** The design follows modern UI/UX principles, focusing on minimalism to reduce cognitive load and ensure the interface is intuitive for users of all backgrounds. Attention to color scheme, typography, and layout enhances accessibility and user engagement.
- **Functionalities for User Interaction:** Key features include a chat interface for query input, allowing for natural language interaction. Users can upload documents or enter URLs for direct data analysis, which the system processes in real-time. This immediate feedback loop encourages exploration and interaction, making complex information retrieval tasks straightforward for users.

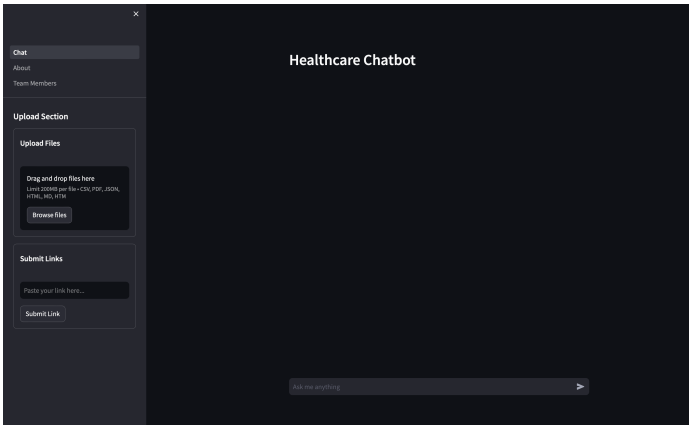

Fig. 2. Basic User Interface

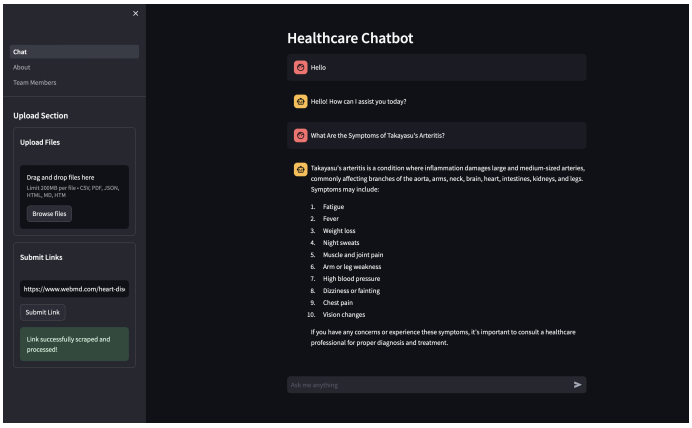

Fig. 3. Sample Chat

VI. Design of Experiments and Evaluation Plan

Our evaluation strategy is centered around rigorously assessing the technical performance of the "Healthcare Mining" system, particularly through the integration of Trulens for in-depth analysis. This plan is structured to ensure that the system effectively meets its goal of providing accurate and relevant healthcare information retrieval.

A. Technical Performance Metrics

We utilize a set of quantitative metrics to evaluate the system's performance, focusing on:

- **Trulens Analysis:** The core of our evaluation, Trulens, is used to assess Groundedness, Questions/Answer Relevance, and Question/Context Relevance. This analysis allows us to gauge how well the system's responses are rooted in the factual content of our database and their relevance to the user's queries. Figure 4 shows the feedback functions we are incorporating in this project to obtain a qualitative analysis of our Retrieval agent's responses.

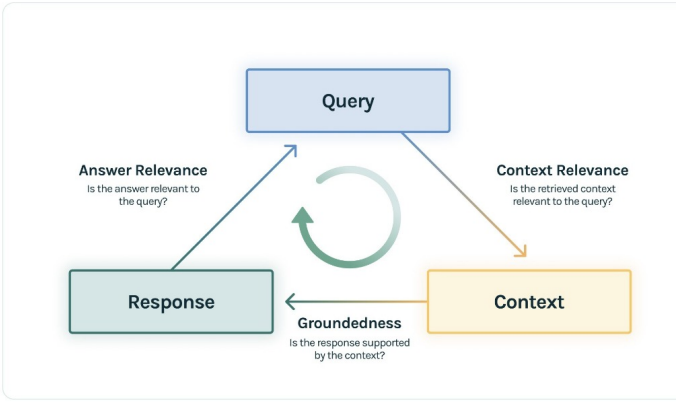

Fig. 4. TruLens Feedback Functions

B. Iterative Refinement Process

Feedback obtained from the technical performance evaluation will guide the iterative refinement of our system. This process involves:

- Adjusting and optimizing the algorithmic approach based on Trulens metrics to enhance the accuracy and relevance of information retrieval.
- Continuously updating the database and retrieval mechanisms to incorporate the latest healthcare information and user feedback.

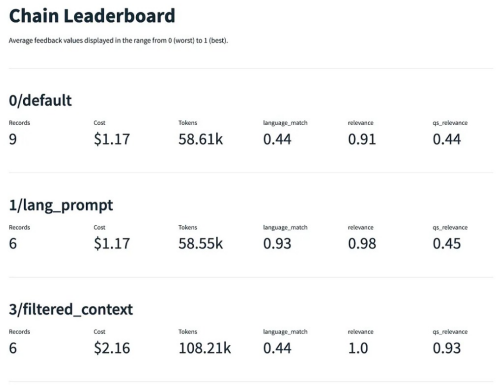

Fig. 5. Comparison between different Retrieval Chains

Figure 5 shows the comparison between some of the retrieval chains that we have tried to implement. This focused evaluation plan is designed to ensure that our "Healthcare Mining" system achieves high standards of accuracy and relevance in healthcare information retrieval, leveraging the insights gained from Trulens analysis for ongoing system improvements.

VII. Project Timeline: Tasks, Descriptions, and Deadlines

Our project is meticulously planned to ensure timely completion of each phase, from initial research to final deployment. Below is an overview of the primary tasks, along with their descriptions and deadlines.

A. Initial Research and Data Collection

Task Description: Identify and gather healthcare data from various sources, ensuring a comprehensive dataset for initial development. **Deadline:** Feburary 2nd, 2024

B. System Design and Prototype Development

Task Description: Design the system architecture and develop a prototype, integrating the OpenAI API for embeddings and setting up the initial UI with Streamlit. **Deadline:** Feburary 16th, 2024

C. Algorithm Implementation and Testing

Task Description: Implement the System, followed by rigorous testing to evaluate performance. **Deadline:** February 23rd, 2024

D. System Evaluation and Refinement

Task Description: Conduct a comprehensive evaluation using Trulens and other metrics, refining the system based on findings. **Deadline:** March 5th, 2024

E. Final Deployment and Project Presentation

Task Description: Deploy the finalized version of the system and prepare for the project presentation. **Deadline:** March 15th, 2024

This timeline is designed to ensure a structured approach to the project, with clear milestones and deadlines for accountability and progress tracking.

VIII. Division of Work

The successful execution of our project, "Healthcare Mining," requires a collaborative effort where tasks are clearly allocated to team members. Below is the division of work among our team:

- **Sujith Ramprasad Tellakula** - Contributed to the project by developing the user interface with Streamlit, facilitating an engaging chat UI. Implemented the conversational retrieval Q&A chain and retrieval agent in langchain, essential for query processing. Also developed chat history functionality to maintain conversational context and integrated streaming of OpenAI's API responses to enhance UI interactivity.
- **Ronit Patil** - Helped build the basic RAG chatbot. Integrated API usage for tokens and total cost per API call, added features such as dynamic file uploading and link scraping. Added two new agents for effective retreival from vectorDB. Implemented TruLens as an evaluation metric to validate responses by chatbot by coding specific feedback functions.
- **Vishnu Batla** - Played a pivotal role by conducting thorough research on X-Ray Models, contributing key insights, and drafting the overall conceptual framework. Recommended suitable X-Ray models with readily available APIs, to facilitate easy integration in the later stages. Also acquired knowledge about Large Language Models throughout the process and explored ways to implement them.
- **Simran Panchal** - Utilized Python libraries like BeautifulSoup and Scrapy to extract data from websites through web scraping techniques. Sent an HTTP request to the target website to initiate the retrieval of web data using Python. Parsed HTML and XML documents efficiently using BeautifulSoup, enabling seamless navigation and extraction of desired information.
- **Jay Mistry** - Used the asyncio library for asynchronous programming and Playwright for browser automation to scrape disease information from WebMD. By parsing HTML content with BeautifulSoup, it extracts data about disease names, symptoms, prevention methods, and causes, saving the collected information into JSON files. The script handles exceptions during scraping and ensures proper closure of the browser after each operation, offering an efficient and robust solution for gathering health-related data from the web.

This structured approach ensures that all aspects of the project are managed efficiently, with clear responsibilities and deadlines.

References

[1] Langchain documentation. https://python.langchain.com/docs/get_started/introduction.

[2] Openai api documentation. https://platform.openai.com/docs/introduction.

[3] Streamlit documentation. https://docs.streamlit.io.

[4] Trulens evaluation with langchain quickstart. https://www.trulens.org/trulens_eval/langchain_quickstart/.

[5] Anand V. Saurkar, Kedar G. Pathare, and Shweta A. Gode. An overview on web scraping techniques and tools. 2018.

[6] Parikshit Sondhi, Jimeng Sun, Hanghang Tong, and Chengxiang Zhai. Sympgraph: A framework for mining clinical notes through symptom relation graphs. In *KDD'12 - 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1167–1175, September 2012. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012 ; Conference date: 12-08-2012 Through 16-08-2012.

[7] Illhoi Yoo, Jinbo Bi, and Xiaohua Hu, editors. *2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019, San Diego, CA, USA, November 18-21, 2019*. IEEE, 2019.