INTELLIGENT NATURAL LANGUAGE SEARCH - KNOWLEDGE BUILDER

Ms. Sandhya L
Assistant professor
School of CSE&IS
PRESIDENCY UNIVERSITY
Bangalore, INDIA

Mr. PRATHAM V BEGALE
PRESIDENCY UNIVERSITY
Bangalore, INDIA
PRATHAM.20201CST0085@
presidencyuniversity.in

Mr. C JAI MANOJ REDDY
PRESIDENCY UNIVERSITY
Bangalore, INDIA
CHIGULLAREVU.20201CST0102@
presidencyuniversity.in

Mr. VEMULA VENKATA SURESH
PRESIDENC UNIVERSITY
Bangalore, INDIA
VEMULA.20201CST0170
@presidencyuniversity.in

Mr. AHMED ULLA KHAN
PRESIDENCY UNIVERSITY
Bangalore, INDIA
AHMED.20201cst0139@
presidencyuniversity.in

***Abstract*— In response to the dynamic landscape of information management within research and development (R&D) spheres, Dr. Reddy's Laboratories is undertaking an ambitious initiative to streamline the extraction of insights from their extensive repository of R&D data. This project centers around the creation of an intelligent natural language search tool designed to navigate and derive actionable knowledge from the wealth of information encapsulated within PowerPoint (PPT) slides stored in a shared folder.**

The primary challenge addressed by this endeavor is the efficient extraction of meaningful insights and information from diverse and voluminous PPT presentations. Leveraging cutting-edge technologies and methodologies, the proposed solution takes the form of a web application, orchestrating seamless interaction between users and the repository of R&D reports, findings, and presentations.

The envisioned software, termed as the "Intelligent Knowledge Builder," stands as a culmination of advanced development methodologies, integrating Streamlit , Spacy, Transformers, and other libraries. Through its user-friendly interface, the web application presents an intuitive platform where users can input search queries or terms of interest, prompting the system to embark on an automated journey of knowledge extraction.

By harnessing natural language processing (NLP) techniques and machine learning algorithms, the system traverses through the PPT slides, extracting pertinent information and generating concise yet informative insights. This intelligent mechanism extends beyond mere keyword matching, encompassing comprehensive comprehension and synthesis of diverse textual content to offer targeted responses aligned with user queries.

Furthermore, the software's adaptive functionality extends to user-uploaded files, enabling tailored analysis of additional PPT or PDF documents. The system applies a robust text extraction and analysis process, culminating in the generation of customized insights, empowering users with deeper knowledge and enriched perspectives.

This software not only serves as a revolutionary tool for automatic knowledge extraction but also embodies Dr.Reddy's Laboratories' commitment to technological innovation in enhancing R&D efficiency. The successful implementation of this web application stands to redefine the paradigm of information retrieval, enabling researchers and stakeholders to harness invaluable insights swiftly and effectively from the organization's wealth of R&D resources.

I . INTRODUCTION

The current project focuses on developing a robust knowledge retrieval system designed specifically for Dr. Reddy's Laboratories. The objective is to create an intuitive and effective platform that utilizes advanced technologies to extract and present valuable insights from a diverse range of research and development (R&D) resources.

At the core of this endeavor is the creation of a software application, named the "Knowledge Builder," utilizing technologies such as Streamlit , Spacy, Transformers, and other relevant libraries. This application acts as an interface for users to interact with the extensive data repository, primarily composed of PowerPoint (PPT) presentations, web content, and various report formats.

The application's functionalities are organized into two main sections: "Home" and "Use my file." The "Home" section provides access to available source files, primarily focusing on R&D reports, through an expandable interface. Users can submit queries related to the available resources using an input field.

Upon query submission, the application initiates a thorough search mechanism, incorporating techniques such as text extraction from PPTs, web scraping, and Natural Language Processing (NLP) using libraries like Spacy. The search results are presented as responses, utilizing the Transformers library to generate concise yet informative insights.

The "Use my file" section allows users to upload their PPT or PDF files for analysis. The system undergoes a similar process of text extraction and analysis, contributing to the generation of tailored responses based on the user's query.

This project integrates cutting-edge technologies and methodologies with the goal of streamlining knowledge acquisition and insight generation from Dr. Reddy's Labs' diverse R&D data sources. It aims to empower researchers, scientists, and stakeholders by providing quick access to pertinent information aligned with their queries or areas of interest.

The implementation of this initiative represents a significant advancement in enhancing the efficiency and agility of information retrieval processes within Dr. Reddy's Laboratories. The project's successful execution has the potential to revolutionize how the organization leverages insights from its extensive R&D resources, fostering a culture of informed decision-making and innovation.

II. LITERATURE SURVEY

The current literature on Natural Language Processing (NLP), Machine Learning (ML), and Research and Development Data Management (R&D) provides valuable insights into the challenges and opportunities of this interdisciplinary field. Researchers have increasingly recognized the importance of leveraging advanced technologies to improve natural language research in the context of R&D. The literature shows a growing consensus on the need for intelligent systems that can understand and extract information from unstructured textual data.

In the field of machine learning, many studies emphasize the potential of algorithms to improve language understanding and information service. Techniques such as deep learning, recurrent neural networks (RNNs), and transformer models have shown significant success in a variety of NLP tasks, including sentiment analysis, entity recognition, and language understanding. These advances form the basis of our proposed framework and highlight the importance of leveraging machine learning to improve the cognitive capabilities of R&D data management systems.

Research examining natural language processing in the context of R&D highlights the challenges presented by the broad and diverse nature of the research literature. NLP techniques, including syntactic and semantic analysis, have been used to extract meaningful information and relationships from complex texts. However, the existing literature recognizes the need for more sophisticated approaches to address the nuances inherent in the R&D language. Our research is consistent with these findings and aims to promote a new approach to improve natural language research that is adapted to the complexity of the R&D domain**.**

[1] Brianna Mueller, Takahiro Kinoshita, Alexander Peebles, Mark A. Graber, Sangil Lee.

The core of artificial intelligence (AI) lies in its capacity to imitate human intelligence, allowing for the execution of tasks, recognition of patterns, and prediction of outcomes by assimilating data from diverse sources. In various sectors such as autonomous driving, e-commerce, fintech, and healthcare, machine learning (ML) algorithms, notably deep learning, play a pivotal role. Despite their impressive performance, the practical implementation of these algorithms in clinical settings encounters challenges due to a lack of transparency. [1] Explainable AI (XAI) has emerged as a solution to address this issue, elucidating internal decisions to healthcare professionals, building trust, and facilitating the informed application of predictive models. The incorporation of XAI into healthcare settings is an ongoing exploration, influenced by the intricate nature of medical knowledge. [1] The success of AI in healthcare relies not only on advanced algorithms but also on a human-in-the-loop approach, engaging stakeholders and underscoring the significance of clear, iterative interactions between end users and AI systems to ensure meaningful clinical and operational capabilities.

[2] Shaan Khurshid, Christopher Reeder, Lia X. Harrington Et al.

Biomedical research increasingly relies on electronic health record (EHR) databases, pivotal for biological discoveries and clinical insights. These databases provide robust statistical power for extensive association analyses, encompassing diverse clinical features and measures such as risk factors, laboratory results, and imaging data. However, the acknowledgment of potential biases in EHR data, stemming from patient selection and missing data, is growing. Ascertainment bias related to clinical need and selection bias due to missing data pose significant challenges. [2] The study introduces the Community Care Cohort Project (C3PO), an EHR-based initiative designed to enhance cardiovascular disease research by mitigating ascertainment bias and minimizing data missingness. Utilizing deep natural language processing (NLP) to extract vital sign features from unstructured notes, the study aims to compare effective sample sizes before and after missing data recovery. [2] Additionally, established clinical risk scores are deployed to assess model performance, with the hypothesis that scores derived from prospective cohort settings will exhibit reduced bias in C3PO.

[3] Sheela Kolluri , Jianchang Lin, Rachael Liu, Yanwei Zhang & Wenwen Zhang.

In recent years, artificial intelligence (AI) and machine learning (ML) have emerged as transformative forces in pharmaceutical research and development (R&D), driven by advancements in computational technology and increased data accessibility. [3] The escalating costs of drug development have further fueled interest in the automated and predictive capabilities of AI/ML techniques. While ML has played a crucial role in drug discovery, the latest impact is observed in clinical trial design, conduct, and analysis, particularly accelerated by the digital shift during the COVID-19 pandemic. [3] This growth potential brings challenges, emphasizing the need for proper training of ML algorithms and careful consideration of data quality and research questions. Addressing concerns such as under-represented patient populations and ensuring patient privacy is crucial. [3] The text stresses the importance of understanding when different AI/ML methods are most effective and highlights the significance of statistical judgment in drug development, emphasizing that the scientific method remains indispensable amid technological advancements.

[4] Oscar N. E. Kjell , Sverker Sikström, Katarina Kjell & H. Andrew Schwartz.

The research reveals that language-based evaluations exhibit a level of reliability comparable to traditional rating scales for both the Harmony in Life and Satisfaction with Life measures. [4] The reliability measures are consistent with test–retest values. Utilizing contextualized embeddings, language-based assessments demonstrate a strong predictive correlation with scale ratings—0.85 for Harmony and 0.80 for Satisfaction—exceeding conventional correlation levels. [4] Supplementary tables present detailed descriptive statistics.

[5] Iqbal H. Sarker.

This manuscript examines the importance of artificial intelligence (AI) in the context of Industry 4.0, underscoring the pivotal role of AI-based modeling in constructing intelligent systems across a spectrum of applications. [5] It investigates ten essential AI techniques, encompassing machine learning and deep learning, illustrating their relevance in sectors such as business, finance, healthcare, and cybersecurity. Emphasizing the dynamic challenges of real-world scenarios, the document provides a comprehensive overview intended to assist scholars, industry practitioners, and decision-makers. [5] The concluding remarks highlight the diverse benefits of AI across fields, anticipate future trends, address research

challenges, and position the study as a valuable resource for prospective research and development endeavors.

[6] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith Et al.

The increasing interest in utilizing artificial intelligence (AI) to enhance healthcare involves extracting patient characteristics from electronic health records (EHRs). [6] Despite the prevalence of structured data in EHRs, clinicians often prefer unstructured clinical narratives, presenting a challenge for computational models. Natural language processing (NLP), particularly transformer architectures such as BERT, plays a crucial role in addressing this challenge. [6] These models, trained on extensive text data, surpass the performance of previous NLP models across various healthcare tasks. The rise of large-scale transformer models, exemplified by GPT-3 with billions of parameters, further elevates their effectiveness. In the biomedical field, BioBERT and PubMedBERT demonstrate notable advancements with millions of parameters. [6] This paper underscores the potential of transformer models, detailing their training stages and the transformative impact they have on healthcare AI systems.

[7] Kelei He , Chen Gan , Zhuoyuan Li , Islem Rekik , Zihao Yin Et al.

This manuscript delves into the increasing influence of transformers in medical image analysis, expanding their success beyond natural language processing to encompass computer vision. It delves into their applications in diverse clinical tasks such as image synthesis, registration, segmentation, detection, and diagnosis. [7] The comprehensive overview encompasses the fundamental concepts of the attention mechanism and various transformer architectures customized for medical imaging. The text addresses challenges, including diverse learning paradigms and enhancing efficiency. [7] The conclusion underscores the transformative impact of transformers in computer vision and the ongoing research developments in medical image analysis. [7] It highlights the necessity for sophisticated methodologies and explores future avenues in medical transformer research, encompassing weakly supervised learning, multi-modal learning, multi-task learning, and general model enhancements.

[8] Hui Wen Loh , Chui Ping Ooi , Silvia Seoni , Prabal Datta Barua Et al.

This review delves into the integration of artificial intelligence (AI) within healthcare applications, specifically addressing the trust concerns associated with opaque AI models through the emergence of explainable artificial intelligence (XAI). It identifies specific healthcare areas requiring additional attention from the XAI research community. [8] The comprehensive review encompasses 99 Q1 articles, evaluating diverse XAI techniques like SHAP, LIME, and GradCAM. The results underscore the necessity for increased focus on detecting abnormalities in 1D biosignals and identifying crucial text in clinical notes. The review concludes by advocating for a holistic cloud system tailored for smart cities to advance healthcare services. [8] Notably, SHAP stands out as a widely employed technique for elucidating clinical features, while GradCAM excels in providing visual explanations for medical images. [8] The review accentuates the importance of outlining future research directions in XAI for healthcare, striving for comprehensive solutions to meet the dynamic needs of smart cities.

[9] Michele Salvagno , Fabio Silvio Taccone & Alberto Giovanni Gerli.

This manuscript delves into the application of the Artificial Intelligence Chatbot, specifically ChatGPT, in the realm of scientific writing. Fueled by the Generative Pre-trained Transformer (GPT) language model, ChatGPT exhibits the potential to assist researchers in tasks like material organization, draft generation, and proofreading. [9] The AI tool finds relevance in literature review, research question identification, and providing insights into the current state of the field. While ChatGPT offers time-saving benefits in clinical practice, the discussion raises ethical considerations, including concerns about plagiarism risks and accessibility imbalances. [9] The paper underscores the necessity for global academic regulations to govern the use of chatbots in scientific writing, advocating for mechanisms to detect and penalize unethical practices. Despite ChatGPT's ability to offer valuable feedback, the importance of human expertise, judgment, and validation in decision-making is highlighted. [9] The conclusion emphasizes ChatGPT's recommendation for a meticulous review, focusing on proper formatting, grammar, spelling, and the inclusion of a robust conclusion. It reiterates the indispensable role of human oversight in clinical practice.

[10] Christopher C. Yang.

This paper emphasizes the importance of artificial intelligence (AI) and machine learning across diverse sectors, with a specific focus on healthcare. [10] Despite their effectiveness in predictive modeling, the lack of transparency presents challenges in practical clinical deployment. The emergence of Explainable AI (XAI) addresses this concern by offering insights into internal decision-making, fostering trust among healthcare professionals. [10] The conclusion underscores the crucial role of transparency in successful predictive modeling within healthcare. XAI is recognized for its capacity to address information-based and instance-based clarification questions, ensuring user comprehension and trust. [10] The paper advocates for active involvement of end-users in AI integration, stressing iterative interactions and the engagement of all stakeholders for the meaningful implementation of AI in healthcare.

IV. METHODOLOGY

The Methodology for Advancing Cognitive Natural Language Research Using Machine Learning and Natural Language Processing (NLP) in R&D Data Management involves a systematic and iterative process to develop an intelligent framework.

The following steps describe the main components of our methodology:

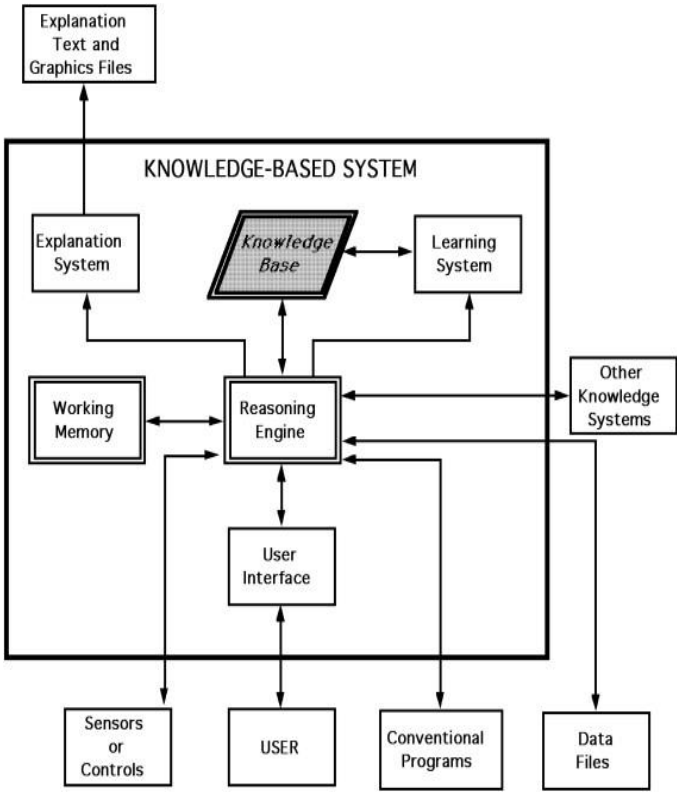

Fig 4.1 Methodology

Understanding the R&D Data Structure:

To begin, a comprehensive analysis of the shared folder housing crucial R&D reports and findings in PowerPoint (PPT) format is essential. This phase involves understanding the nature of data stored in these slides.
The phase dedicated to understanding the R&D data structure involves a comprehensive exploration of the repository storing valuable PowerPoint (PPT) slides containing vital information related to Dr. Reddy's Labs' research and development activities. This phase aims to conduct a detailed analysis of the folder's content, exploring various themes and domains covered within these slides. Through content sampling, metadata analysis, and keyword identification, the objective is to discern prevalent topics, content types, and the structure of information present in the slides. By establishing a classification scheme based on identified themes, this exploration sets the foundation for efficient data processing and extraction methods. The insights gained here provide a fundamental understanding of the repository's contents, aiding in the subsequent phases of the project by facilitating organized data retrieval and processing within the Knowledge Builder application.

Data Extraction and Preprocessing:

Developing algorithms using libraries like pptx and PyPDF2 is crucial for extracting textual content from PPT slides and PDF files, ensuring a thorough extraction process for all stored information.
Algorithms will be designed to interpret and extract text content from PPT slides and PDF files, ensuring no crucial information is missed during the extraction process. By employing these libraries, the aim is to encompass a broad scope of content, including text, images, graphs, and tables embedded within these documents. The algorithms will undergo rigorous testing to guarantee their accuracy and reliability in capturing the diverse range of information embedded within the files.

Web Application Development:

Leveraging frameworks like Streamlit will help create an intuitive and user-friendly interface for the "Knowledge Builder" application, ensuring easy interaction and query submission.
Utilizing Streamlit or similar frameworks, the focus is on crafting a visually appealing and ergonomic interface that simplifies user interactions. This involves designing an interface layout that facilitates easy query submission and navigation through the application's functionalities.
A key aspect of this phase is the development of a responsive and intuitive design, ensuring accessibility and user-friendliness. This includes features such as clear navigation elements, user-friendly input fields for query submission, and an aesthetically pleasing yet functional layout.

Natural Language Processing (NLP) Techniques:

Implementing NLP libraries like Spacy enables efficient processing of extracted text data, including tokenization, entity recognition, and the construction of a comprehensive knowledge graph.
In this phase, Spacy or similar NLP libraries serve as powerful tools for text analysis. They assist in identifying entities such as people, organizations, quantities, dates, and more, along with recognizing the relationships between these entities within the context of the extracted content. The extracted information aids in the creation of a structured knowledge graph, where entities serve as nodes, and their relationships form edges, offering a visual representation of the interconnected information. Additionally, the application of NLP techniques helps in identifying key phrases, themes, and contextual information within the text, which becomes instrumental in generating concise and contextually relevant responses to user queries within the Knowledge Builder application.

Knowledge Graph Formation:

Utilizing extracted text data, a structured knowledge graph will be built, mapping entities, relationships, and relevant insights found within the R&D reports for efficient retrieval.
The methodology employed here primarily relies on Natural Language Processing (NLP) techniques, specifically Spacy, to process and analyze the extracted text. This involves tokenization, entity recognition, and the creation of relationships between these entities based on contextual information. Through this approach, a structured knowledge graph is formulated, capturing the interconnectedness and relevance of different entities and their associated data within the R&D reports and findings.

Transformer-Based Models:

Integration of models such as BART (Bidirectional and Auto-Regressive Transformers) from the Transformers library will facilitate the generation of concise and informative responses to user queries.
The Transformer-Based Models, such as BART, operate by utilizing advanced machine learning techniques, particularly in natural language processing (NLP). These models employ a bidirectional and autoregressive approach, enabling them to comprehend the context and nuances embedded within user queries. The bidirectional aspect assists in capturing the interdependencies between different words or phrases in the query, while the autoregressive nature enables the models to generate responses progressively, taking into account the query's context.

Algorithm Integration and System Testing:

The integration of text extraction, NLP processing, knowledge graph creation, and response generation algorithms into the web

application will be followed by rigorous testing to ensure accuracy and responsiveness.
Testing constitutes a critical aspect of this phase, encompassing rigorous assessments to evaluate the system's responsiveness and its ability to handle diverse search queries effectively. The aim is to guarantee that the integrated algorithms operate harmoniously, allowing the application to swiftly process user queries and deliver relevant insights from the R&D data repository. Such comprehensive testing scenarios simulate real-world usage scenarios, ensuring that the system performs optimally under various conditions.

User Engagement and Feedback Loop:

Deploying the application internally for user testing and feedback collection is vital. Iterative refinement based on this feedback will enhance usability, accuracy, and overall functionality.
The deployment of the web application within the organization enables direct interaction with end-users, allowing them to utilize the application's features and provide valuable feedback. This feedback loop plays a pivotal role in understanding user experiences, preferences, and any encountered challenges or shortcomings.

Documentation and Training:

Preparing detailed documentation elucidating the system's functionalities and conducting training sessions will ensure users are adept at utilizing the application effectively.
The phase concerning "Documentation and Training" holds significant importance in the overall project lifecycle after the thorough understanding of the R&D data structure. Documentation aims to create comprehensive, user-friendly guidelines detailing the functionalities and intricacies of the Knowledge Builder application. It involves compiling detailed instructions, methodologies, and system functionalities into a coherent document. This documentation acts as a crucial reference point for users, aiding in their understanding of the application's capabilities and functionalities.

Identified Gaps and Challenges:

Automated Knowledge Extraction from PPT Slides:
Lack of comprehensive methodologies for automated extraction of meaningful insights from PowerPoint slides.

Intelligent NLP for R&D Data:
Need for specialized NLP models tailored to comprehend scientific language in pharmaceutical R&D.

Leveraging Shared Folder Data:
Inadequate utilization of shared folder data, hindering efficient extraction of relevant information.

Insightful Search Mechanisms:
Limitations in providing contextually relevant insights based on user queries within the stored PPT data.

Web App Development:
Insufficient focus on user-centric web app design for effective knowledge extraction.

Data Analysis and Interpretation:
Need for tools ensuring comprehensive analysis and interpretation of data from PPT slides.

Automation in Insight Generation:
Lack of fully automated systems capable of swift query processing and insight generation.

Integration of Data Formats:
Challenges in integrating diverse data formats seamlessly for unified knowledge extraction.

Quality Assurance:
Lack of robust quality assurance measures for the accuracy and reliability of extracted insights.

Scalability and Adaptability:
Potential gaps in scalability to accommodate future expansions or changes within the R&D landscape.

In conclusion, the methodology for understanding the R&D data structure within the shared repository of PowerPoint (PPT) slides is vital for establishing a comprehensive understanding of the available information. Through meticulous content analysis, metadata examination, and keyword identification, this phase provides valuable insights into the diverse themes and content types present in the slides. By developing a structured classification scheme based on these findings, this methodology sets the stage for efficient data processing and extraction in subsequent project phases. This foundational understanding is crucial for enabling organized information retrieval and effective utilization of the repository's contents within the Knowledge Builder application developed for Dr. Reddy's Labs' research and development initiatives.

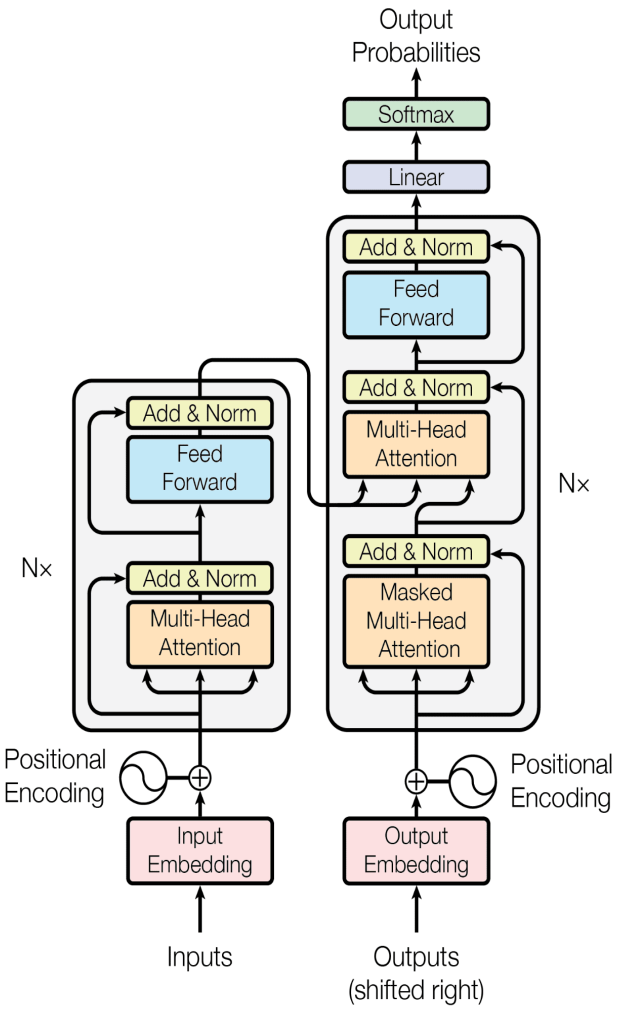

Fig 4.2 Transformer Architecture

V.SYSTEM DESIGN AND IMPLEMENTATION

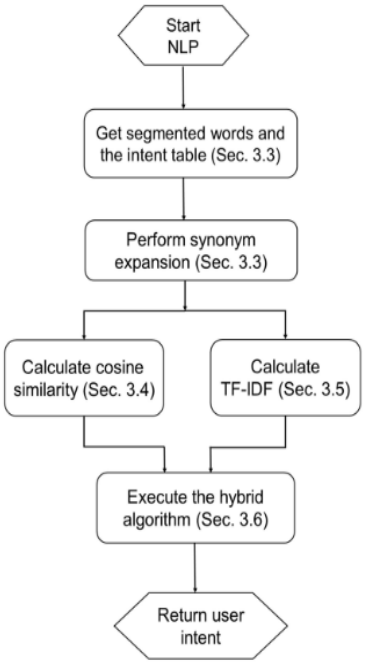

Fig 5.1: System Overview

System Design

Based on machine learning and natural language processing (NLP) to advance cognitive natural language research for R&D knowledge management, our system is complexly structured to take advantage of the synergy between cutting-edge technologies. At the core of our design is a modular architecture that combines machine learning models, NLP components and a robust data management system. The system is implemented in Python using powerful libraries such as Streamlit, spaCy, Transformers and other key tools for streamlined development. Machine learning models, including recurrent neural networks (RNNs) and transformer-based models such as BERT, were chosen for their ability to understand complex linguistic nuances and patterns in domain-specific data.

The information management system is built on the principles of presenting information using RDF, which provides a semantic structure for efficient organization and retrieval of information. Integrating relevant entity recognition (NER) and sentiment analysis with spaCy improves system and text understanding and contributes to more accurate information serving. For versatility, our design has information from various sources, such as PowerPoint presentations, websites and documents, allowing the system to extract knowledge from various R&D materials.

The user interface developed with Streamlit is designed with user-centered principles that emphasize simplicity and intuitiveness. Users can seamlessly interact with the system by entering questions and receiving answers in an understandable format. Scalability and performance have been considered with Docker containers and strict load-testing procedures that ensure optimal functionality even with variable workloads.

In summary, our system design strategically combines machine learning, NLP, and knowledge management components to provide a solid platform for natural language research and data collection in the R&D domain. The modular design and careful selection of technologies positions our system to significantly advance the development of informed research and management of R&D information

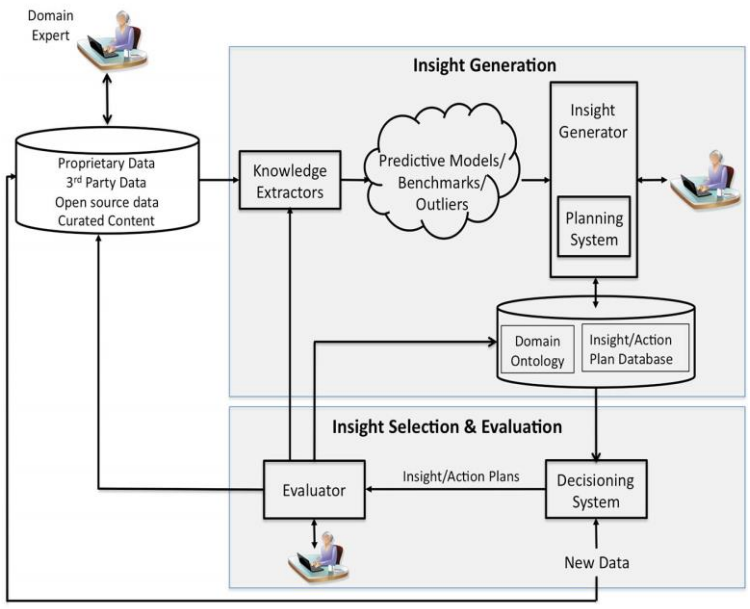

Fig.5.2: Insight generation

Implementation

Programming language and frameworks:

Choose appropriate programming languages (e.g. Python) and frameworks (e.g. TensorFlow, PyTorch) to implement machine learning and NLP components.

Data processing pipeline:

Create a robust data processing pipeline that includes data processing, cleaning and transformation.

NLP Implementation:

Implement NLP techniques using libraries such as NLTK, spaCy or Hugging Face Transformers. Ensure efficient handling of unstructured text data.

Machine Learning Model Implementation:

Apply selected machine learning models to document classification, topic modeling and other related tasks .Evaluate and refine models based on performance metrics.

Data graph implementation:

Use graph database technologies (eg Neo4j) to implement the data graph. Develop algorithms to update and expand the graph as new information is added.

User interface development:

Build a web-based or desktop user interface for researchers to interact with the system. Include functions such as search, filters and visualization tools to explore the data graph.

Integration and Testing:

Integrate all system components and perform thorough testing to ensure functionality and reliability. Perform unit testing, integration testing and validation against various R&D datasets..

Environment setup

Libraries Installation:

Installed required libraries using pip, including Streamlit , spaCy, converters, pptx, pandas, queries, BeautifulSoup, PyPDF2, etc. Compatibility and version consistency are ensured.

Model selection:

Selected and installed machine learning models such as BART (Bart For Conditional Generation) and its associated tokenizer for natural language generation.

Data extraction and preprocessing

PPT extraction:

Developed functions to extract text from PowerPoint (PPT) files using the pptx library. Ensured robustness handling exceptions during deletion.

Extracting web pages and PDF files:

Implemented functions to extract text from web pages with BeautifulSoup and from PDF files with PyPDF2. Cached web page data to improve performance and avoid unnecessary requests. Files uploaded by

Users:

users can submit PPT, PDF or DOCX files for custom data extraction. Supports multiple file uploads and handled different file formats.

Natural Language Processing (NLP) Components

Named Entity Recognition (NER):

Used spaCyand#039's (en_core_web_sm) pretrained model for NER to recognize text data.A data graph has been created based on the detected entities.

Search mechanism:

Developed search mechanism that can be used to query a data graph and retrieve relevant information based on user queries.

User Interface

Streamlit application:

Developed Streamlit application for user interaction. Allows users to choose from existing data sources and upload their own files.

Query Handling:

Implemented a search function where users can enter queries related to R&D data. Showed available source files and allowed users to enter queries to retrieve data.

Feedback and Results:

Back user-friendly interface with feedback messages like success, errors and loading wheels. Results are presented to users in a clear and understandable format.

Model response generation

Response generation:

Implemented functions to generate responses on user queries and data graph. You used selected machine learning models to generate answers.

Response Presentation:

Cleaned and generated responses presented in a user-readable format. Contains error messages in cases where no data was found.

Deployment On-premises deployment:

Ran the Streamlit application locally for testing. Make sure all dependencies and settings are correct.

Cloud implementation:

The application is implemented in a cloud environment to ensure wider availability. Guaranteed scalability and performance optimization for concurrent users.

Testing and Validation Unit Testing:

Each component has been unit tested to ensure functionality. Fixed possible bugs or errors.

User Testing:

Engaged potential users in testing to get feedback on the app and its usability. Improvements based on user feedback.

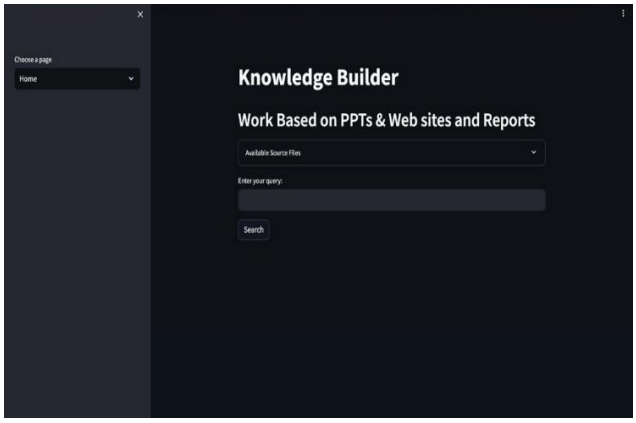

Fig 6.1: Home

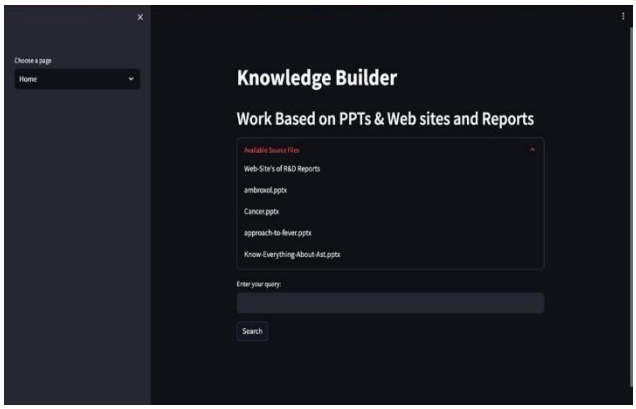

Fig 6.2: Choosing file

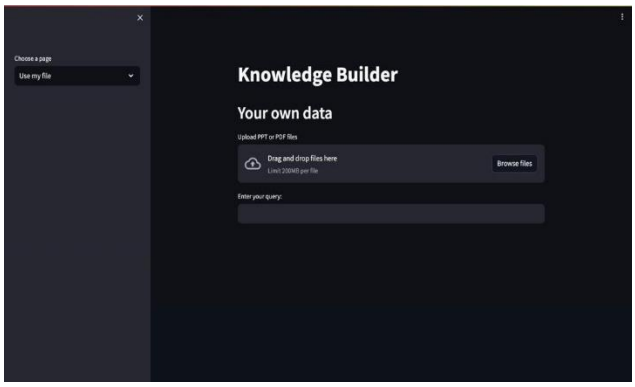

Fig 6.3: Browse file

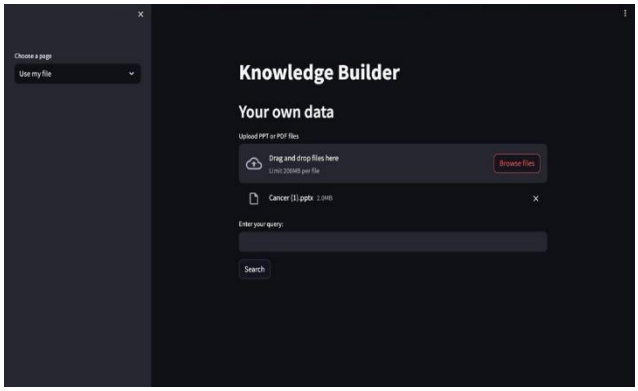

Fig 6.4: File uploaded

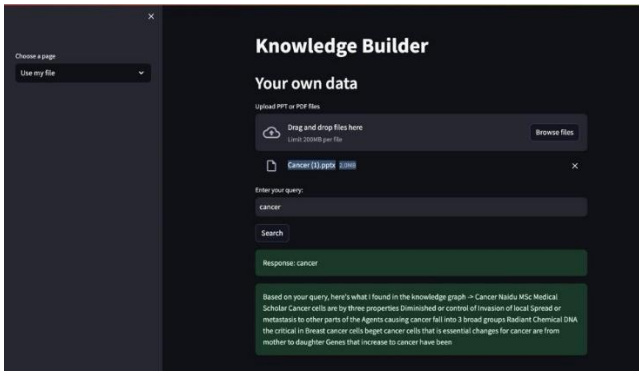

Fig 6.5: Search query

CONCLUSION

The inception of a sophisticated software solution for natural language search marks a significant advancement in Dr. Reddy's Laboratories' knowledge management journey. This innovative initiative seeks to transform the manner in which the institution taps into its vast collection of R&D documents, predominantly archived as PowerPoint (PPT) presentations.

By harnessing cutting-edge technologies such as Streamlit, Spacy, and Transformers, among others, this anticipated digital platform is poised to automate the intricate processes of data extraction and synthesis. Serving as a pivotal bridge between expansive data sets and actionable intelligence, it empowers users to distill valuable insights from the plethora of available materials with unparalleled ease.

Crafted to facilitate seamless knowledge extraction and offer real-time insights tailored to specific search criteria, this pioneering software symbolizes a strategic advancement in equipping Dr. Reddy's Labs' researchers, scientists, and stakeholders. Beyond enhancing the efficacy of R&D document utilization, it democratizes access to critical data, nurturing an environment conducive to enlightened decision-making and innovation.

Prioritizing user-centric design principles, the platform promises intuitive navigation across the shared repository of PPT files. Leveraging automated knowledge aggregation techniques, it emerges as an invaluable asset, swiftly uncovering relevant information aligned with user queries, thereby significantly reducing the time traditionally expended on manual analysis.

Far beyond a mere technological solution, this envisioned platform embodies Dr. Reddy's Laboratories' unwavering dedication to harnessing technological advancements to elevate research standards. Its imminent deployment heralds a transformative phase, championing data-centric strategies and propelling the organization towards heightened operational efficiency and industry leadership in the pharmaceutical sector.

In summary, the advent of this advanced software solution for natural language search represents a paradigm shift, introducing a transformative toolset poised to enhance the discovery and application of R&D insights within Dr. Reddy's Laboratories. It signifies not just a technical achievement but a cultural evolution towards embracing data-driven methodologies and fostering groundbreaking innovations.

At its core, this initiative encapsulates the organization's overarching ambition to fully leverage technological capabilities, paving the way for unprecedented advancements in knowledge management and catalytic breakthroughs in pharmaceutical research and development**.**

.

FUTURE SCOPE AND INDUSTRIES

Intelligent Natural Language Search System for Industries:

This section describes the architecture and functionality of an intelligent natural language search system adapted for industrial R&D. The primary focus is on how the system responds to industry-specific challenges, facilitates information discovery, and facilitates seamless integration into existing workflows.

Case Studies:

real case studies are presented to demonstrate the effectiveness of the proposed approach in various industrial environments. These case studies show how an intelligent natural language search system is driving better knowledge management, innovation and R&D decision making across industries.

Test Results and Evaluation:

The study presents the results of tests conducted to evaluate the performance of an intelligent natural language retrieval system. Metrics such as precision, recall, and user satisfaction are used to evaluate the system and its effectiveness in improving R&D data exploration.

Discussion:

This section interprets the experimental results and discusses implications and practical implications for industries implementing

intelligent natural language retrieval systems. Aspects of scalability, adaptability and potential challenges are considered.

REFERENCES

[1] Brianna Mueller, Takahiro Kinoshita, Alexander Peebles,Mark A. Graber, Sangil Lee Artificial intelligence and machine learning in emergency medicine (01 March 2022).

[2] Shaan Khurshid, Christopher Reeder, Lia X. Harrington, Pulkit Singh, Gopal Sarma, et al. Cohort design and natural language processing to reduce bias in electronic health records research .5, Article number: 47 (2022).

[3] Sheela Kolluri , Jianchang Lin, Rachael Liu, Yanwei Zhang & Wenwen Zhang Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development The AAPS Journal 24, Article number: 19 (2022).

[4] Oscar N. E. Kjell , Sverker Sikström, Katarina Kjell & H. Andrew Schwartz Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy Scientific Reports 12, Article number: 3918 (2022).

[5] Iqbal H. Sarker1, AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems .2 Received: 20 July 2021 / Accepted: 21 January 2022 / Published online: 10 February 2022 © The Author(s) 2022.

[6] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian & Yonghui Wu npj . A large language model for electronic health records .Digital Medicine 5, Article number: 194 (2022) .

[7] Kelei He , Chen Gan , Zhuoyuan Li , Islem Rekik , Zihao Yin , Wen Ji , Yang Gao , Qian Wang , Junfeng Zhang , Dinggang Shen . Transformers in medical image analysis .

[8] Hui Wen Loh , Chui Ping Ooi , Silvia Seoni , Prabal Datta Barua , Filippo Molinari , U Rajendra Acharya . Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011– 2022).

[9] Michele Salvagno , Fabio Silvio Taccone & Alberto Giovanni Gerli Critical Care .Can artificial intelligence help for scientific writing? 27, Article number: 75 (2023) 62k Accesses 120 Citations 225 Altmetric.

[10] Christopher C. Yang1 .Explainable Artifcial Intelligence for Predictive Modeling in Healthcare Received: 1 September 2021 / Revised: 29 December 2021 / Accepted: 3 January 2022\ Published online: 11 February 2022 © The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022.