INTERNSHIP REPORT

(MAY 2023 - AUGUST 2023)

Summer Internship Report submitted
by

M.RAHUL MANIKANDAN
(Bachelor of Engineering at
SSN College of Engineering)

Under the Guidance of

PROF. SATYANARAYANAN SESHADRI
(Department of Applied Mechanics-IIT MADRAS)

TO

ROBERT BOSCH CENTRE FOR DATA SCIENCE AND ARTIFICIAL INTELLIGENCE-IIT MADRAS

RBCDSAI IITM

Table of Contents:

Acknowledgement

I want to express my sincere thanks to the **Robert Bosch Centre for Data Science and Artificial Intelligence** for granting me this valuable internship opportunity. The Center's guidance and support have significantly influenced my learning experience.

I want to express my gratitude to **Prof. V. Rajini**, head of the EEE department, and **Prof. S.Tamil Selvi**, my mentor, from SSN College of Engineering for allowing me to undertake this internship. Your support and permission have been of great significance to me.

I am truly grateful to **Prof. Satyanarayanan Seshadri** for entrusting me with this project. His constant support and guidance have been a major source of motivation throughout this project.

I extend my gratitude to **Prof. Raghunathan Rengaswamy** and **Prof. Preeti Aghalayam** for consistently providing feedback and guidance. Their input has been crucial in shaping my approach and aligning my efforts with the project's goals.

I want to express my gratitude to **Ram Kishore Sankaralingam**, **Ajay Koushik V** and **Arun Muthukkumaran** for continuous support and guidance during my internship. Their helpful ideas and teamwork have been essential in tackling challenges and making progress.

Introduction

In recent years, the concerning path of climate change has emphasized the need for quick and collective action across societies. A major cause of this problem is the increasing release of greenhouse gasses due to human activities.Carbon dioxide (CO_2) is the biggest culprit, making up about 65% of these emissions. Since 1970, CO_2 emissions have risen by nearly 90%, mostly because of burning fossil fuels and industrial activities. This makes it really important to closely watch and reduce emissions from industries, farming, and transportation.

Understanding how each part of an industry creates CO_2 is crucial. This helps us find the key factors causing emissions. Figuring these out lets us improve current technologies or even invent new ones to cut down on emissions. To do this, we need to look at different parts working together. This way, we can spot the most important sources of CO_2 emissions and come up with strategies to meet different goals, like getting to net-zero carbon emissions.

Throughout this internship, I’ve been part of a team focused on developing a tool that analyzes CO_2 production across various industry sectors by leveraging fundamental principles. This tool gives us a clear picture of how much CO_2 industries are producing. With this information, we can make smart decisions about how to lower emissions.

My task was to use Natural Language Processing (NLP) techniques to gather useful information from sustainability reports. I used tools like LangChain and OpenAI API to make this process easier. The aim was to quickly get the necessary data from these reports and use it to better understand emissions. This way, I can connect the detailed information in sustainability reports by developing a Question and Answer Interface to our bigger plan of analyzing emissions and finding effective ways to improve the situation.

Motivation

I was part of a team that faced a practical challenge—collecting data from sustainability reports manually. These reports were often massive, stretching beyond 1000 pages. Going through all that data by hand was not just time-consuming, but also prone to mistakes.

Understanding the effort that went into this painstaking process and realizing how important accuracy was in sustainability reporting, I saw a need for a smarter approach. The idea of building an automated tool struck me as a way to make things easier. My goal was to speed up the data collection and make it more accurate by automating the process.

I wanted to create something that could simplify my work and reduce the chances of errors.The excitement to revamp the data collection process using technology was the driving force behind my initiative. The chance to utilize advanced techniques like Natural Language Processing (NLP) and new AI tech got me really motivated.

I aimed to build something that not only made my work smoother but could also be useful for others in the sustainability field. So, the motivation behind this tool was all about making my work efficient, accurate, and possibly inspiring positive changes beyond just my tasks.

Problem Statement

The manual extraction of data from extensive sustainability reports poses significant challenges. With reports often exceeding 1000 pages, this process is time-consuming and error-prone. The resulting inaccuracies undermine the reliability of the extracted data and its subsequent analysis.

This inefficiency hampers accurate sustainability reporting, impacting decision-making and strategy formulation. A solution is needed to streamline data extraction, reducing errors and enhancing efficiency. By integrating technologies like Natural Language Processing (NLP) and AI, we aim to create an automated process that ensures accurate and timely data extraction from these reports.

Methodology

The methodology employed for creating an efficient and accurate data extraction process through a sustainable data extraction chatbot is elaborated below. This approach merges Natural Language Processing (NLP) techniques with advanced AI technologies, including the utilization of OpenAI's GPT-4 API and the integration of LangChain for comprehensive text analysis using Python.

- ➢ **Data Collection and Preprocessing:**
  - ○ **Sourcing Sustainability Reports:** Begin by obtaining sustainability reports from various organizations. These reports serve as the primary data source for information extraction.
  - ○ **Text Preprocessing:** Clean the obtained text data to remove irrelevant content, special characters, and formatting inconsistencies. This step ensures the subsequent analysis is carried out on structured text.

- ➢ **Text Chunking and Embeddings:**
  - ○ **Utilizing LangChain for Chunking:** Leverage LangChain, a robust language processing tool, for the purpose of dividing preprocessed text into meaningful chunks. LangChain's algorithms and functionalities ensure efficient and accurate chunking.

  - **Text Embeddings:** Implement NLP techniques to convert the generated chunks into numerical embeddings. These embeddings capture the semantic essence of the text, enabling similarity analysis.

- **Knowledge Base Creation:**
  - **Storing Embeddings:** Store the text embeddings within a knowledge-based vector. This storage method enables rapid retrieval and comparison of information during user interactions.

- **User Interaction and Question Processing:**
  - **User Queries:** Allow users to input questions related to sustainability data.
  - **Embeddings for User Queries:** Transform user queries into embeddings for semantic comparison with the knowledge-based chunks.

- **Semantic Similarity Search:**
  - **Leveraging LangChain for Similarity Search:** Deploy LangChain's semantic similarity capabilities to match user query embeddings with stored knowledge-based chunks.

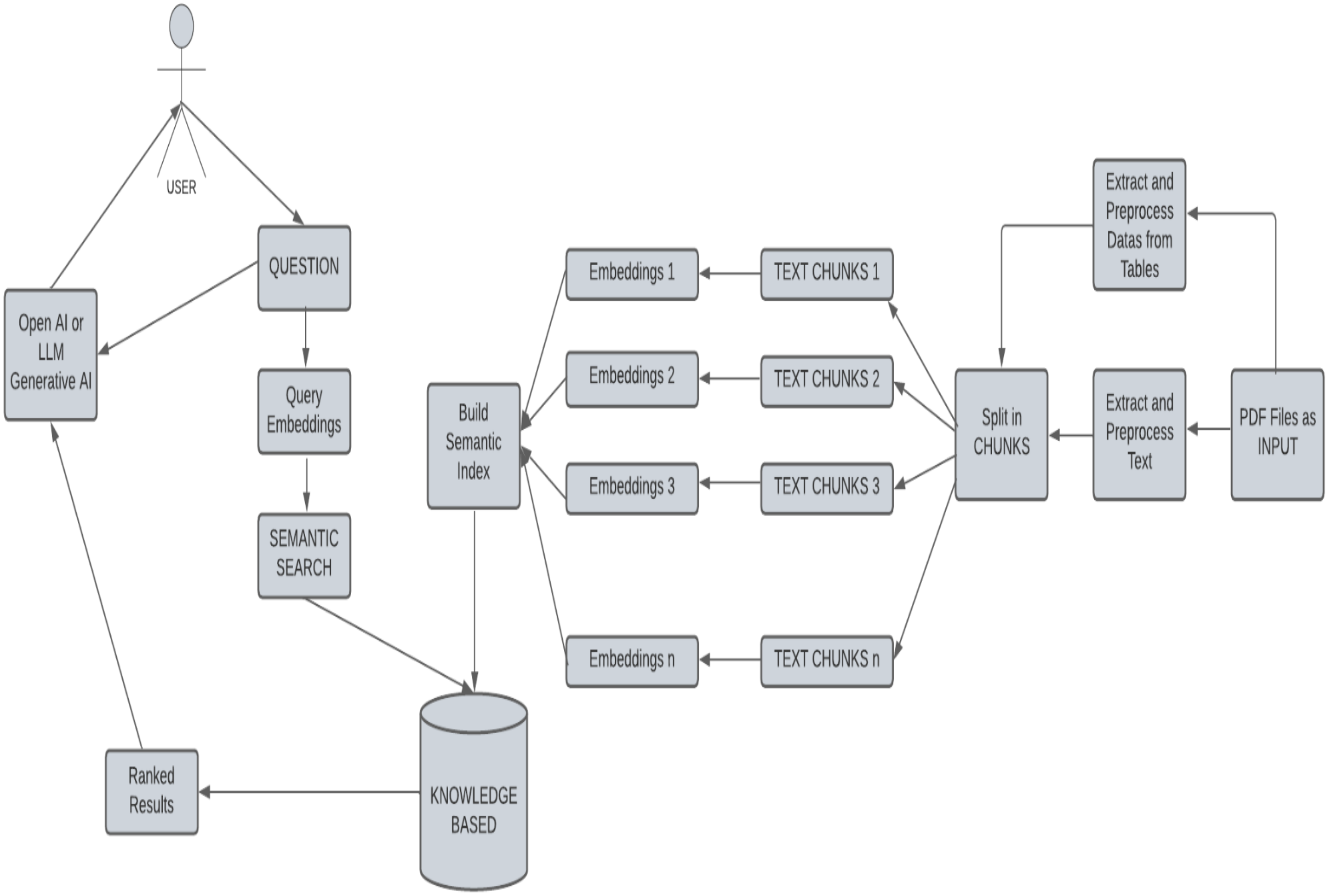
Identify the most relevant chunks aligned with the user's question.

- ➢ **OpenAI API - GPT-4 and Cost Analysis:**
  - ○ **Direct Integration for Testing:** Initially, integrate OpenAI's GPT-4 API directly for testing purposes. Generate answers based on user queries and retrieved knowledge chunks.
  - ○ **Challenges of Direct API Use:** Identify limitations in terms of high costs and token usage constraints associated with the direct utilization of the OpenAI API for extensive data extraction.

- ➢ **Comparative Analysis and Optimization:**
  - ○ **Method Cost Comparison:** Compare the costs incurred through direct OpenAI API integration with the LangChain method for chunking, embedding, and similarity analysis.
  - ○ **Cost-Efficient Approach:** Recognized that the LangChain method offers significant cost savings due to efficient resource utilization compared to the direct API integration.

The developed methodology capitalizes on the integration of NLP techniques, LangChain's capabilities for chunking and similarity analysis, and OpenAI's GPT-4 API for answering user queries. By adopting a cost-efficient approach centered around LangChain's functionalities, the chatbot not only streamlines sustainability data extraction but also ensures accurate and insightful answers. This innovative solution aligns sustainability with cutting-edge AI and NLP technologies while optimizing costs for enhanced efficiency.

Flowchart:



Results and Inference

The implementation of the sustainable data extraction chatbot yielded promising outcomes in extracting information from various sources within sustainability reports, including text narratives and tables. The chatbot's capabilities were tested on a diverse set of sustainability reports, demonstrating its efficacy in capturing relevant data across different formats.

- **Text Extraction:**

The chatbot successfully extracted information from textual sections of sustainability reports. It identified key metrics, objectives, and achievements articulated within the reports' narratives. This included details about environmental goals, social impact initiatives, and corporate sustainability strategies. The chatbot's ability to comprehend and retrieve information from complex language structures showcased its proficiency in handling textual data.

- **Table Data Extraction:**

The chatbot was able to navigate through tabular data present in sustainability reports. It identified and extracted numerical values, percentages, and other relevant data points from tables. This encompassed financial data, emissions statistics, and performance metrics. The chatbot's aptitude in recognizing patterns and discerning data relationships within tabular formats demonstrated its versatility in handling structured data.

The chatbot's proficiency in comprehending and extracting data from diverse sources within sustainability reports highlights its potential in automating the data extraction process. By effectively handling text and tables, the chatbot contributes to comprehensive and accurate data acquisition for sustainability analysis.

In summary, the sustainable data extraction chatbot showcased its ability to effectively retrieve information from textual narratives and tabular data within sustainability reports. Its versatility in handling different data formats contributes to streamlined data extraction and supports informed decision-making in the realm of sustainability analysis and reporting.

Future Work

- **Extracting Information from Graphs using Image Processing:**

  Exploring the extraction of information from graphical elements like charts and graphs using image processing techniques is a promising avenue. Developing algorithms to interpret and translate graphical data into quantifiable values could further expand the chatbot's data extraction capabilities. This would allow the chatbot to directly extract data points from visual representations, contributing to a more comprehensive data collection process.

- **Integrating Output into the Climate Action Tool:**

  A significant future step involves integrating the output generated by the sustainable data extraction chatbot into the climate action tool. This integration enables the seamless transfer of data obtained from sustainability reports into the tool. For instance, by feeding the extracted coal consumption data into the climate action tool, accurate CO2 emission estimates can be calculated, aiding in the mitigation of carbon emissions from industries.

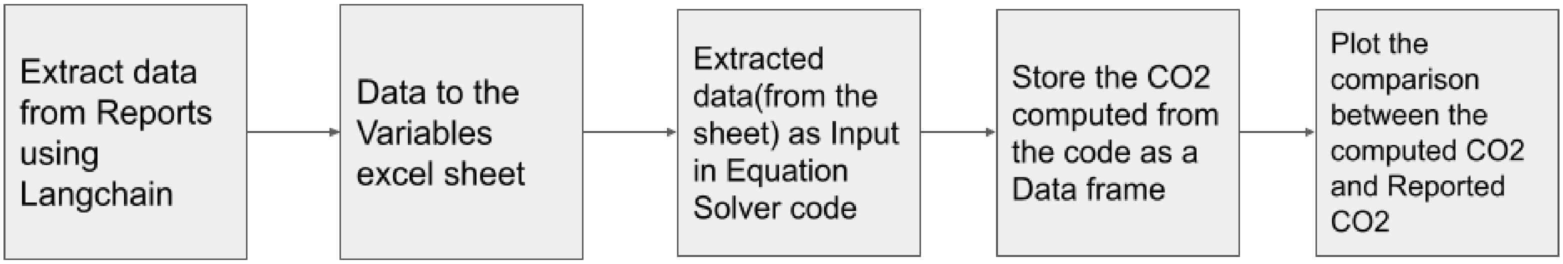
- **Automating Validation and Comparative Analysis:**

  Automating the validation process by linking the data extracted from sustainability reports with the climate action tool's CO_2 emission calculations is a critical aspect of future work. This automation eliminates the need for manual data input, making the validation process more streamlined and accurate. The chatbot-generated data can be directly fed into the tool, generating CO_2 emission estimates that can be automatically compared with the reported CO_2 emissions from industries.

- **Automated Comparative Plotting:**

  Automating the process of plotting and visualizing the comparison between estimated CO_2 emissions based on the extracted data and reported CO_2 emissions from industries is a valuable enhancement. The development of scripts that generate these comparative plots using the chatbot's output and the tool's calculations will provide an immediate visual representation of the tool's accuracy.

Automated Validation Process


Extract data from Reports using Langchain
Data to the Variables excel sheet
Extracted data(from the sheet) as Input in Equation Solver code
Store the CO2 computed from the code as a Data frame
Plot the comparison between the computed CO2 and Reported CO2

- **Iterative Learning and Refinement:**

  The sustainable data extraction chatbot's performance can be further improved through iterative learning and refinement. Ongoing training and updates based on new sustainability reports will enhance the chatbot's ability to handle diverse report formats, ensuring accurate data extraction and validation.

Concentrating on upcoming work domains like automating sustainability data retrieval and CO_2 emissions computation can lead to substantial automation and streamlining. Introducing parameter optimization to authenticate industrial sustainability reports takes these innovations a step further. This not only improves tool efficiency but also guarantees precise sustainability reporting, reinforcing climate action by adjusting parameters to align reported CO_2 emissions with the model's calculations.

Conclusion

In conclusion, my research internship has been a journey of exploring sustainable data extraction. The urgency of climate change served as the backdrop, prompting the development of an automated solution to extract crucial sustainability data from extensive reports.

The motivation stemmed from the need to simplify data collection from voluminous reports. Building an efficient tool to process these reports aligned with the demand for accurate sustainability information. This drive to contribute to climate action initiatives fueled my commitment to this project.

The challenge lay in automating data extraction from diverse formats like text, tables, and graphs. Manual methods were inadequate for the complexity. The solution, utilizing NLP techniques and LangChain, enabled data extraction from various sources.

The methodology combined NLP and LangChain for data extraction. OpenAI's GPT-4 API added advanced language generation capabilities. Results demonstrated the chatbot's ability to extract data from text and tables, aligning well with sustainability's diverse nature.

Future work includes exploring graphical data extraction using image processing and integrating chatbot output into a climate action tool. This could automate validation, comparing tool-calculated CO_2 emissions with reported emissions.

In conclusion, this internship illuminated efficient sustainability data extraction. By merging AI technologies and innovative approaches, the journey resulted in a tool with potential for sustainability reporting. Ongoing improvements and collaboration affirm its value in addressing global challenges.