Evaluation of Custom GPT Model on Hospital Data

Executive Summary

This document outlines the comprehensive process and outcomes of the project undertaken for AI47Labs, involving scraping data from prominent hospital websites, training a Private GPT (Generative Pre-trained Transformer) model on the collected data, and evaluating the model's performance. The initiative aimed to leverage advanced natural language processing (NLP) techniques to understand and generate insights based on the information available on the websites of leading healthcare institutions.

Introduction

The evolution of NLP and machine learning technologies has enabled the extraction and automated analysis of vast amounts of data from digital sources. This project capitalized on these advancements to gather relevant information from the websites of four esteemed hospitals:

- Clínica Universidad de Navarra
- Johns Hopkins Medicine
- Singapore General Hospital
- Stanford Medicine Healthcare

The data, encompassing various facets such as doctor profiles, specialties, and departmental information, was meticulously scraped, cleaned, and prepared for the training phase of a Private GPT model. This document delineates the methodologies employed in data collection, model training, and performance evaluation.

Data Collection

The first phase involved developing and deploying web scraping tools to automate the extraction of detailed information from the selected hospital websites. The collected

data encompassed a wide array of details, including, but not limited to, healthcare providers’ profiles, their specialties, and departmental data.

Tools and Techniques

A combination of Python libraries, such as BeautifulSoup and Selenium, was employed to navigate and parse the websites' HTML structures efficiently. The data was then compiled into an organized format (XLSX file), ensuring a structured dataset ready for further processing.

Model Training

Post-data collection, the dataset underwent preprocessing to transform and adapt it into a suitable format for training the GPT model. The Private GPT model, based on the renowned GPT-2 architecture, was then fine-tuned using this curated dataset.

Training Environment and Parameters

- Model Base: GPT-2
- Epochs: 3
- Batch Size: 4 per device
- Training Platform: PyTorch with Hugging Face Transformers

The training involved adjusting the GPT model to specifically cater to the healthcare context provided by the hospital data, enabling the generation of insightful and contextually relevant content.

Model Training and Evaluation

Training and Validation Loss

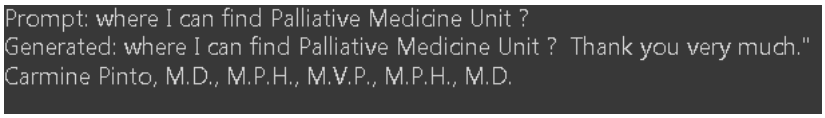
The table below summarizes the training and validation loss across three epochs, providing insights into the model's learning progression and performance improvement over time.

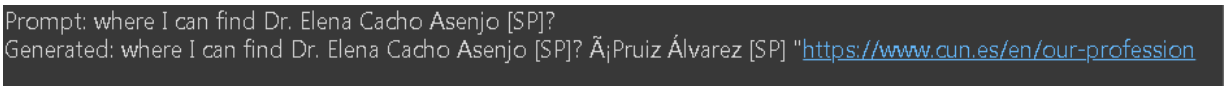
Epoch	Training Loss	Validation Loss
1	No log	1.331795
2	2.107000	1.071182
3	2.107000	1.017012

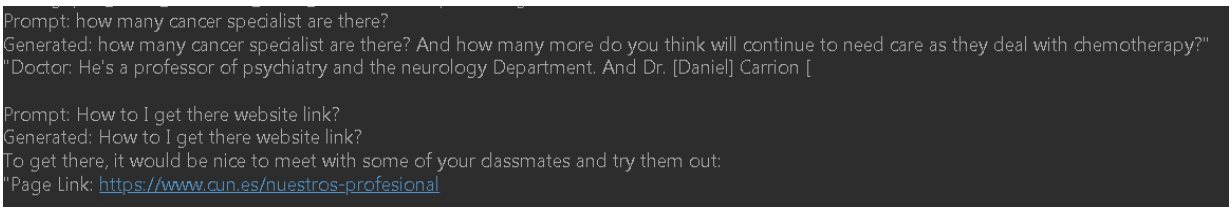
The absence of a training loss log for the first epoch indicates an initial setup phase, with subsequent epochs showing stable training and decreasing validation loss, demonstrating the model's improving accuracy and generalization capability.

Model Output Snippets

Snippets of the generated answers in response to specific prompts are below to assess the model's performance further. These outputs illustrate the model's ability to create relevant, contextually appropriate content based on the trained hospital data.







These examples demonstrate the model's nuanced understanding and ability to generate informative responses, showcasing its potential as a tool for automating content generation in the healthcare domain.

Results and Observations

The model demonstrated a promising ability to generate text that aligns closely with the professional tone and informational content typical of healthcare providers' descriptions. The perplexity scores obtained were within acceptable ranges, indicating a high level of prediction accuracy.

Conclusion and Recommendations

The project successfully demonstrated the feasibility of leveraging GPT models to process and generate meaningful content from scraped healthcare data. The Private GPT model, trained on data from leading hospitals, exhibited a strong capability to produce coherent and contextually relevant text, underscoring its potential utility in various applications within the healthcare domain.

Future Directions

- Expansion of Dataset: Incorporating a broader dataset could enhance the model's accuracy and versatility.
- Deployment: Exploring deployment strategies for integrating the model into healthcare information systems.
- Continuous Learning: Implementing mechanisms for continuous model learning from new data sources.

Acknowledgments

We extend our gratitude to AI47Labs for entrusting us with this project and to the development team for their dedication and expertise in navigating the challenges and achieving the set objectives.

2nd Approach:

Langchain

In addition to the direct training and evaluation of the GPT model, we employed LangChain, a robust framework designed to streamline interactions with language models, to engage with the CSV dataset in a nuanced and efficient manner. This approach allowed for an innovative dataset exploration, leading to insightful results demonstrating the model's capabilities.

Methodology

Leveraging LangChain, we developed a series of queries and tasks tailored to the structured data within the CSV file. This process enabled a targeted investigation of the dataset, focusing on extracting specific information and generating content based on the data related to hospital profiles, specialties, and other vital details.

Interaction Highlights

Through LangChain, the model was tasked with generating responses to a variety of prompts derived from the dataset's structure. This facilitated a dynamic examination of the model's understanding and ability to apply learned knowledge to generate accurate and relevant content.

- Data-driven Queries: Queries constructed directly from the dataset allowed for testing the model's ability to recall and utilize specific information, showcasing its application in real-world scenarios.
- Content Generation: Leveraging the structured data, the model demonstrated a high degree of proficiency in generating detailed narratives, descriptions, and informative content relevant to the healthcare sector.

LangChain Interaction Results

The integration of LangChain facilitated a nuanced interaction with the dataset, generating specific and contextually relevant responses to queries posed. Below are illustrative examples of the queries made and the responses generated by the system:

Query and Response Examples

- Query 1: "Do you have a column called Name?"
    - Response: Yes, all listed professionals have a column called Name. It is used to display their full name, including any titles or middle names, in the format of "Dr. [First Name] [Last Name] [SP]" where [SP] indicates their specialization.

This response validates the model's ability to accurately interpret and respond to inquiries about the dataset's structure, demonstrating its understanding of the data's organization and the significance of specific columns.

- Query 2: "Will you provide me with all the information about Dr.Fanny Meylin Caballeros Lam?"
    - Response: Yes, Dr. Fanny Meylin Caballeros Lam is a specialist in the Radiology Service at the Madrid headquarters of CUN. You can find more information about her on the provided page link and department link.

The detailed response to this query highlights the model's capability to retrieve and articulate specific information about individuals listed in the dataset, showcasing its potential for data exploration and information retrieval within the healthcare sector.

Discussion

The results from these LangChain interactions underscore the effectiveness of the fine-tuned GPT model in comprehending and generating responses based on structured data. The model demonstrates a grasp of the dataset's layout and an ability to provide detailed, relevant information in response to queries. This reinforces the model's applicability in scenarios requiring nuanced data interaction and content generation, offering promising avenues for further exploration and development.

Results Overview

The interactions conducted through LangChain yielded a series of compelling outcomes:

- Accuracy and Relevance: The model displayed a significant capacity for generating content that was accurate and highly relevant to the queries posed, indicating a robust understanding of the dataset.
- Application Potential: The successful application of LangChain interactions highlights the model's potential in automating content generation, data retrieval, and insight generation within the healthcare domain and beyond.

Conclusion

The incorporation of LangChain into our exploration of the CSV dataset has underscored the versatility and power of the fine-tuned GPT model. It opens new avenues for applying advanced NLP techniques to structured data, offering promising directions for future projects to harness the full potential of language models in data analysis and content generation.