



Artificial Intelligence

PROJECT

Submitted By:

21I-2697 M Talal Qureshi

21I-1703 Hassan Saeed

21I-1709 M Owais Zahid

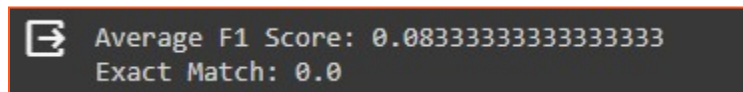
Documentation for Domain-Specific Chatbot Development

Introduction:

This project involves developing a domain-specific chatbot tailored to provide specialized assistance and information retrieval within particular fields such as medicine, law, engineering, or finance. The primary goal is to fine-tune a pre-trained BERT language model on domain-specific literature to enable the chatbot to understand and respond accurately to user queries related to the book's content.

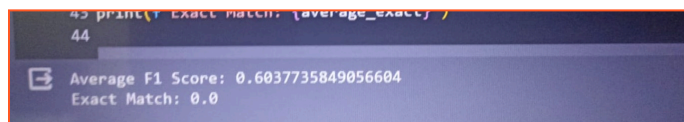
Data Preprocessing:

- **Data Segmentation:** The text data from the PDF source <https://mu.ac.in/wp-content/uploads/2021/05/Data-Structure-Final-.pdf> was segmented into different structures (chunks, lines, paragraphs) to find the optimal granularity for processing:
- **Chunks:** Initially, the data was segmented into chunks, which resulted in an F1 score of 0.08, indicating poor model performance due to insufficient context



```
43 print('Exact Match: {average_exact} /  
44  
45  
Average F1 Score: 0.08333333333333333  
Exact Match: 0.0
```

- **Lines:** Subsequently, data segmented into lines improved the F1 score to 0.60, suggesting better context capture but still lacking in coherence.



```
43 print('Exact Match: {average_exact} /  
44  
45  
Average F1 Score: 0.6037735849056604  
Exact Match: 0.0
```

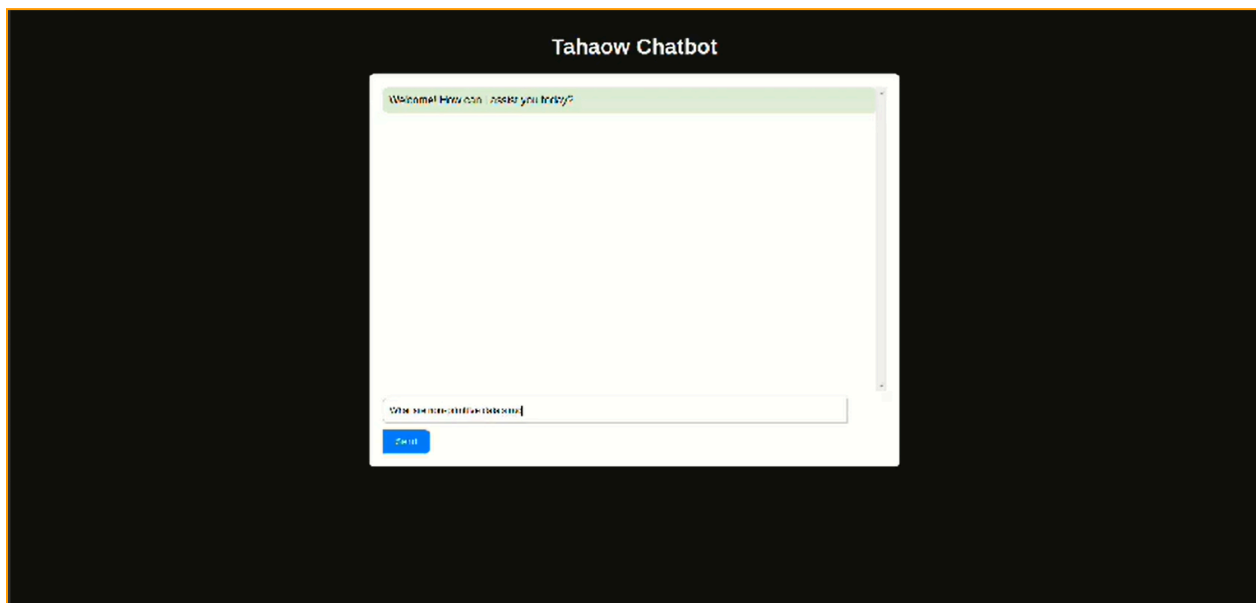
- **Paragraphs:** Finally, data segmented into paragraphs yielded the most satisfactory results, providing a balanced context for the model to generate accurate and coherent responses.

Model Training and Fine-tuning:

- **Model Choice:** The BERT model, specifically the bert-large-uncased-whole-word-masking-fine-tuned-squad, was chosen for its robust performance in question answering tasks.
- **Fine-tuning Approach:** The model was fine-tuned on the extracted text data, focusing on improving its understanding of domain-specific terminology and concepts. This was achieved by adjusting the model's parameters through transfer learning techniques.

GUI Development:

- **Interface Design:** A simple graphical user interface (GUI) was designed and implemented to facilitate user interaction with the chatbot.
- **Integration:** The GUI was integrated with the backend using Flask, enabling seamless communication between the user inputs and the chatbot responses.



Implementation Details

Text Extraction:

- PDF Processing: Text was extracted from a PDF using PyMuPDF and Requests. The PDF URL was accessed, and its content was streamed and read into memory.
- Text Saving: Extracted text, including page numbers for reference, was saved into a TXT file, which was then used for model training and fine-tuning.

Answer Retrieval:

- Relevance Filtering: To retrieve the most relevant paragraphs for a given query, a TF-IDF vectorizer and cosine similarity measures were used.
- Answer Extraction: The model utilized tokenized input to identify the most likely start and end of an answer within the relevant text, and responses were then compiled from these segments.

Results and Evaluation:

● Performance Metrics:

The effectiveness of the chatbot was evaluated using the F1 score at different stages of text segmentation, with the paragraph-based approach showing the best results.

● User Feedback:

Preliminary feedback from test users indicated that the chatbot effectively provided relevant and precise answers, enhancing user satisfaction and engagement.



Average F1 Score: 0.7917241379310345
Exact Match: 0.0



Question: What is an Searching?

Answer: the process of finding some particular element in the list sorting an item in a list the value to be searched is val = 5

Question :

Why do we choose BERT over Roberta,GPT,LAMA ? BERT was chosen over other models for several reasons:

- **Performance in Question Answering Tasks:**

The specific task of the chatbot involved understanding user queries and providing accurate responses related to domain-specific literature. BERT, especially the bert-large-uncased-whole-word-masking-fine-tuned-squad model, has demonstrated robust performance in answering tasks, making it a suitable choice for this project.

- **Pre-trained Representations:**

BERT's pre-trained representations capture bidirectional context from large amounts of text data, enabling it to understand and generate responses that take into account the context of the input query. This is crucial for providing accurate and coherent answers, particularly in domain-specific contexts where context is essential.

- **Transfer Learning Capabilities:**

BERT's architecture allows for efficient transfer learning, where the model can be fine-tuned on domain-specific data to adapt its understanding and response generation capabilities to the specific domain. This was important for improving the model's understanding of domain-specific terminology and concepts, ultimately enhancing its performance in the given task.

From an aerial view Roberta model looks better but for a small data-set and less resources scientists preferred to use bert as we researched.

Conclusion:

The development of a domain-specific chatbot presents a significant advancement in specialized fields, offering targeted assistance that enhances productivity and knowledge retention. This project not only demonstrated the feasibility of fine-tuning a sophisticated language model on domain-specific literature but also highlighted the importance of detailed dataset preprocessing and user-friendly GUI design in creating effective digital solutions.