Clinical Note Summarizer using FLAN-T5

Submitted by: Sean Samuel

1. Why This Project?

Doctors often spend a significant amount of time documenting clinical conversations after each visit. This can contribute to burnout, delays, and reduced time with patients.

I wanted to explore how deep learning can be applied to ease this burden - by automatically summarizing doctor-patient dialogues into brief, structured notes that resemble what doctors usually write. The goal is to help streamline documentation and improve clinical workflows.

2. What This Project Does

This project uses a language model to generate concise summaries of medical dialogues. The summaries mimic the kind of notes doctors write in Electronic Health Records (EHR), especially following a structure similar to SOAP (Subjective, Objective, Assessment, Plan).

By processing full-length medical conversations, the model outputs short, relevant summaries suitable for documentation and review.

3. Dataset

I used the MTS-Dialog dataset, a high-quality medical dialogue dataset that pairs real clinical conversations with professionally written summaries.

4. Model and Workflow

I used the 'google/flan-t5-small' model from Hugging Face. Each dialogue was converted into a summarization prompt and tokenized. The model was trained using Hugging Face's Trainer API and evaluated with ROUGE scores to check summary quality.

5. Example Output

Input:

Doctor: Are you experiencing any chest pain?

Patient: Yes, especially when I take deep breaths.

Generated Summary:

Patient reports chest pain when breathing deeply.

6. Screenshots

Sample Dialogue + Summary

Dialogue:
Doctor: Are you experiencing any chest pain?
Patient: Yes, especially when I take deep breaths.

Generated Summary:
Patient reports chest pain when breathing deeply.

ROUGE Evaluation Scores

ROUGE-1: 0.45
ROUGE-2: 0.32
ROUGE-L: 0.41

7. GitHub Repository

https://github.com/Seansamuelsamuel/clinical-note-summarizer

8. Evaluation and ROUGE

At the end of training, I used a metric called ROUGE (Recall-Oriented Understudy for Gisting Evaluation) to check how close the generated summaries were to the original human-written ones.

ROUGE compares the overlap of words and phrases between the model's output and the reference summary.

- ROUGE-1 looks at matching individual words
- ROUGE-2 checks for two-word pairs (bigrams)
- ROUGE-L finds the longest matching word sequence

My results were:
- ROUGE-1: 0.45
- ROUGE-2: 0.32
- ROUGE-L: 0.41

These scores show that the model is doing a decent job of generating relevant summaries that are similar to what a doctor might write manually.