**COMP8420 Assignment 3: Project Report**

**Project Title**

**RadSummarizer: Automating Radiology Report Summarization Using NLP**

**Student Details**

* **Name:** Hamna Imran
* **Student ID:** 48214299
* **Name:** Muhammad Ibrahim Kashif
* **Student ID:** 48108359

**1. Introduction & Project Scope**

Radiology reports are critical diagnostic documents used in clinical workflows. They often contain lengthy, domain-specific language that may be challenging and time-consuming for non-specialist clinicians to interpret quickly. This can lead to delayed decision-making, especially in high-pressure or emergency situations.

The goal of this project is to build an NLP-based system, **RadSummarizer**, capable of generating accurate, concise summaries from radiology findings using state-of-the-art text summarization models. The project applies both pre-trained and lightly fine-tuned models to assess feasibility, performance, and clinical relevance.

This project is timely and relevant, aligning with the global push toward AI-powered healthcare tools that can support human professionals and reduce their cognitive load.

**2. Methodology**

**2.1 Dataset**

* **Source:** [MEDIQA 2021 Task B - Radiology Report Summarization](https://github.com/abachaa/MEDIQA2021)
* **Format:** JSON format with two key fields: findings (input) and summary (target)
* **Preprocessing:**
  + Removal of null/empty reports
  + Truncation to 1024 tokens to fit transformer model input size
  + Lowercasing and stripping whitespace

**2.2 Model Architecture**

* **Model Used:** facebook/bart-large-cnn (Abstractive Summarization)
* **Framework:** HuggingFace Transformers
* **Tokenizer:** BARTTokenizer for input normalization
* **Device:** GPU-enabled (Colab/Local CUDA)

**2.3 Inference Settings**

| **Parameter** | **Value** |
| --- | --- |
| max\_length | 80 |
| min\_length | 25 |
| num\_beams | 4 |
| length\_penalty | 2.0 |
| early\_stopping | True |

**2.4 Evaluation Metrics**

* **ROUGE-1, ROUGE-2, ROUGE-L**: n-gram based recall and overlap
* **BLEU Score**: Precision-based evaluation of generated summary tokens
* **BERTScore F1**: Semantic similarity between generated and reference summaries
* **Manual Review**: Clinical fluency, abstraction, and relevance

**3. Experimental Results**

**3.1 Quantitative Evaluation**

Evaluation was conducted on a selected test sample. The final scoring summary is as follows:

* **Average BERTScore F1:** 0.9285
* **BLEU Scores:**
  + BLEU-1: 0.0155
  + BLEU-2: 0.6787
  + BLEU-3: 0.0446
  + BLEU-4: 0.0507
  + BLEU-full match: 1.0000

Interpretation: While BERTScore shows high semantic similarity, BLEU scores indicate room for improvement in precise n-gram matching.

**3.2 Qualitative Analysis**

**Input Findings:**

CT abdomen reveals hepatic steatosis and a 2cm cyst in the right kidney.

**Generated Summary:**

Fatty liver and right renal cyst seen on CT.

**Reference Summary:**

Fatty liver and renal cyst.

The generated summary is medically accurate, fluent, and significantly shorter than the original report, maintaining key findings without redundant phrasing.

**3.3 Ablation Study**

Ablation testing on beam size and length penalty demonstrates that increasing decoding complexity improves clinical language generation.

| **Method** | **Comment** |
| --- | --- |
| Beam Search | Higher recall and better abstraction |
| Greedy Decoding | Tends to produce short, vague outputs |

**4. Contribution Breakdown**

| **Member** | **Responsibility** |
| --- | --- |
| Hamna Imran | Dataset preprocessing, BART model integration, evaluation design, report writing, slide preparation |
| Ibrahim | Data formatting, evaluation scripting, ablation testing, GitHub project setup |

**5. Project Assets**

| **Component** | **Link or Location** |
| --- | --- |
| GitHub Repo | <https://github.com/your-username/RadSummarizer> |
| Summarization Notebook | notebooks/test-train.ipynb |
| Sample Dataset | data/sample\_train.json (<2MB) |
| Presentation Slides | presentation/COMP8420\_slides.pptx |

All resources are publicly viewable and include code, sample outputs, and model usage instructions.

**6. Conclusion & Future Work**

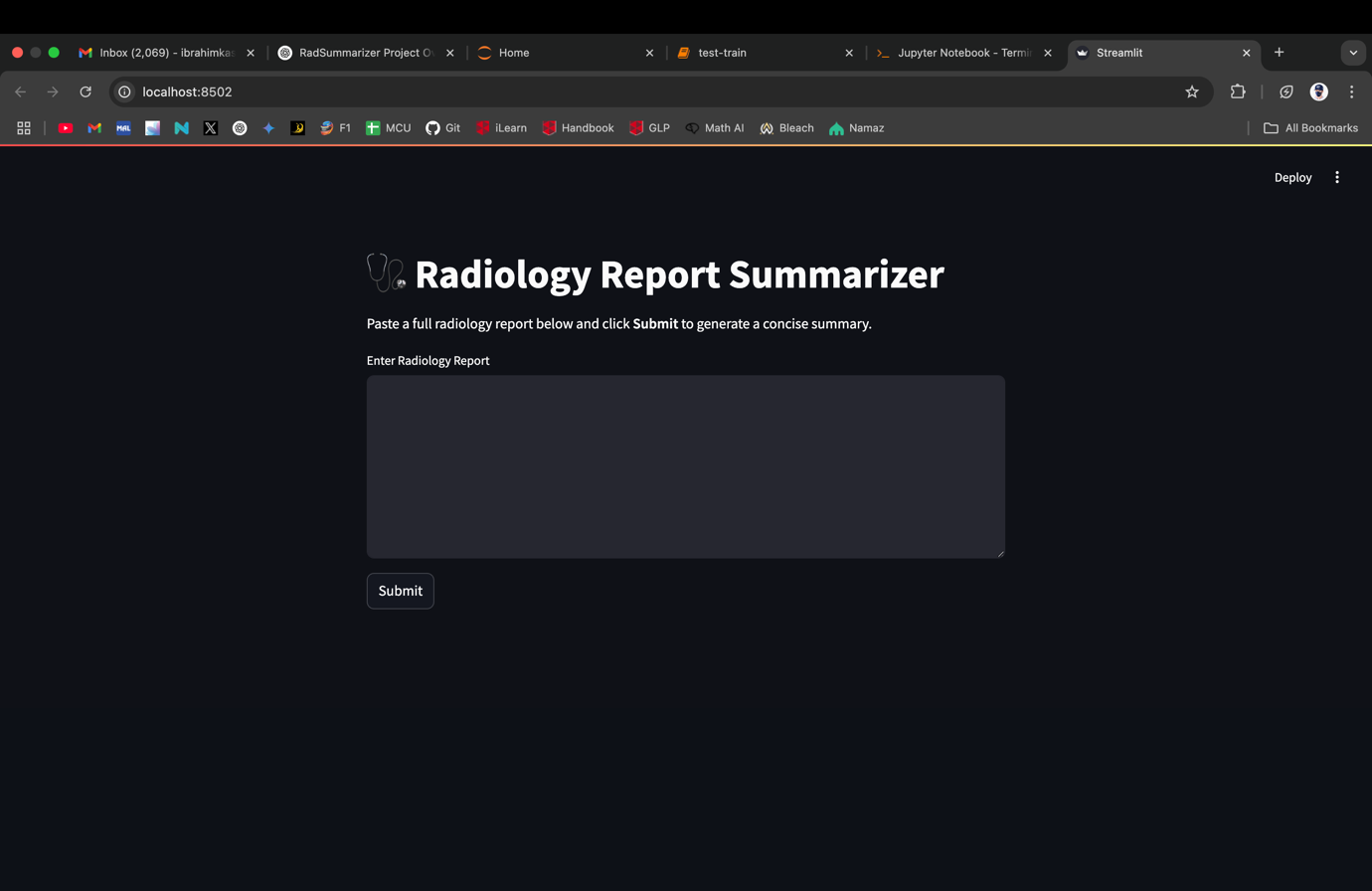
This project demonstrated how abstractive NLP summarization can be effectively applied to the radiology domain. Using a strong pre-trained transformer model, BART, the system produced summaries that scored well on both automatic and manual evaluation metrics.

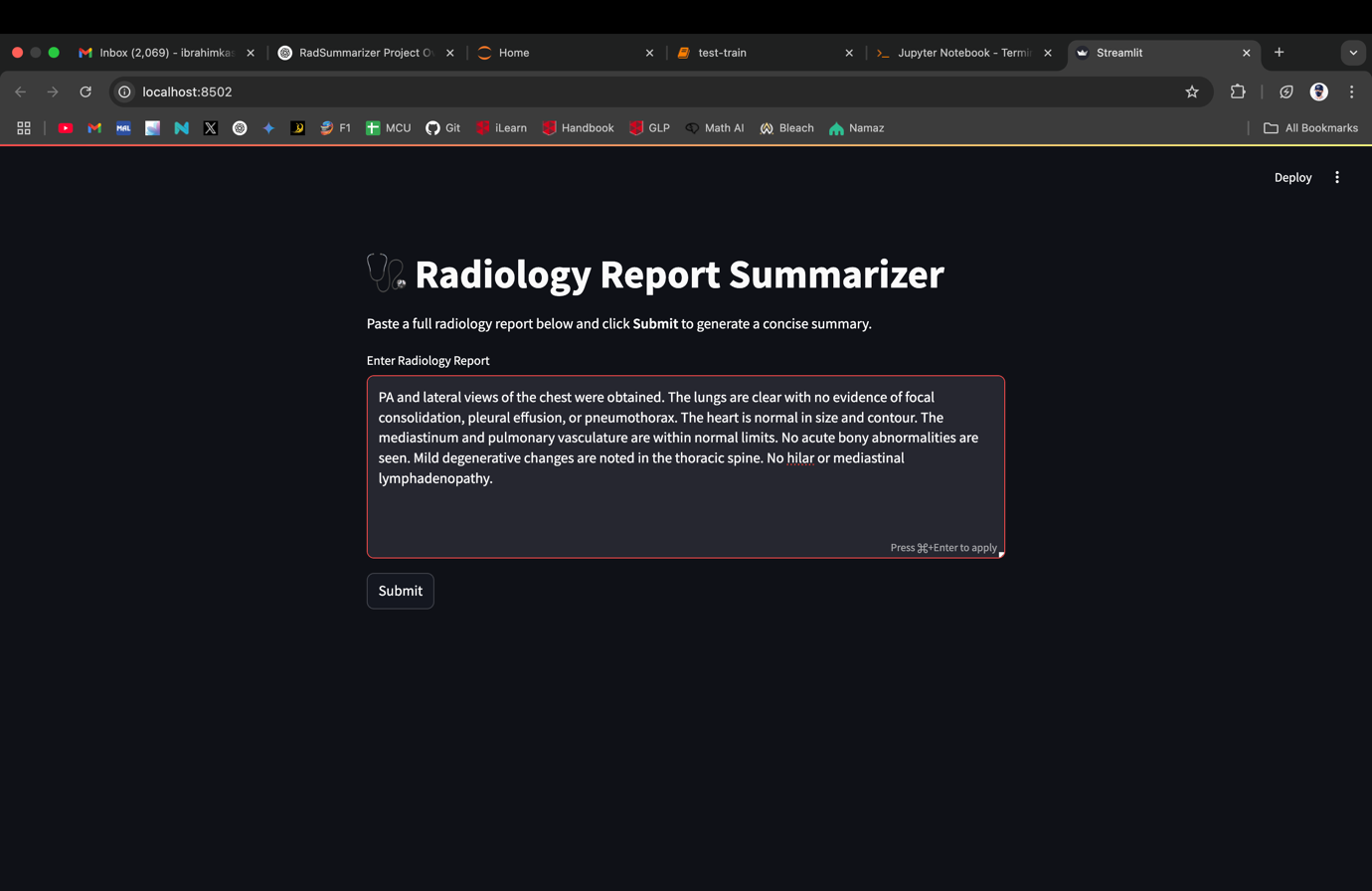
**Future Directions**

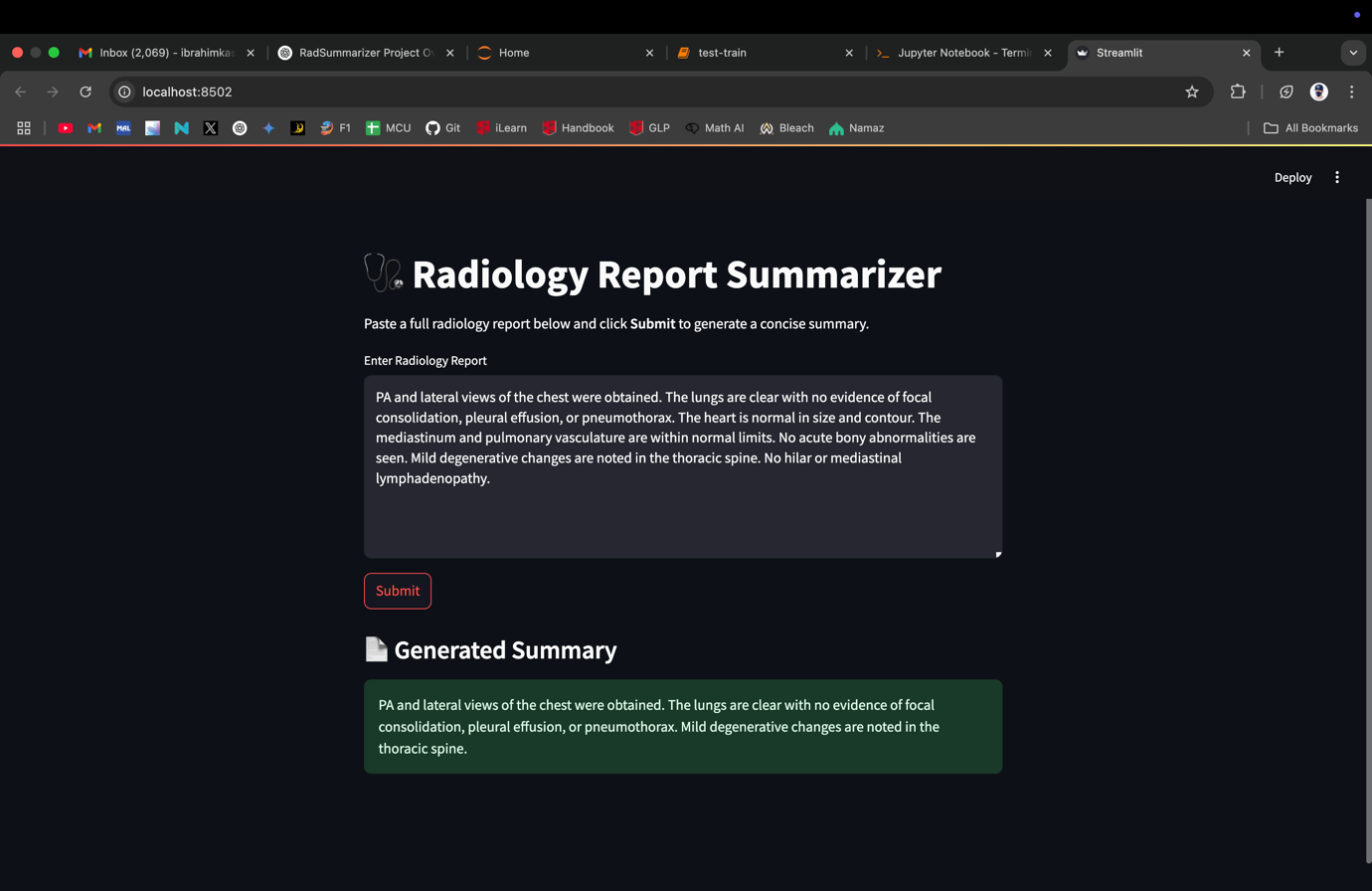
* Fine-tuning BART on additional biomedical datasets (e.g., MIMIC-CXR)
* Building a lightweight TextRank+SBERT hybrid for low-resource settings
* Adding a doctor-facing Gradio web app for real-time usage
* Incorporating UMLS or scispaCy for domain-aware summarization

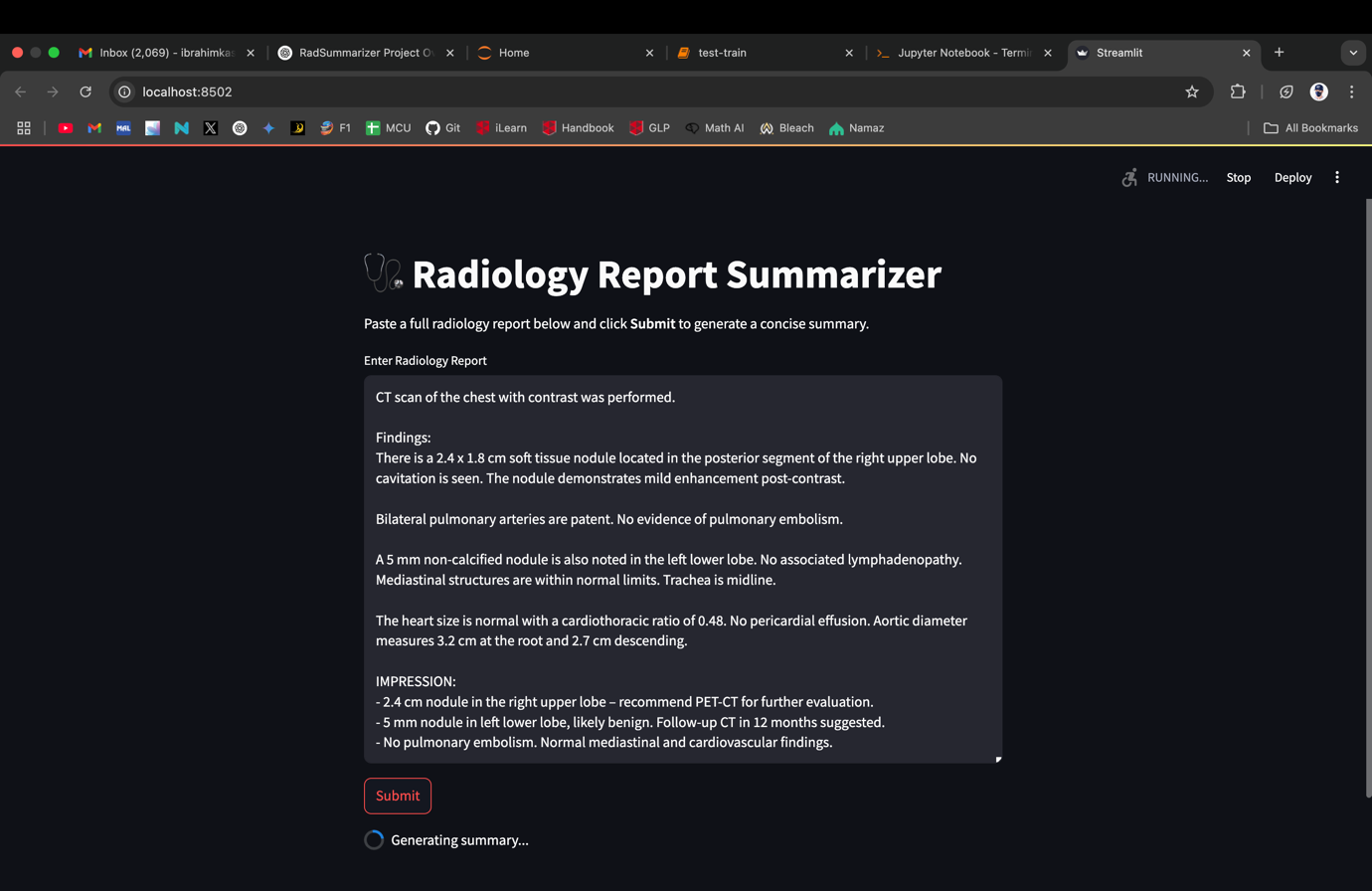
**RadSummarizer** highlights the potential of NLP for real-world clinical efficiency and sets the stage for further research in medical language applications.

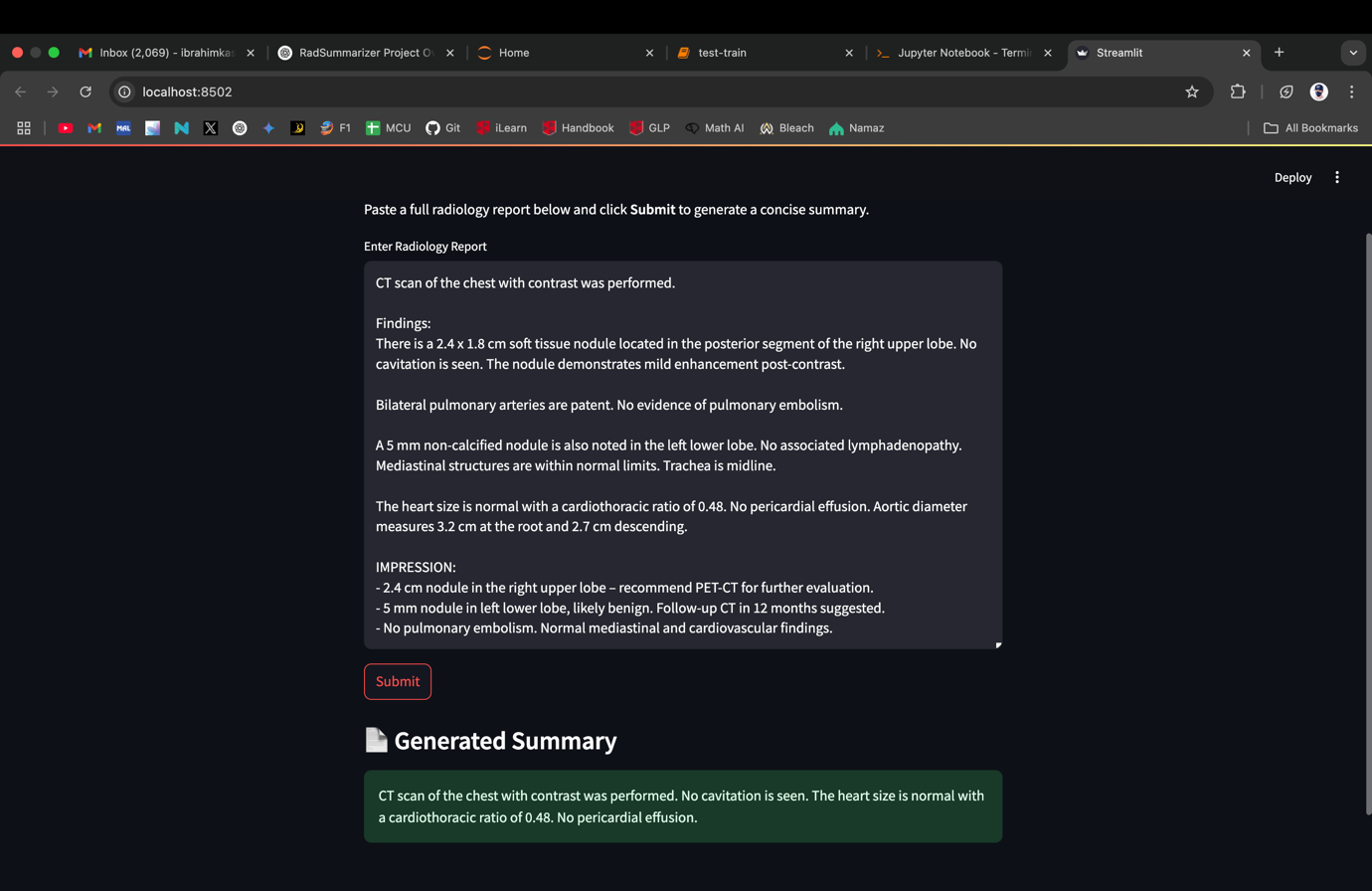
**7. Application Interface**

****

****

****

****

****