WPI

Generating the "Hospital Course" Section of Patient Discharge Summaries Using a Summarization Approach

Dalton Macres

DS504

November 19, 2023

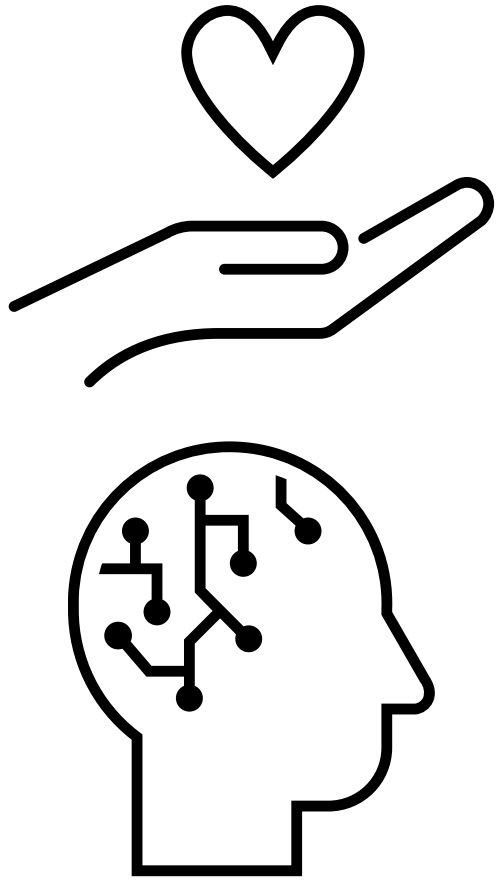
Agenda

1. Project Overview
2. Motivation
3. MIMIC-III Dataset
4. Project Approach
5. Results
6. Streamlit App
7. Future Directions
8. Resources

Project Overview

Project Overview

This project focused on deidentified patient healthcare data, specifically on patient notes. With the recent successes of large language models (LLMs) for natural language processing and understanding, this project aimed to evaluate various summarization techniques to auto-generate a section of patient discharge summaries.



Motivation

Motivation

- Doctors spend nearly 50% of their time doing paperwork
- Discharge summaries, a necessity for patient continuity, can often take days for a doctor to complete

GOAL: To investigate the efficacy of using LLMs to auto-generate a portion of the discharge summary (Brief Hospital Course) in efforts to reduce the turnaround time and associated costs, and thus to improve the continuity of patient care.

https://www.forbes.com/sites/brucelee/2016/09/07/doctors-wasting-over-two-thirds-of-their-time-doing-paperwork/?sh=359754c85d7b
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8322483/

MIMIC-III Dataset

1. MIMIC-III Overview
2. NOTEEVENTS Table
3. Obtaining the Dataset

MIMIC-III Dataset Overview

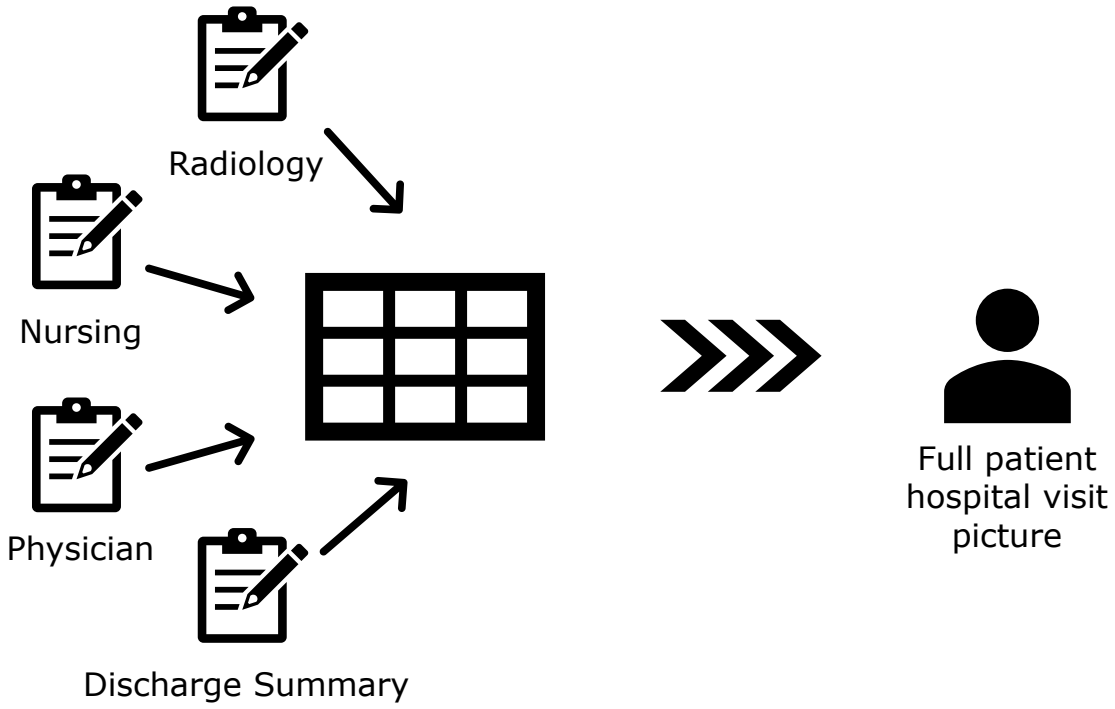
Medical Information Mart for Intensive Care (MIMIC), specifically MIMIC-III

- Deidentified health data from critical care patients at Beth Israel Deaconess Medical Center from the years 2001-2012
- Data highlights
  - ~60K admissions
  - ~2.1M note events
  - ~730M total records in the database

https://physionet.org/content/mimiciii/1.4/

NOTEEVENTS Table

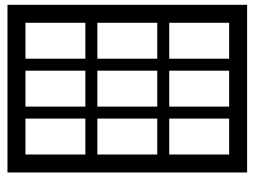
- Contains all notes for patients
  - Category, description, and text of the note
  - Subject & hospital admission linking keys
  - Other metadata, including note creation and upload timestamps into the EMR system



Obtaining the Dataset

- MIMIC-III dataset access requested and accepted through PhysioNet
- AWS CloudFormation stack was launched to deploy the MIMIC-III data in AWS Athena
- Query the data directly on AWS or locally using access keys

SELECT * FROM NOTEEVENTS



https://aws.amazon.com/blogs/big-data/perform-biomedical-informatics-without-a-database-using-mimic-iii-data-and-amazon-athena/

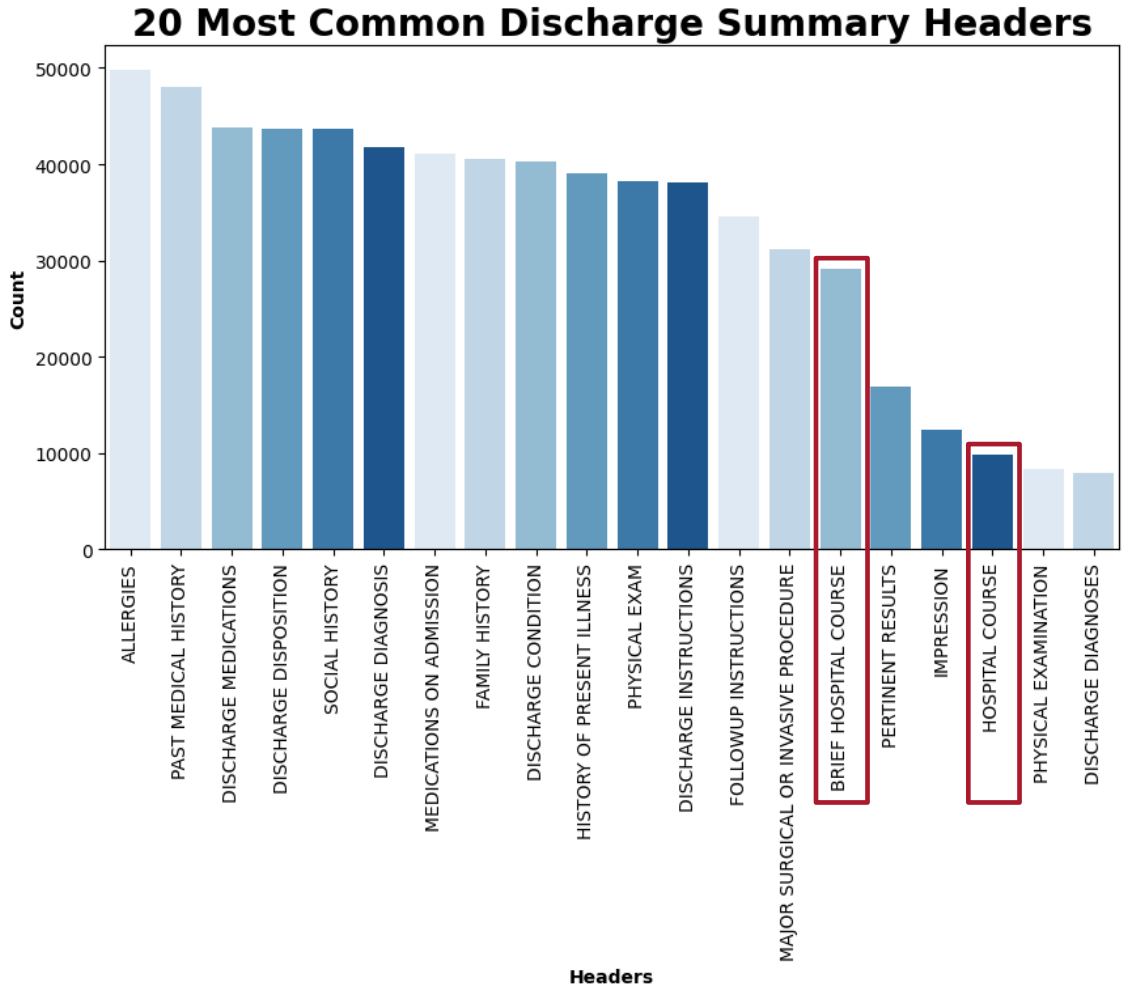
Approach

1. Target Text Identification
2. Initial Dataset Generation & EDA
3. Extractive vs. Abstractive Summarization
4. Extractive Summarization Method
5. Fine Tuning Pegasus and BART for Abstractive Summarization
6. Recursive Summarization
7. Measuring Performance

What are the top 20 headers/sections in discharge summaries?

- Headers give an indication of the context of discharge summaries
- Multiple headers can represent the same type of section
  - Brief Hospital Course and Hospital Course
- (Brief) Hospital Course sections tend to summarize all other patient notes

Hypothesis: Patients' notes can be used to auto-generate the (Brief) Hospital Course section of Discharge Summaries



Initial Dataset Generation

1. Notes were aggregated by hospital visits along with additional metadata
    1. category, description, charttime, subject_id
2. “Brief Hospital Course” & “Hospital Course” were identified and extracted from patient discharge summaries and used as target text
3. Processed both note texts and target text
    1. Removed obfuscated date & clinician name formats
    2. Removed successive underlines and newline characters
4. Dataset was split into train, validation, and test sets
    1. 70% (24,933), 15% (5,356), and 15% (5,356), respectively
5. Pushed dataset to Hugging Face Hub 🤗

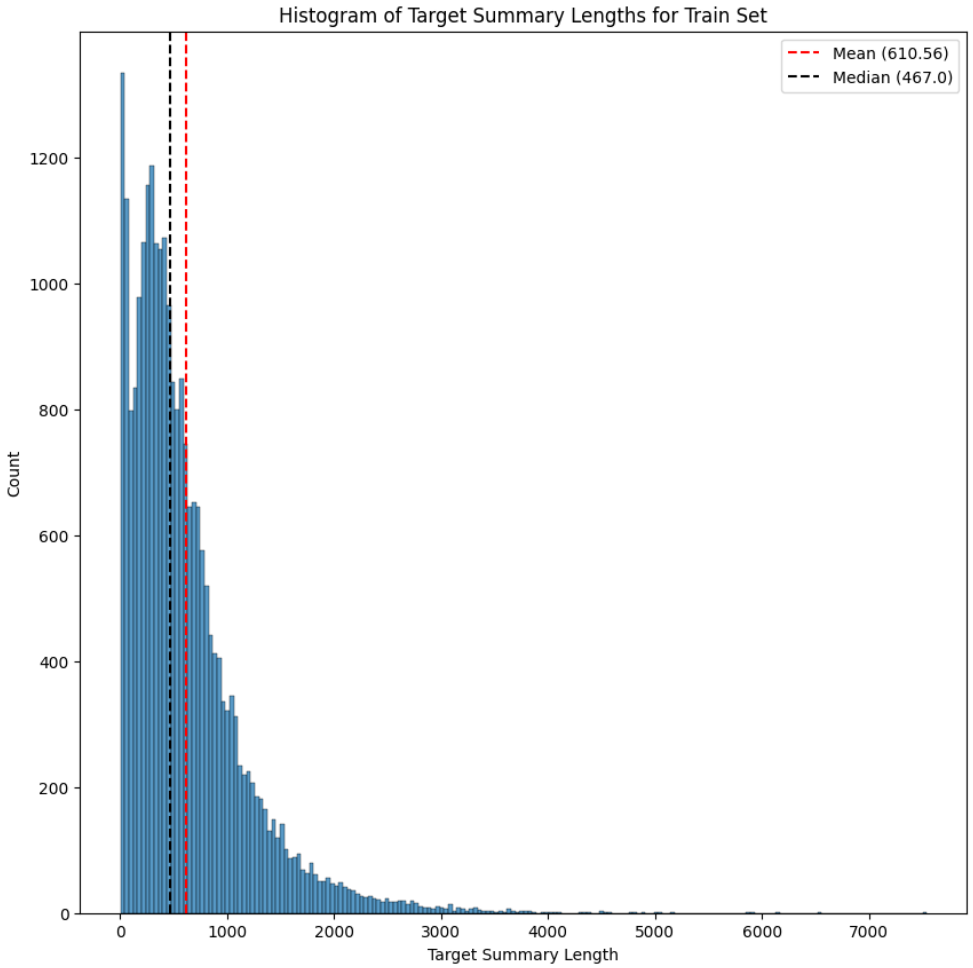
Initial Dataset Exclusions

- Exclusions:
  - Patients/hospital visits with multiple discharge summary reports
  - Discharge summary addendums (only the note was dropped)
  - Patients/hospital visits where the only note was a discharge summary report
  - Patients/hospital visits where the discharge summary report did not contain “Brief Hospital Course” & “Hospital Course” sections

Train Set Target Text EDA

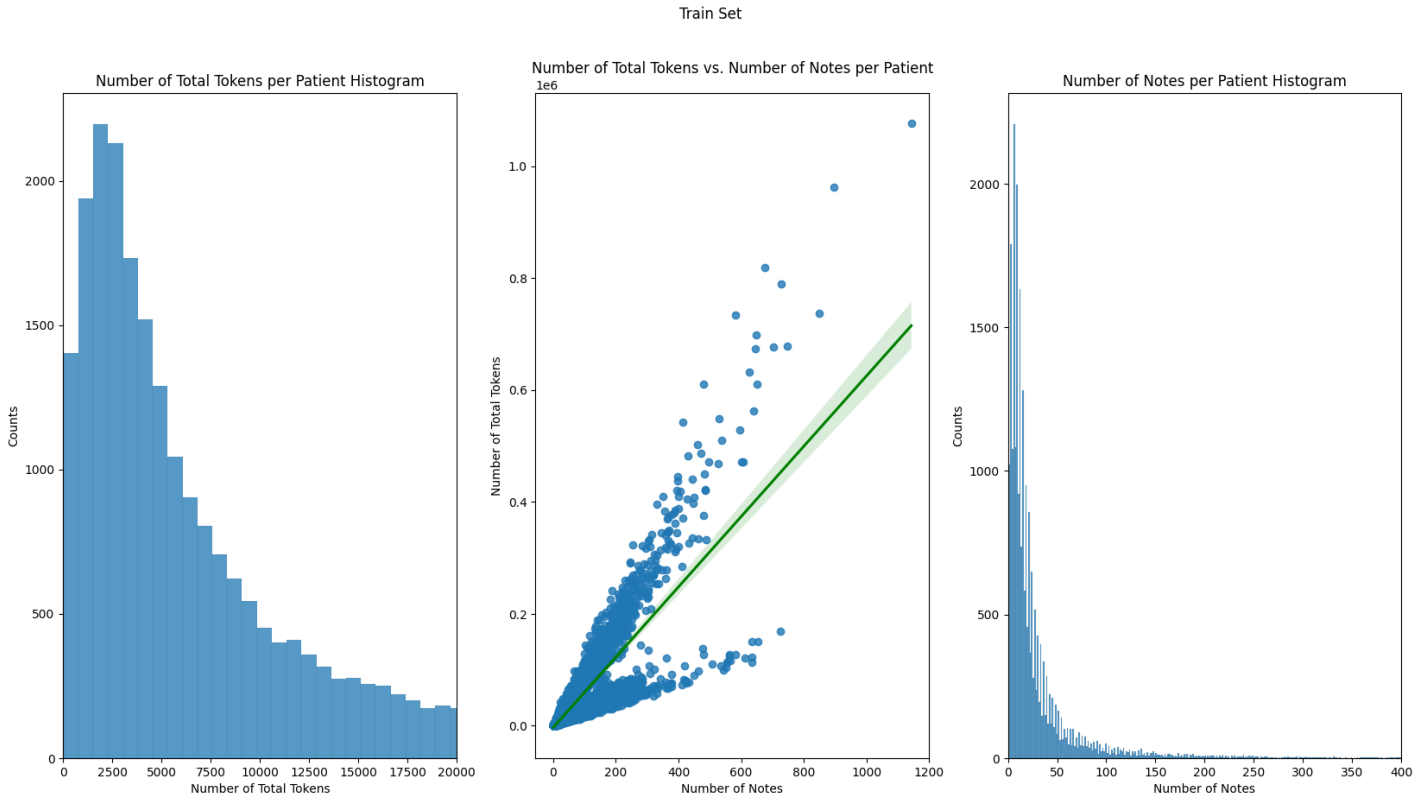
- Train set target text token lengths

- Provides insight to “max_length” parameter
  - Determines the token length of the generated summary



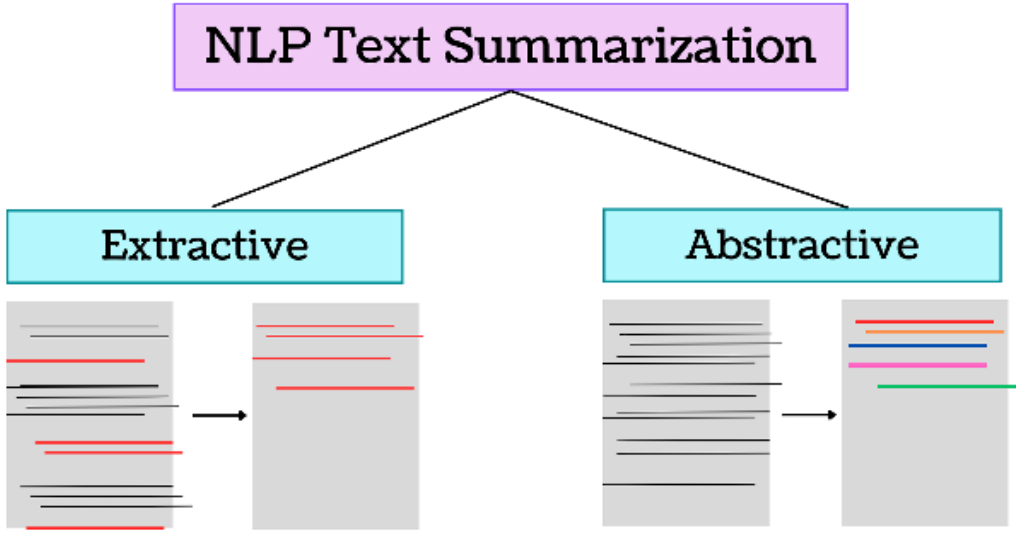
Train Set Note Text EDA

- Context size (max input tokens) for models of interest range from 512-1024 tokens
- Total token length for nearly all patients exceed 1024 tokens
  - How can we train a model with large context lengths?
    - We cannot: **Extractive summarization to cut down length, followed by abstractive summarization with LLMs.**



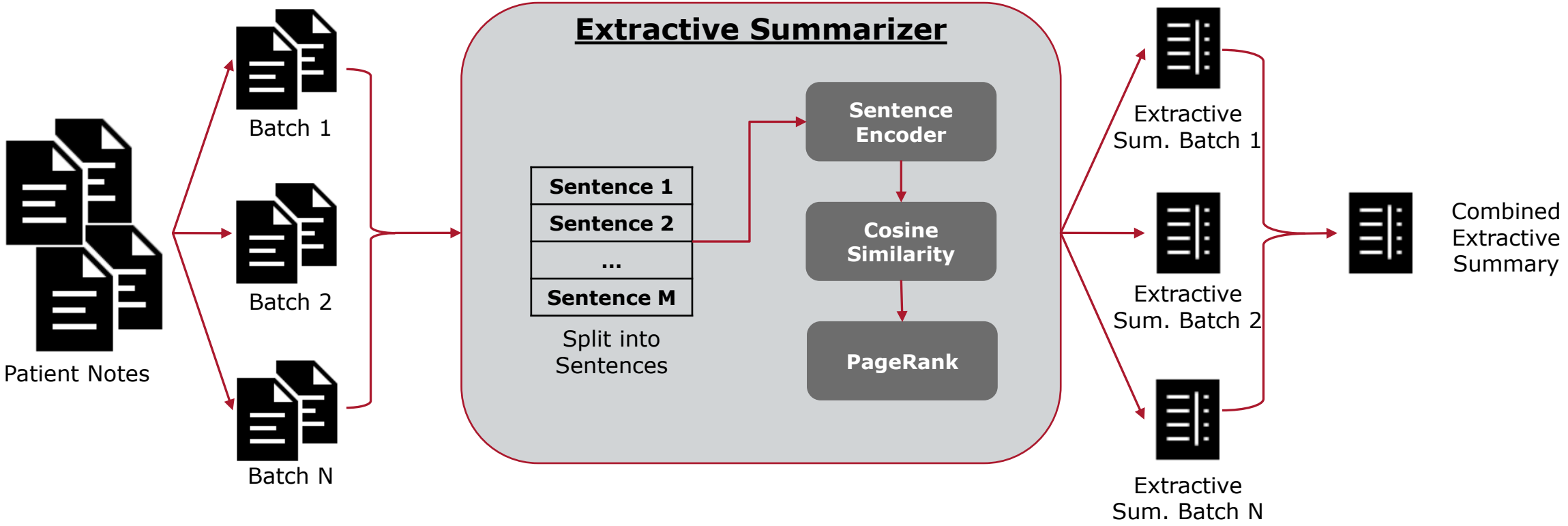
Extractive vs. Abstractive Summary

- **Extractive:** Selects the most important sentences from a text
  - Language is the exact same of selected sentences
  - Uses ranking algorithm to determine the top sentences
- **Abstractive:** Understands the text content and produces a summary based on this understanding
  - Language is different, but similar, to the text
  - i.e., asking ChatGPT to summarize a document



https://medium.com/analytics-vidhya/text-summarization-in-python-using-extractive-method-including-end-to-end-implementation-2688b3fd1c8c

Extractive Summarization Method

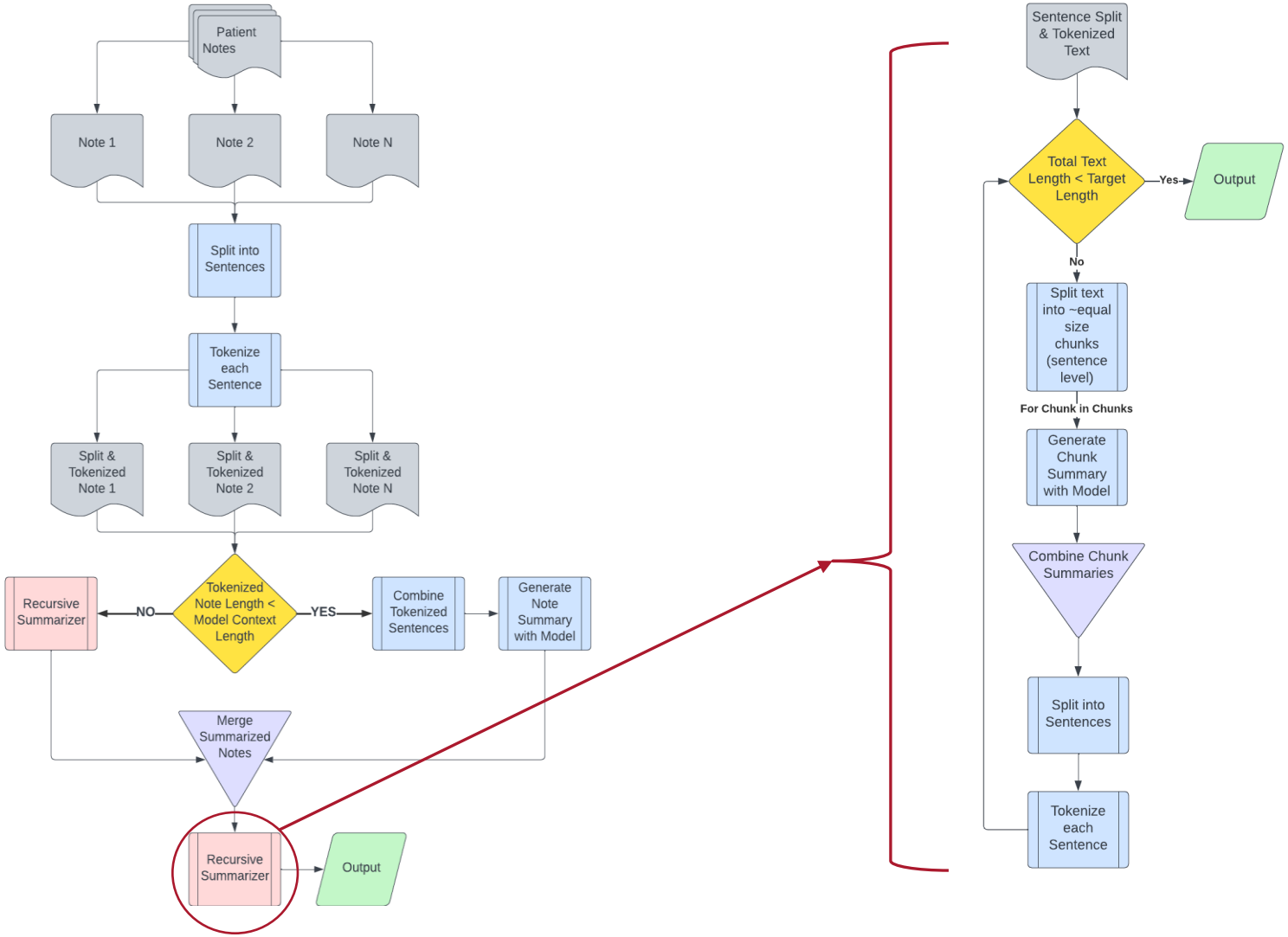


Training/Fine-Tuning with Extractive Summaries

- Pegasus-Large and BART-Large models were fine-tuned using the extractive summaries as the input text and the "Hospital Course" discharge summary sections as the target text
- Leveraging Jupyter Notebooks on Google Colab allowed for increased GPU RAM and decreased train time`

Model	N Epochs	Weight Decay	Gradient Accumulation Steps	GPU	Train Time (h:mm:ss)
Pegasus	3	0.01	16	V100 (16 GB GPU RAM)	5:54:29
BART	3	0.01	16	V100 (16 GB GPU RAM)	4:25:10

Recursive Summarization



Measuring Performance: ROUGE Score

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) Score

- Overlap of N-grams between reference and generated texts (Rouge1 & Rouge2)
- Longest common substring between reference and generated texts (RougeL)
  - RougeLSum at a summary level (splits by newline '\n')

Compare the generated text from the model to the target text

- ROUGE scores generated for the various models/methods on the test set

https://huggingface.co/spaces/evaluate-metric/rouge

Results

ROUGE Score Results

Method	Rouge1	Rouge2	RougeL	RougeLSum
Fine-Tuned Pegasus-Large on Extractive Summaries	0.004913	0.002295	0.003949	0.003925
Fine-Tuned BART-Large on Extractive Summaries	0.023803	0.011787	0.018235	0.018248
Base Pegasus-Large on Extractive Summaries	0.078553	0.009239	0.052366	0.052339
Base BART-Large on Extractive Summaries	0.093933	0.010894	0.058937	0.058920
Extractive Summaries	0.172416	0.023146	0.073196	0.073197

** Overall recursive summarization scores not captured due to significantly long (12+ days) run time to generate the summaries

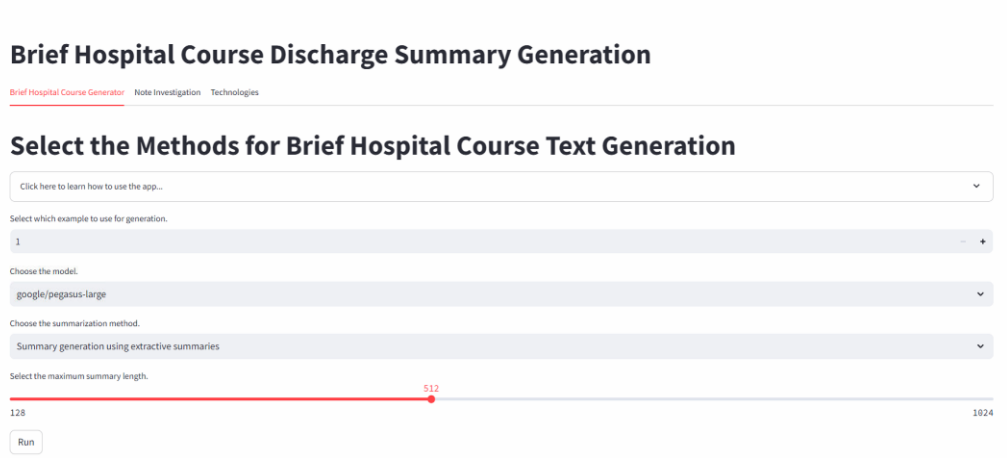
- Pure extractive summarization performs the best
- Fine-tuning the models has a negative impact on ROUGE scores
- All methods produce relatively low ROUGE scores
  - 0.5 is an excellent ROUGE score, whereas 0-0.35 is low

https://docs.oneai.com/docs/rouge-metrics-for-summary-headline#:~:text=Rouge%20metrics%20measure%20the%20word,50%20are%20considered%20excellent%20results.

Streamlit App

Streamlit App

- User can select:
  - Model
  - Method (Regular vs. Recursive)
  - Generation parameters
    - Maximum Length
    - Maximum Chunk Size
- Investigation of notes
  - Target text (Brief Hospital Course section)
  - Extractive summary text
  - Individual note text



Future Directions

Future Directions

- Extractive summarization with KMeans clustering (rather than PageRank)
- Create custom tokenizer and train from scratch using extractive summaries as input
  - Investigate a custom tokenizer due to the distinct nature of patient notes
- Recursive summarization with Louvain community detection algorithm
- Incorporate other tables (labs, vitals, medical history) in efforts to give more context to the model for better summaries

https://towardsdatascience.com/summarize-podcast-transcripts-and-long-texts-better-with-nlp-and-ai-e04c89d3b2cb

Resources

Resources

- MIMIC-III Dataset
- Custom Hugging Face 🤗 Dataset
- BART-Large Model
- Pegasus-Large Model
- GitHub Repo

WPI

Questions?