NLP Project Proposal – Text Summarization for Medical Documents

a. **What are you doing? Give us a one-sentence description of what you hope to do in your project.**

   i. Our project will be the evaluation of ML models for the task of text summarization, specifically the summarization and simplification of medical PDF documents for the benefit of older adults.

b. **Most importantly: what is your data? (identify specific data set(s) for this step, link to them and describe them in detail)**

   i. While we haven't narrowed down exactly which dataset would be best for our project, we have identified a few promising candidates below.

   ii. PMC-Patients – a first-of-its-kind dataset created in the last year of 167k patient summaries extracted from PubMed Central. Each entry consists of a patient's age, gender, medical summary, and relevant articles for the patient's health history.

   iii. MIMIC – health-related data from patients in critical care at Beth Israel Deaconess Medical Center. MIMIC contains several databases of medical data, but the most relevant ones for us are critical patient care data, hospital and critical care data, and emergency department data. All databases follow the format of listing a patient's age, length of stay, medical procedures, diagnoses, and patient care.

   iv. MeQSum – a dataset of 1,000 summarized consumer health questions consisting of a medical-related question and its corresponding summary.

c. **What tools will you be using?**

   i. Due to the nature of our project, most of our implementation can be done using popular Python libraries such as PyTorch, NumPy, and Matplotlib. We'll be referring to tutorials and demo models from HuggingFace as a reference for fine-tuning our model.

d. **What models will you be using?**

   i. We plan on using the GPT-2, BERT, T5, BART, and BigBirdPegasus language models for fine-tuning and evaluation. Because the models themselves are already publicly available, most of our work will involve fine-tuning them for medical data in particular and evaluating their performance pre- and post- fine-tuning.

   ii. For a baseline comparison, we can implement a N-gram language model to compare against other models.

   iii. While we won't be able to implement fine-tuning or training of larger/more advanced LLMs such as GPT-4 & Sonnet 3.5, it may be helpful/possible to compare performance against these models as well.

e. **Any other resources that you'll need to use? Components you'll need to implement?**

        i. Previous answers contain all resources we intend to use, however we expect to encounter challenges along the way that will necessitate the use of further resources that we will seek out on a case-by-case basis.

f. What visualizations/results/etc will you be producing?

   i. We will produce results based on each candidate model, evaluated on how well they summarize text from a given medical document corpus. This evaluation will be based on the accuracy and simplicity of the summary using metrics such as ROUGE and BLEU scores, and the Flesch-Kincaid Grade Level or other measurements of text complexity.

   ii. We will hopefully be able to produce some useful and intuitive visualizations that graph these results, using libraries such as matplotlib.

g. What are preliminary sources/tutorials/etc that might be helpful?

   i. We can look into videos and existing literature that detail how to evaluate the performance of a model at summarization. Additionally we can look into different ways of tuning models to see what approach works best for our intended results. HuggingFace has several publicly available tutorials for fine-tuning and model evaluation.

h. What is your timeline/working plan?

   i. Our current working plan is to meet once a week to touch base and get some work done as a group, and in that time communicate what we need to individually get done between meetings, while discussing any further topics in our group chat when not in person. We plan to meet with the TAs the week after submitting our project proposal to refine our final idea and get started working on the project. Beyond that, we intend to make steady progress each week and evaluate our next goals on a week-by-week basis to ensure our scope remains realistic and everyone has a proportional workload.

i. What do you aim to have completed by the project check-in due date?

   i. A specific goal, if our proposal is to continue as planned after we meet with the TAs, would be to identify the data source(s) that we will be using, and format it for model fine tuning.

   ii. We'd also like to get our basic codebase and repository structure set up, allowing us to visualize/understand/play around with this data before getting into more complex tasks.

j. Who in your group will be in charge of which component?

   i. We will work on preprocessing data and some preliminary testing with models together. We will then assign each group member a different model to fine-tune/test with: GPT-2, BERT, T5, BART, BigBirdPegasus.