Capstone Three: Project Proposal

For my final project, I want to use medical text data to perform sentiment analysis, classify the type of diagnosis, and communicate with AI chatbot APIs to act as an AI doctor. Furthermore, in the sentiment analysis, we will perform tasks such as predicting the next word and classifying the "tone" or "severity" of the diagnosis.

- Problem statement
    - Use the provided medical text data to perform sentiment analysis to draw conclusions such as the type of diagnosis and severity. Also, we will communicate with AI chatbot APIs to create additional labels for the data. The provided medical training text includes the category of diagnosis (5-class). Therefore, we can use these labels to assess the accuracy of the APIs before generating additional labels for the training and testing data. Lastly, we will analyze the text data to predict the next word using a LSTM (long short-term memory) network. This will display the "intelligence" of the model and emphasize the capabilities of an AI doctor.
- Context
    - AI chatbots are very popular right now with the recent contributions of ChatGPT along with other platforms. Members of the general public are beginning to see some of the capabilities of AI by using these platforms which makes this a hot topic. Furthermore, using this technology for medical diagnosis is at the forefront of research for many companies as asking the internet medical questions (e.g. WebMD) is nothing new for many people. Therefore, enhancing chatbot capabilities for medical use is going to be very important moving forward as it can be a good *supplemental* tool in addition to typical medical attention. In this project, with the tools and realistic constraints we have, are going to test these principles and model a simple "AI doctor".
- Criteria for success
    - The criteria for success will be twofold: (1) the classifications made by the APIs are reasonable and consistent and (2) the predictions of our "AI doctor" achieve performance better than guessing.
- Scope of solution space
    - N/A
- Constraints
    - Performance is going to be subjective. For the API performance, we can assess the training set's diagnosis type label accuracy. However, this will be the only API classification that we will be able to assess. All other API classifications will be treated as *true labels.*
    - Due to the complexity of building a model such as an AI doctor and the data we have, this project will be used more as a proof of concept. This will not be a production-ready model.
- Stakeholders
    - N/A

- Data sources
    - https://www.kaggle.com/datasets/chaitanyakck/medical-text?select=train.dat
    - The medical text data here contains 14438 training records and 14442 testing records. The data is purely text-based with the training data containing diagnosis class labels. The five class labels are 1. Neoplasms, 2. digestive system diseases, 3. nervous system diseases, 4. cardiovascular diseases, and 5. general pathological conditions. There is no other information provided about the individual patients. For this reason, it is going to be difficult to add additional features to the data.