Project Proposal

Task description: The aim of this project is to solve two problems as mentioned below.

Problem #1
The goal is to categorize/classify student responses to the following question:

"What one action can students in your lab groups take to improve your educational experience at UW?"

Categories have already been established within a community of learners of framework as:
Category 1: Individual Accountability(IA)
Category 2: Positive Interdependence(POSI)
Category 3: Promotive Interactions(PI)
Category 4: Interpersonal and Social Skills(ISS)
Category 5: Group Processing(GP)
Note: no student responses were classified as Category 5

The purpose of this section of the NLP project is to classify student responses into the five categories above for maximum accuracy, where accuracy is defined as a classification that is identical to what the researcher determined in initial coding and analysis.

Problem #2
The goal is to identify the best number of clusters that responses to the following question organize into using NLP methods:

"What one action can faculty take to improve your educational experience at UW?"

No assumptions are made about how many clusters (groups) these responses will fall into. The goal of this portion of the NLP project is to identify the optimal number of clusters to support future coding of these responses. This will be accomplished by representing the students' responses into three categories:: topic, sentiment analysis, and semantic similarity.

Data description: This project includes 1329 data samples (survey responses) for each problem #1 and problem #2. Data used represents the student population from 19 courses junior and sophomore level classes in electrical and mechanical engineering at a large public research institution reported existing and preferred levels of peer and faculty support within in-person and an online learning environment.

- ***Input (X):*** Student survey (text) response to the questions.
- ***Output (Y):*** Classify the responses by labels (IA, PI, PIOS, ISS)
- ***Data source information:*** IRB (Internal Review Board) approval (provide application/study number here) was obtained to recruit and survey undergraduate students for this study. All participation was voluntary, and students were informed that their survey responses would remain confidential. (Data is provided by my supervisor Professor Denise Wilson)

Note: Expected to get more data for testing. I will be adding that information on the main report.

Proposed approach outline: I will be creating a multi-label text classification model that analyzes textual feedback and predicts one or multiple labels associated with it. I will start by splitting the data into training, and test sets. Next step will be to choose an nlp model which can predict labels with better results than a baseline for problem #1. As a sanity check, I will also be regularizing the model and will tune the hyperparameters for improved performance. In problem #2, I will use the Latent Dirichlet Allocation method for topic modeling, the RNN method for sentiment analysis, and the BERT method for

semantic similarity. The approaches listed for each category are tentative, as I will be exploring further options as I learn more by solving problems one at a time.

Baseline solution: As a baseline, I will be using majority class label like Naive Bayes or Latent Dirichlet Allocation and as my main approach I will be using Bidirectional Encoder Representation from Transformers (BERT) which is the *deeply bidirectional*, *unsupervised* language representation.

Evaluation criteria: My evaluation criteria will be to use K-fold cross validation and calculate model accuracy. I will also plot bar plots that show the total feedback (PI, IA, PIOS, ISS) counts for different labels.

Software: I will be using Jupyter Notebook to implement the model for this project using python language.