Presentation 2 Instructions

The following metrics will be used to assess your progress in your project presentation 2. Create a presentation slide and a recording of the presentation slide. Put both in D2L under your group.

1. The completion of your data collection

You have added 1000 new records to the existing benchmark that you used. Show an overview of the overall data three ways: distribution of labels (i.e., target variable) in the original benchmark, in the new 1000 records, and in the extended benchmark (i.e., original+1000)

2. The quality of your data labeling.

To demonstrate the labeling of your data, you can report the agreement between the data labelers. You can use two metrics: percent agreement and Cohen kappa. You should also explain how you resolved the disagreements, e.g., for a give record two labelers have two labels. How did you pick the correct label? Give examples in your presentation.

3. The quality of data preprocessing

How you preprocess data to ensure that noises are removed or any unnecessary information that is not needed for ML is filtered out.

4. The quality of features you generate to develop the machine learning models

For example, a straightforward feature for natural language text is bag of words. Show examples of such features in your presentation. Also show what additional features you can pick and think of.

5. An overview (theoretical is fine) of how a machine learning model can use the features

For example, if you are using bag of words as features, you can show a pipeline on how such features are going to be used in your model to train as well as to test (do you need to preprocess to words like using TF-IDF, standard scaler? what else do you need for the features to be usable in your model?)

Each metric carries similar weight.