COMPSCI 4NL3: Competition Launch

Winter 2025
Due: March 13th

1 Overview

In this part of the project, you will analyze the annotations that your classmates completed for your data, compute agreement, determine ground truth labels, and build baseline models. The deliverables will be your report, a single slide advertising your project, and a codabench page. Setting up codabench will likely be the most time consuming step, so make sure to budget time for that.

2 Annotation Analysis

If you have not already received them, collect the annotations from your classmates for your dataset. Compile the data into a single file containing all data points and labels. Next, you will need to compute agreement. Agreement metrics help assess the consistency of annotations. Some commonly used metrics include:

1. Cohen's Kappa: For pairwise agreement between two annotators
2. Fleiss' Kappa: For agreement among multiple annotators
3. Krippendorff's Alpha: Useful when not all annotators annotated all data
4. Percentage Agreement: Simpler measure that does not account for chance agreement

These metrics will help you determine if your annotations are reliable or if there are significant disagreements among annotators. Read more about each to see which one is most suitable for your dataset and compute that metric. You may use existing libraries to do this for you. The one that is the most appropriate for most of you will be Cohen's Kappa unless you decided to annotate more of your own data as a team.

Next, decide on the ground truth labels for your dataset. These will be the final labels used for training and evaluating your models. Some possible methods of determining ground truth include:

1. Majority Vote: The most common label is taken as the ground truth. This requires having more than 2 annotations per data point. If there are only 2 annotators and you need a third vote, you could label some of the data points yourself to provide the tiebreaker.
2. Weighted Voting: Assign weights to annotators based on experience or reliability.
3. Adjudication: Manually review data points where annotators disagree and discuss between the annotators until a resolution is reached on the correct label.

You will need to show how you computed the ground truth labels and explain why you chose that method. Next, you should analyze the distribution of ground truth labels. Create a visualization to show the amount of data with each label. Are the classes balanced or imbalanced?

3 Codabench

Create a Codabench page to benchmark models for your task. This platform will allow you and others to submit models and compare their performance. Steps include:

1. Read the Guide: Learn how Codabench works

2. Task Definition: Describe your task and provide a clear description of the dataset.

3. Split Data: Choose a training, validation, and test split for your dataset.

4. Data Upload: Upload your dataset for participants to download or use via an API.

5. Evaluation Metric: Specify which metrics (e.g., accuracy, F1-score) will be used to evaluate submissions.

6. Baseline Model: Set up a baseline model that participants can use as a reference point.

7. Configuration: Configure the task in the way that you see fit. You do not need to use the compute queue, you can allow participants to directly upload a submission file with their model outputs and you compute the score.

3.1 Create Baselines

Develop baseline models to compare against more advanced models. This will include:

1. Simple Baselines: These are simple models that provide a baseline performance. There are a few options, and you should choose whichever of these two gives the best performance: (1) Random Baseline: Randomly assign labels to the data. (2) Majority Baseline: Always predicts the most common label in the dataset.

2. Trained Model: Train at least one model (e.g. Logistic Regression, Random Forest, Feedforward Neural Network) on your dataset. Evaluate it's performance using a validation set and compare it to your simple baselines.

These will set the lower bound for performance on your task. Make sure to include these with the initial task launch on Codabench.

4 Slide Advertisement

You should make one slide with Google slides to advertise your task to the class. This should serve as a quick description of your task. You may want to include an example and show what data you have and what the model should predict. Why should others work on your task? How does your Codabench page work? What is the baseline performance? We don't have time to have everyone give a presentation. I will go through the slides in class and on the day they are presented I will ask if anyone in class would like to present their own slide.

5 Deliverables

You should submit a report explaining how you computed agreement, why you chose that metric, how you decided on ground truth labels, what baselines you chose, a link to your codabench page and a link to your Google slide (make sure it is viewable to anyone who has the link).