COMPSCI 4NL3:
Project Proposal

Winter 2026
Due: January 20th

1 Overview

For the course project, you will gain experience with the full natural language processing life cycle from a task definition to a fully functional application. Note that you ARE NOT doing all of this right now. This is just to let you know what is coming later. You will work in teams of 3-4 people to complete the following steps throughout the semester:

1. **Define a task:** You should select a task that is interesting and challenging. Next, define the input data type and evaluation metrics.
2. **Collect unlabeled data:** Collect data that fits within the constraints of your task. It should be possible for you to determine the label by reading the data. You may use data that has been previously labeled if you wish. If you do not, it may take considerably more time for your group to complete your corpus.
3. **Develop annotation guidelines:** Write the set of rules that another person could use to label your data, and follow these rules to label a portion of your data. You may choose to set up a simple command line interface to annotate the data.
4. **Compute agreement and finalize data:** Check the inter-annotator agreement and resolve disagreements if necessary. Compute and basic statistics about your dataset and make a post about it on the teams channel.
5. **Create benchmark:** Train basic models to set the expectations for performance on the task and create a Codabench page.
6. **Build models:** *Individually* submit models to 3 other teams' competitions and try to beat the baseline approaches.

2 Proposal

The first milestone is to complete your project proposal. The format of the proposal will be a PDF document which includes the following items:

1. Team members **(3-4 people required)**
2. Task title and overview, including the significance and what makes it challenging. You may not work on sentiment analysis. You should choose a task where you believe you and your team can annotate the data with reasonable agreement. This means it does not have to be super high, but should be significantly better than chance.
3. Task definition (type of data, classification or regression, number of classes, single label or multi-label)
4. Data source(s) and plan for data collection. This may include how you are going to scrape the data and follow terms-of-service, API access and handling rate limiting, using open-source data, or any other relevant details. If assigning labels by hand, how long does it take per instance? Include links to the data if relevant. If you download your dataset from Kaggle, you must include the Kaggle link and the original data source link from where the dataset was downloaded and posted on Kaggle (I recommend against using Kaggle datasets, as they often are not well documented). Include any meta-data available for your corpus. If you are not collecting it yourself, include details about how the data was collected, annotated,

preprocessed, etc. Do not work on the datasets for which no source information is available. Please indicate whether you will use a small subset of the data or features for your project or the entire dataset, and why.

5. Expected size of the dataset (number of data points) and 3 example data points with labels assigned by your group. You should ideally have enough data for several hours of annotation. This corresponds to a few thousand data points. Some projects that require more annotation time per instance may have fewer.
6. A **team contract signed by all team members**. This is a description of the expectations and/or roles that team members agree upon for the semester. Part of your final project grade will be based on whether your teammates agree that you followed this contract. See next section for details.

3 Team Contract

The Team Contract[1] specifies the overall purpose of the team, the responsibilities, and the ground rules or norms that team members agree to follow. Your contract should include:

1. What is our team's purpose or mission?
2. What are the duties/roles of each team member? What is expected of each team member?
3. How will the team handle the leadership/facilitation/management activities?
4. Anything else you think will be helpful to set the groups expectations.

The team should describe what is expected from each team member and each member should sign the contract **(print name, sign, and date)**. Your contract should not say "Person A will do all the work". It should focus more on how you will work together rather than who will do each individual task. You won't know all your individual tasks yet. When discussing how to handle conflict, you should write something about how you will meet, discuss, and resolve the conflict. You should not write "Person A will be shunned and banned from group activities". You will need to work together for this project.

4 Deliverables

You should submit your written proposal (PDF) on A2L

5 Requirements

You will be assessed based on the completion of the following requirements. Your proposal should contain all the required items with sufficient detail.

1. (30%) Data collection sources and plan (item 4). This should address all relevant details for your dataset and provide examples. This mainly focuses on describing your dataset, how it has been processed, where it came from, and how you plan to further process it.
2. (30%) Team contract (item 6). A brief contract should include the expectations and roles agreed upon. Check the footnotes for examples of what your contract might look like.
3. (30%) Team and task details (items 1, 2, 3, 5). This should include your task overview, definition, team members, and expected size of the dataset, with some estimation of how long it will take to annotate.
4. (10%) Document is well-organized and professionally presented.

[1] Borrowed ideas from `https://crlte.engin.umich.edu/wp-content/uploads/sites/5/2020/05/Team-Contract.pdf` and `https://uwaterloo.ca/centre-for-teaching-excellence/catalogs/tip-sheets/making-group-contracts` and you may also look at `https://cns.utexas.edu/sites/default/files/uploads/documents/2023-03/Examplegroupcontract.pdf` for ideas.

6 Next Steps

The next project step will be released at a later date. In the meantime, you may want to start collecting your data. Data collection (e.g. scraping) can sometimes be time consuming and it is best to start earlier if possible.