1. Overview: Annotation and Modeling Experiment

Using a **sample of 1000 comments**, you must conduct an annotation task from start to finish:

1. Task design: Write up annotator guidelines/instructions.
2. On your group's dataset, collect annotations from **two other groups** for your tasks. In order to get everything done in time, note you need to tackle all these steps:

• As a group, find your annotators.

• Group sends annotation guidelines and spreadsheets to annotators. Make the spreadsheet easy to use.

• Annotators hand off their completed spreadsheets back to researcher.

• Researcher double checks basic data cleanliness (are all examples annotated? are all labels allowable ones?) and submits.

1. Collect feedback from annotators about the task, including annotation time and obstacles encountered.

• **IMPORTANT**: You should do a few test runs where you go back and forth with annotators to ensure adequate agreement before having them annotate everything.

• For this, it is recommended that you first get ~50 comments annotated. After that, gather initial comments to fine-tune the guidelines based on annotator feedback and initial agreement scores. Talk to the annotators and discuss issues they had. Then, have the 950 comments annotated

1. Calculate the final inter-annotator agreement and other related statistics. Manually review disagreeing annotations for adjudication to create the gold standard dataset.
2. Develop machine learning models using your dataset.

Many more details on these steps are spelled out in the next section.

Roles: Every **group** in the class will take on two roles: **researchers** and **annotators**. As a **researcher**, you perform the above steps in your comments. At the same time, your group will also serve as **annotators** for your fellow students.

Tasks: Every researcher will have the same set of comments, and will define their *own* unique annotator guidelines/instructions. Therefore, the researcher will have to draw up guidelines and explanations to allow their annotators to do a good job at reliably identifying whatever categories the researcher has a goal of collecting.

Annotator etiquette: Everyone must collect annotations from at least two groups. Therefore, everyone should be ready to annotate. If someone asks you to annotate for them, please accept, unless you have already accepted more than two annotation jobs! You should try to do a good job of annotating for your fellow students.

It's OK if this is hard: Designing annotation task guidelines and collecting annotations is tricky. This assignment is small-scale as far as annotation projects go, but hopefully, it illustrates the challenges that underlie it.

Datasets. The City of San Antonio is interested in understanding the communities' thoughts about new technology that can positively or negatively impact them. This work builds on a prior collaboration where we developed machine learning models to analyze text inputs from the city's surveys about technological innovation that were posted here https://www.saspeakup.com/N4568. We want to continue that work to gain insights at scale using social media. For this class, we will focus on Reddit data from reddit.com/r/sanantonio. We will use the models/data we gather to gain insights that can be shared with the city.

For the format, the text of one comment is on each line, along with the comment the question is associated with. The original comments sometimes had tab or newline characters; in our preprocessing, we normalized all those to spaces. We did no other preprocessing. The text is encoded as UTF-8. Your web browser may not view it correctly; to view, we recommend using a smart text editor that lets you explicitly set the encoding (e.g., VS Code, Vim, Emacs, etc.), and when you view it on the terminal, be aware your terminal software may not handle it either. In Python 3, you should be able to read the data correctly using the encoding='utf8' option for open() (see online docs).

2. Details: Annotations

2.1 Task Design

Conceptualize and design a classification task for the comments. Take a look at the dataset. Write **annotation guidelines** for your task. You will give your annotators these instructions so they can read them before annotating. It should be a written document, at least a page long, for each task (or more if necessary). The guidelines should include:

• A list of the class under consideration, including the exact string you want them to use when typing into a spreadsheet.

• Descriptions of the category and what it means.

• Example comments that are illustrative of the categories. **Do not use examples copied directly from your dataset.** We specify this so that you can't make the task too easy. Please make up synthetic examples or use examples drawn from another dataset.

• A discussion of tricky corner cases and criteria to help the annotator decide on them. If you look at the data and think about how annotators could do the task, you will realize a bunch of such issues!

You must provide guidelines for two classification tasks. **The first task** that must be annotated as "Technology-related" or "Not Technology-related."

Next, **for the second task**, you should explore something that **your group comes up with**. The goal is to analyze what the underlying text is about. This could be general topics or something more implicit (e.g., sarcasm, emotion, etc.). I recommend meeting with me to discuss your ideas. I list a few examples below, but this can be anything:

- Transportation systems, vehicles, and autonomy (mobility)
- Data governance and city data platforms and dashboards (access to government and data)
- Wireless communications and broadband applications
- Cybersecurity and privacy for public and private sectors
- Public safety and security
- Community resilience, adaptability, and sustainability (Environment, sustainability & resilience)
- Public utilities for energy, water, and waste management
- Agriculture and rural productivity and quality of life
- Smart building technologies and IoT applications
- Education and workforce development
- Community well-being: Diversity, Equity, Integrity, and Trust (DEI&T)
- Housing
- Sarcasm
- Political vs. not political

Your guidelines must provide a description of your task. Moreover, they must provide examples (that should not appear in your dataset for annotation but should be similar to examples that may appear).

Overall, you are in charge of defining your guidelines! This process of boiling it down to something specific, actionable, and thus measurable is called *operationalization*. Everyone will be defining slightly different guidelines, so feel free to make yours specific or unique in some way. But you will want to make it clear and as straightforward as possible for your annotators to do the task.

Deliverable: Include a copy of your annotation guidelines in your writeup. Due **April 5th.**

2.2 Annotation collection

It's time to collect annotations! Find two "volunteers" from class to be your annotators. Please use Slack to find help! Every group must annotate for someone else if asked to make this easy! :)

It would be best if you first did a pilot study with the groups performing your annotations. I recommend a small set of 50 examples. You can then calculate the agreement and decide if you need to:

1. Meet with the groups to train them.

2. Modify your guidelines.
3. Make changes to what you are annotating.

Please collect your annotations through an Excel document with three columns for both of your tasks:

1. Comment text
2. Label from annotator
3. Notes from annotator

You should create two identical spreadsheets with just the text column, then send each, along with the guidelines, to each annotator. **Annotators should annotate independently of you and of each other.**

DELIVERABLE: Include two CSV files in your **April 19th** submission, one for each annotator.

IMPORTANT: I recommend you schedule time to meet with me as a group to review initial annotations to ensure they are high quality and discuss potential pain points.

3. Details: Modeling

Now, you must develop models using the final gold standard annotations you received for your tasks. There is one major requirement for the model development. You must try at least three features (e.g., number of capital words, number of exclamation points, lexicon features). PS3 will be related to the final project; pay special attention to it to help you understand how to accomplish the final project. I recommend splitting the data into a training, validation, and test dataset. You can also use cross-validation instead of a validation split. The Macro and Micro Precision, Recall, and F1 scores must be reported on your validation dataset and test set in your final report for each set of features you use.

DELIVERABLE: in your **May 8th** submission, include one CSV file containing three columns based on your test set: the comment text, the gold label, and the predictions of your best model. You must also supply code to reproduce your best model.

FINAL DELIVERABLE: Create a small technical report describing your method (e.g., what features you used and models you explored) along with evaluation metrics on a development dataset and a comprehensive error analysis. The writeup should be two (min) to four (max) pages long. This is due **May 8th.**