Annotation Plan for Sprint 3

1. Purpose

- We aim to annotate a dataset of ~ 1,600 LinkedIn posts to facilitate potential downstream NLP tasks with said data, such as content recommendation, automated classification, and entity extraction.
- Data will be annotated according to its topic/category.

2. Tools and Methods

- **Data Source:**
  Our dataset consists of LinkedIn influencer posts collected from a publicly available Kaggle dataset. These posts have been pre-filtered based on relevance to our study.
- **Data Preparation:**
  Prior to annotation, the data has been cleaned to remove duplicate entries, as well as filtering out posts with limited textual information.
- **Annotation tool:n:**
  Excel will be used to annotate each example. Each member will fill a separate excel document with their annotations/categorizations for each example, then each of these will be collated later for analysis.

3. Annotators

- At this time we intend for all of the data to be annotated manually entirely by the experts conducting this project (Kartik, Muhammad, Zhengyi, Tim).
- In cases where it is applicable, majority vote will be applied to settle disagreements.
- In cases where there is a tie, ChatGPT will be used to decide the final label.

5. Annotator Overlap Plan

- **Assignment Strategy:**
  - Each of the 4 team members will be assigned an equal portion of the corpus for annotation, roughly ¾ (1200 documents) of the total corpus (to allow for overlap between 3 team members, to avoid ties).
  - Each document will be randomly assigned with 3 annotators.
- **Overlap for IAA:**

    - We intend to randomly assign annotators to each example; this means we cannot tabulate an exact percentage of examples on which annotators will overlap, but can expect overlap on a significant proportion of the corpus.

6. Annotation Schema

- A schema for annotation is provided here: https://github.ubc.ca/zyshan/COLX523_linkedin_corpus/blob/main/documents/Annotation_Schema.pdf

7. Data Volume and Timeline

- **Expected Output:**
    - By the next sprint deadline (Saturday, March 9th, 2025), we intend to have ~1,600 annotated examples prepared for use in our project.
- **Resource Considerations:**
    - Expected time commitment per team member is roughly 3-5 hours during the week of March 2nd - March 9th.
    - For those examples on which ChatGPT is to be used as an annotator disagreement tiebreaker, Tim Christilaw will be responsible for its application and will utilize GPT-4o to balance time efficiency and model performance.

8. Pilot Study

- The team has already met to discuss annotation strategy and refine our final list of topics/categories as outlined in the schema, according to a small pilot study conducted via Zoom with a subset of the data.
- The Pilot Study unveiled a few difficulties with our original annotation plan, namely that one of the categories was too broad and general, while also being relatively uninformative about the actual post content, leading to a large majority of our annotated practice examples sharing this category label. This was addressed by splitting the problematic category into several more focused categories.