Project Proposal: Annotation of LinkedIn Posts for Categorization and NLP Applications

1. Project Overview

This project aims to annotate an English dataset of approximately 1500 LinkedIn posts with predefined category labels to enable topic modeling and slot filling applications. The goal is to create a structured corpus that captures the semantic and functional aspects of LinkedIn posts.

By systematically categorizing posts, we can facilitate downstream NLP tasks, such as content recommendation, automated classification, and entity extraction.

2. Preliminary Annotation Scheme

We propose annotating each post with one or more of the following three categories (these are preliminary ideas and may be expanded upon in the final product):

1. **Professional Growth**: Includes achievements, job advancements, networking advice, and hiring posts.
2. **Industry & Culture**: Covers trends, lessons, company culture, diversity, and inclusion.
3. **Engagement & Learning**: Focuses on interactive content, educational resources, and humor.

Each post will be labeled with at least one category but can have multiple, if applicable.

3. Acquiring the Corpus

In the interest of time and efficiency, as well as copyright issues (LinkedIn doesn't allow scraping), we opted *not* to manually scrape data or extract it using an API, because a high-quality dataset complete with additional metadata was available and downloaded at the following link: https://www.kaggle.com/datasets/shreyasajal/linkedin-influencers-data
We intend to randomly sample from this dataset to pare it down to ~ 1500 examples. We will store the original and annotated data in csv file format in the data folder of the project repository.

4. Enhancing the Corpus

To improve the usefulness of the dataset in downstream NLP or ML tasks, we propose incorporating one or more of the following additional metadata, which our dataset gives us access to:

- **Engagement Metrics** (likes, comments, shares) – Useful for understanding the impact of different categories.
- **Post Length** (word count) – Helps analyze how verbosity affects engagement.
- **Hashtags & Mentions** – Can provide insights into topic modeling and named entity recognition.
- **Time of Posting** – Could reveal trends about when different types of content perform best.

5. Sample Data and Annotation Process

To ensure high-quality annotations, we will:

- Develop a structured schema with guidelines to categorize each post.
- Use a double-annotation approach, where each post is labeled by at least two annotators for consistency.
- Resolve disagreements through discussion and majority voting.
- Re-visit our annotation schema if inter-rater agreement is < 80%.

6. Solvable Task & Hypothesis

This annotation task is designed to be **non-trivial yet feasible**, focusing on **entire texts** rather than token-level or syntactic annotations. We believe the project may help to answer the following questions in follow-up NLP tasks:

- *Can LinkedIn posts be automatically classified into meaningful categories with high accuracy?*
- *Which post categories receive the highest engagement?*
- *How do lexical distributions differ among our categories?*

7. Expected Applications

- **Topic Modeling**: Understanding emerging themes in professional social media.
- **Slot Filling/NER**: Extracting key entities (e.g., job titles, company names, product types/names, event details).
- **Information Retrieval:** finding posts based on relevance to a user's topic-related query
- **Recommendation Systems:** generating a list of similar posts based on a post selected by the user.