NLU CSCI 4907/6515 - Spring 2026

Overview

The goal of the project proposal is to demonstrate that you have a clearly defined task, a viable data source, and a technical roadmap. Your proposal should be concise ~3 pages (more is okay) and follow the structure below. You are free to use the platform you are most comfortable with (e.g. google doc, latex, …)

Deadline March 16th, 11:59PM EST

---

1. Introduction

- **Motivation:** Why is this problem important?
- **Task Definition:** Formally define the NLU task. Is it classification, sequence-to-sequence generation, or structured parsing?
- **Contributions:** What is the specific "added value" of your project? Are you comparing two architectures, creating a new dataset, or applying an existing model to a new domain (like poetry)?

2. Related Work

- **Literature Review:** Briefly summarize 3–5 key papers that have addressed this or similar problems.
- **Gap Analysis:** How does your project differ from or build upon these works?
- **References:** Use a standard citation format (e.g., APA[1])

3. Data Strategy (Added Requirement)

- **Source:** Where is the data coming from? (e.g., Hugging Face Datasets, Twitter API, Project Gutenberg).
- **Preprocessing:** Mention specific NLU preprocessing steps (e.g., tokenization, handling special characters, or data augmentation).
- **Synthetic Data (if applicable):** If generating data, explain the prompt engineering or LLM pipeline you will use to ensure quality.

4. Plans for Models

- **Baselines:** What is the simplest model you will compare against? (e.g., a Naive Bayes classifier or a basic LSTM).

[1] [Optional] I recommend you use Zotero for citation manager and it has extension for some platforms like google doc

- **Proposed Architecture:** Describe the primary model(s) you intend to use (e.g., BERT-based transformers, GPT-4 API, or T5 for multilinguality).
- **Implementation:** Specify the libraries you plan to use (e.g., PyTorch, Transformers, spaCy).

5. Plans for Evaluation

- **Metrics:** Which metrics are appropriate for your task?
    - *Classification (Hate Speech):* Precision, Recall, F1-score.
    - *Generation (Poetry/Synthetic Data):* BLEU, ROUGE, or human evaluation for "creative" quality.
- **Validation Set:** How will you ensure your model isn't just memorizing the training data?

6. Team Members & Responsibilities

- **Please list all the members and what each will be responsible for**
    - **Member A:** (e.g., Data collection, preprocessing, and GitHub maintenance.)
    - **Member B:** (e.g., Model implementation and hyperparameter tuning.)
    - …

7. Project Timeline & Milestones

- **Phase 1 (Weeks 1-2):** Data collection and baseline implementation.
- **Phase 2 (Weeks 3-4):** Primary model training and error analysis.
- **Phase 3 (Week 5):** Finalizing results and writing the report.

8. GitHub Repository

- Provide a link to a private or public repository.
- **Requirement:** Your repo must include a `README.md` and a `requirements.txt` file.