Group Proposal

Natural Language Processing Final Project

Minwoo Yoo

October 20, 2025

1. What problem did you select and why did you select it?

- **Main objective:** Apply NLP methods to estimate media slant in newspaper text and classify articles by topic.
- **Motivation:** The project is related to my research on how local economic conditions shape variation in local newspaper slant.

2. What database/dataset will you use?

The project will use two complementary datasets:[1]

- **Local newspapers (NewsLibrary):**
  - **Coverage:** U.S. local newspapers spanning the late 1980s through the 2000s.
  - **Scale:** Approximately 20–30 million article records are expected.
  - **Structure:** Newspaper name, publication date, article headline, and the first 80 words of text (per NewsLibrary access policy).
  - **Role in pipeline:** *Application & validation.* After training the model on labeled political speech, apply estimated coefficients/scores to newspaper articles; evaluate performance and analyze slant distributions by topic.
- **Congressional speeches (Congressional Record):**
  - **Coverage:** Late 1980s–1990s congressional floor speeches.
  - **Structure:** Speaker identity, party affiliation, chamber, and date; text segmented into speech units.
  - **Role in pipeline:** *Supervised training signal.* Use party labels to train the slant estimator (e.g., TF–IDF features with logistic regression / Naive Bayes, or a text-scaling variant). Learned coefficients (word/bigram weights) are then transferred to newspaper text.

[1] The NewsLibrary dataset is not yet available but will be shared by a professor by early to mid-November.

3. **What NLP methods will you pick from the concept list? Will it be a classical model or will you have to customize it?**

The project will focus on classical NLP methods introduced in the course.

- **Text preprocessing:** Clean and normalize newspaper text through tokenization, stopword removal, and stemming.
- **Feature representation:** Represent articles as bigram-based Bag-of-Words and TF-IDF matrices.
- **Modeling:** Apply logistic regression and Naive Bayes to estimate slant and classify articles by topic.
- **Topic modeling:** Use Latent Dirichlet Allocation (LDA) to identify key themes and compare slant across topics.

4. **What packages are you planning to use? Why?**

- **NLTK and SpaCy:** For tokenization, stopword removal, and stemming.
- **scikit-learn:** For feature extraction (Bag-of-Words, TF-IDF), supervised learning, and model evaluation.
- **gensim:** For topic modeling using LDA.
- **pandas and numpy:** For data handling and matrix operations.
- **matplotlib and seaborn:** For visualization of model performance and media slant distribution.

5. **What NLP tasks will you work on?**

Since this is an individual project, I will complete all parts of the analysis on my own. The project focuses on two main NLP tasks:

- **Slant estimation:** Measure the degree of media slant using supervised classification models.
- **Topic classification:** Categorize articles into topics such as politics and economy to examine variation in slant.

6. **How will you judge the performance of the model? What metrics will you use?**

Model performance will be evaluated using standard metrics introduced in the course.

- **Accuracy:** Proportion of correctly classified articles.

- **Precision, Recall, and F1-score:** To evaluate balance between false positives and false negatives.
- **Confusion matrix:** To visualize classification errors.
- **Cross-validation:** To ensure robustness and prevent overfitting.

7. Provide a rough schedule for completing the project.

Assuming the dataset is received by early November, the tentative schedule is as follows:

Week	**Tasks and Goals**
Week 1 (early Nov)	Receive and inspect dataset; perform cleaning and preprocessing.
Week 2	Construct Bag-of-Words and TF-IDF representations; begin exploratory analysis.
Week 3	Train supervised models for slant estimation and topic classification; tune hyperparameters and validate.
Week 4	Apply topic modeling (LDA) and analyze how slant differs across topics.
Week 5 (late Nov–early Dec)	Summarize findings, generate visualizations, and prepare the final report and presentation.