Capstone Development (DS-5999)

Project Proposal

Please answer the following questions to help guide you in scoping out your Capstone project. Each answer should be about a paragraph for questions 3-7.

1. **Who are you working with on this problem? This could be a company, a faculty member, research group, etc. If it is just you, that's fine, just say "NA" here.**

   I am working with Professor Scott Crossley.

2. **Will you need to meet with me during the semester? If you said "NA" above, the answer to this question is "Yes." Not meeting with me requires that some other faculty member or practitioner is volunteering to provide mentoring and/or oversight. If you say "No" here, please let me know who this person is who will perform this task. It is perfectly fine for you to meet with both me and someone else if you prefer that, and you can change your mind later.**
   I can provide the following:
   a. **Accountability on timelines, planning and meeting deadlines.**
   b. **Help in how to solve difficult logical coding/analysis challenges.**
   c. **Assist in formulating and calibrating metrics that effectively communicate results.**
   d. **Scoping out your problem and ensuring it is not too big or too small.**
   e. **Anything having to do with analytics, statistics, and classification/regression ML tasks.**
   f. **I will be of less help in the following areas: Image recognition, NLP as well as specifics in deep learning.**

   No, I will not need to meet with you during the semester since I will be meeting with Professor Crossley for guidance and oversight.

3. **Describe the problem you are solving. Be sure to include why this problem is unique or novel. If you are working in a research group or team, be sure to detail your precise contribution in relation to what the whole group is doing.**

   This project analyzes how different news sources linguistically frame discussions of Tibetan freedom and human rights using computational NLP techniques. The problem addresses a critical gap in understanding media bias: while manual content analysis of news coverage exists, few studies have applied modern NLP methods to systematically quantify framing differences across ideologically diverse sources on politically sensitive

topics. What makes this unique is using topic modeling to measure sentiment and framing patterns across 6,400 articles spanning 16 years, directly comparing Western media, Chinese state media, and neutral international sources using quantifiable linguistic metrics rather than subjective analysis. This project bridges computational social science with political communication studies, offering both technical innovation in NLP application and substantive insights into how media shapes public understanding of complex international issues, while developing a replicable framework for analyzing media bias on geopolitical topics.

4. **Describe the data you need for your project. Be as detailed as you can be here – even including key column names you require is encouraged.**
   NOTE: You should be sure that you have the necessary access to this data before the beginning of the semester. Access to data is a key point of delay in these projects!

   I need 6,400 news articles from 2008-2024, sampled at 100 articles per source per year across 4 source categories: Western media (BBC, Guardian, or NYT), Chinese state media (China Daily, Global Times, or Xinhua English), neutral international media (Al Jazeera, Deutsche Welle, or Reuters), and optionally Tibetan-focused media (Tibet Post or Phayul). Each article must contain the keyword "Tibet" and include the following fields: full article text (body content), headline, publication date (for temporal analysis), source/outlet name (for cross-source comparison), URL (for reference), author if available, and article category/section. Data will be collected through web scraping using BeautifulSoup/Scrapy, news APIs where available (Guardian API, News API), and potentially the GDELT Project database for supplementary structured event data. I am currently testing web scraping approaches and will prioritize sources with APIs or easily accessible archives to ensure data access is secured before the semester begins.

5. **Describe your approach or primary task. What are you going to do with the data above to answer the question above?**

   My analysis will proceed through four main stages: (1) Data preprocessing and exploration including text cleaning, named entity extraction using spaCy, and temporal tagging linked to major events; (2) Sentiment analysis using fine-tuned BERT models to classify overall sentiment and aspect-based sentiment toward specific entities (Chinese government, Dalai Lama, autonomy policies), followed by statistical significance testing to compare Western vs. Chinese media patterns; (3) Discourse and frame analysis extracting terminology patterns (autonomy vs. independence vs. separatism), building co-occurrence networks, analyzing semantic roles to identify agency, and conducting temporal correlation analysis between sentiment and political events; and (4) Topic modeling using Latent Dirichlet Allocation (LDA) to identify thematic clusters and dynamic topic modeling to track how topic prevalence changes over time across source categories. The final output will be an interactive dashboard using Plotly/Dash visualizing sentiment trends, topic evolution, and cross-source comparisons with statistical measures, accompanied by a comprehensive research paper documenting methodology and findings.

6. **Describe what you have done so far, along with an estimate of how much time you have invested in this project already.**

I have completed a literature review, investing approximately 8-10 hours so far. This review covered academic papers on media framing analysis, computational approaches to political discourse, and existing NLP studies on media bias. Through this process, I identified methodological gaps in current research, developed my research questions, and formulated testable hypotheses about sentiment differences across news sources. I have not yet begun data collection, web scraping development, or technical implementation, which will be my focus before and during the semester.

7. **Describe what you think will be the biggest challenges you will face in executing this project. Identify 1-3 challenges.**

The three biggest challenges I anticipate are: (1) Data collection and access, as Chinese state media sources may have limited English-language archives or geo-restrictions, and some Western outlets have paywalls or restrict scraping, which could result in imbalanced datasets—I will mitigate this by prioritizing sources with accessible APIs, having backup sources for each category, and using GDELT Project as a supplementary source if needed; (2) Model fine-tuning and domain adaptation, as pre-trained BERT models may not perform optimally on political discourse about Tibet without fine-tuning on specialized terminology—I will address this through transfer learning from models pre-trained on news data and creating a high-quality labeled dataset of 500-1000 articles for fine-tuning; and (3) Maintaining analytical objectivity given Tibet is politically sensitive with strong opinions on all sides—I will ensure objectivity by establishing clear pre-defined criteria for all analytical decisions before seeing results, using statistical significance testing rather than subjective interpretation, and clearly documenting all methodological choices and limitations in the final paper.