DSC 180A Quarter 2 Project Proposal

Haoyu Fu, Yixuan Zhang

December 2023

1 Broad Problem Statement

In a world increasingly shaped by data-driven insights, our Quarter 2 project embarks on an ambitious journey to explore the efficacy of advanced predictive models, specifically integrating a large language model with the LangChain framework, in deciphering political preferences. This initiative is pivotal in an era where understanding political tendencies is crucial yet challenging due to the dynamic nature of political landscapes. Our project transcends traditional survey methodologies, offering a novel approach that not only promises efficiency and accuracy but also significantly reduces the high costs typically associated with conventional survey research. The implications of this research could be groundbreaking, potentially transforming how political surveys and research are conducted, thereby making a substantial contribution to both academic research and practical applications in political analytics.

2 Technical Problem Statement

Building upon the foundational work of Quarter 1, where our focus was limited to online data analysis, our current project proposes a more comprehensive approach by integrating offline survey analysis. This integration is crucial in validating the efficacy of the predictive models against real-world data. The core hypothesis is that the combination of a sophisticated language model under the LangChain framework can not only emulate but also accurately predict individuals' political inclinations based on their responses to a set of meticulously curated political questions. This approach aims to address the shortcomings of Quarter 1, where the lack of direct comparison between model outputs and individual political stances was a significant limitation. Our project, therefore, is a critical step in testing the feasibility of AI-driven predictive models in accurately mirroring and predicting complex human political views.

3 Primary Output and Data Analysis Strategy

Our Quarter 2 project is structured to yield the following primary outputs:

3.1 Comprehensive Report and Scholarly Paper

- The report will provide a detailed account of the project's methodology, data analysis techniques, findings, and their implications in the broader context of political survey research.
- The scholarly paper will delve deeper into the theoretical aspects, showcasing our findings in a format suitable for academic discourse.

3.2 Interactive Website

- To engage a wider audience and ensure the accessibility of our research, an interactive website will be developed.
- This platform will present the project's methodology and findings in a user-friendly manner and offer interactive elements, such as visualizations and a demonstration of the model in action.

3.3 Data Analysis Tools

- Depending on the project's success and the insights gained, we may develop an application or a set of analytical tools for other researchers or practitioners to conduct similar analyses.

Our project involves both data analysis and data generation, with a two-fold approach:

3.4 Analysis of Collected Data

- Advanced statistical and machine learning techniques will be used to analyze responses collected from our surveys and Twitter data.
- This includes pattern recognition, sentiment analysis, and comparative studies between human responses and model predictions.

3.5 Generation and Analysis of Predictive Data

- The large language model, equipped with the LangChain framework, will generate predictive data (i.e., responses to political questions).
- We will analyze how closely these predictions align with the actual survey responses, using metrics like accuracy, precision, and recall.

4 Project Success Justification and Data Viability

In justifying the projected success and affirming the viability of the data for our Quarter 2 project, it is imperative to delve into the multifaceted aspects of data accessibility, quality, and the foundational learnings derived from our previous quarter's endeavors. The data, a linchpin for the entire project, is assuredly available and can be procured through established and reliable channels, such as Qualtrics surveys for structured data collection and Twitter's public APIs for obtaining a rich corpus of real-time, unstructured political discourse. This dual approach ensures a comprehensive and diverse dataset, critical for the robustness of our analysis. Furthermore, the quality of the data is underpinned by meticulous preliminary analysis, which serves as a litmus test for its relevance, authenticity, and freedom from biases or inconsistencies. Our methodology for data collection, rooted in validated survey techniques and advanced digital tools, guarantees data integrity and accuracy. Additionally, the lessons gleaned from our Quarter 1 project, particularly the recognition of the need for a more nuanced and direct comparison of predictive models with real-world data, have been instrumental in shaping a more refined and calibrated approach for this project. This evolution in our methodology is anticipated to significantly enhance the predictive accuracy and reliability of our model, thus solidifying the foundation for the anticipated success of our Quarter 2 project.