Sam Cohen PPOL 6801

***Text as Data: Final Project Proposal***

***Background & Research Question(s)***

Officials from all of the world’s countries have gathered at the United Nations General Assembly (UNGA) in New York every year since 1946. Here, representatives of UN member countries discuss the most pressing issues of the day. The UNGA general debate typically gives each member country of the UN the chance to voice their opinions, concerns, and visions for the future of the international order. The speeches given here offer valuable insight into how different nations and regions view diplomacy, international trade, geopolitical crises, and more. For my final project, I propose the following research question(s): *What topics and themes exist in UN speeches? How can these topics, sentiments, and trends tell us about different cultural and socioeconomic aspects of their respective countries? How well do these line up with historical events or trends?*

***Data***

For this project, I will be utilizing the UN General Assembly Debate Corpus, containing opening speeches from each companies UN representative, from 1946 to 2023. The web source for this corpus can be found at the footnote below.[[1]](#footnote-1) This corpus contains information at the country-year level—that is, the year the speech was given, the speaker’s country, the text of the speech, as well as other metrics such as democracy score, regime types, gender equity scores, etc. While the data was retrieved in a tidy form from Kaggle, the full source of the corpus is from Slava Jankin, Alexander Baturo, and Niheer Dasandi and their paper: "Words to unite nations: The complete UN General Debate Corpus, 1946-present," from the Journal of Peace Research (forthcoming).

In addition to text analysis, I will also compare sentiment and topic trends to other social, economic, and government metrics by country and year. For instance, I will compare sentiment scores and topic instances to indicators like democracy scores, freedom index scores, GDP per capita, Gini coefficient, etc. This data will come from the World Development Indicators database from the World Bank.

***Methodology and Outcomes***

I will employ Latent Dirichlet Allocation (LDA) topic modeling to decipher which topics in UNGA speeches are prevalent, and see how these change over time. I will also use sentiment analysis (specifically, Lexicoder Sentiment Dictionary) to see if temporal changes line up with certain historical events. For instance, after the collapse of the Soviet Union in 1991, did Russia’s speeches contain more negative sentiment? Did the fall of the Berlin Wall lead to more positive speeches from German UN reps?

In the end, I hope to uncover temporal trends of different topics, and map those along positive versus negative sentiment so see if certain topics rise or fall with negative or positive sentiments. I also hope to break this down by region, country, and/or socio-economic indicator to illuminate any nuances at the local level. This will involve creating choropleths, column charts, and line graphs.

1. Slava Jankin, Alexander Baturo, and Niheer Dasandi. (2024). NLP with UN General Debate Corpus 1946-2023 [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/9304324. <https://www.kaggle.com/datasets/namigabbasov/united-nations-general-debate-corpus-1946-2023?resource=download> [↑](#footnote-ref-1)