**TransciptCOVID: A COVID-19 Transcript Analysis Across Countries**

**DSGA 1015: Text as Data**

Angela Marie Teng (at2507), Lakshmi S. Menon (lsm454)

**Abstract**

Despite the importance of using unsupervised techniques in predicting the outcome and efficacy of public health-related policies, the swift onset of the coronavirus pandemic has created little opportunity for researchers to study the topics and readability of political texts. This paper demonstrates an application of structural topic models and latent dirichlet analysis on a corpus of coronavirus-related presidential transcripts. The main objective is to exploreThe researchers aim to answer the question, “How do the topics of each country’s coronavirus political address corpus influence their policy responses? Additionally, how do the ways in which country leaders engage with and address coronavirus-related topics differ?” The researchers find that \_\_\_\_\_

Word Count: (excluding appendices and references)

**Introduction**

The coronavirus pandemic has affected the entire world, resulting in thousands of deaths, and a nearly global shut down. Many countries have been praised, while others have been criticized, for their varied responses and measures taken to control the spread of the virus. While containing the virus and protecting lives has been the primary goal for most, if not all, countries, the means by which they do so have been affected by their own beliefs, resources, and priorities. Through this project, we aim to study the responses of five different countries by examining their COVID-19 update press conferences, in order to analyze how different policies and political beliefs affect different leaders’ responses to the same national emergency.

**Literature**

**Theory and Hypotheses**

**Data and Models**

1. **Data Collection**

# In order to measure countries’ public response in a consistent manner, we decided to limit our data to transcripts of press conferences/briefings held by country leaders. We scraped our data from the records at *rev.com*, a transcription service. We identified transcripts related to coronavirus based on the website’s classification, as they had a dedicated page for ‘Coronavirus Briefing & Press Conference Transcripts’1. We decided to limit our analysis to five countries: United States, Canada, United Kingdom, Australia, and New Zealand. This decision was made based on the texts available on the website. In order to maintain consistent among countries, we decided to use only leader speeches, which meant that we did not include conferences by US State Governors and Mayors, as well as briefings by the World Health Organization. We did, however, include briefings by the ‘Coronavirus Task Force’ in the case of the United States, as these often included the US President Donald Trump, and hence reflected the country leader’s response. We first scraped the URLs and titles of all articles available under the coronavirus category, then narrowed them down to country leaders’ speeches by checking for either a country name or leader name in the title. The final list of terms used to match country leader transcripts is shown below:

# 

After narrowing the texts based on the shown terms, we were left with 120 articles. The distribution by country is shown below:

\*\*\*\* INSERT COUNTRY DISTRIBUTION \*\*\*\*\*

Using the narrowed list of URLs, we scraped the transcript text from the *Rev* website. We scraped the transcript as a single large text without segmenting it by speaker. This was mainly done in order to preserve context. We also believed that this would still be in line without our goal of predicting the country response, as questions asked by reporters and responses by other speakers such as health officials still reflect the concerns and response of that country. Additionally, we also scraped the date of each transcript. We used all relevant articles from the first available one, on Feb 25, 2020, up to the most current ones on the date of scraping, which was April 22, 2020. Our final dataset was a table with 120 rows of articles, and 5 columns, containing the URL, title, date, country, and text for each article

1. **Data Processing**
2. **Modeling**

The researchers used a number of methods to model the topics discussed in each country’s coronavirus transcript. The structural topic model is a generative model that applies word counts to identify the topical prevalence and topical content of a given document.

**Results**

**Discussion**

**References**

1<https://www.rev.com/blog/transcript-tag/coronavirus-update-transcripts>