*Congressional Hearing Database: A Shiny App for Viewing and Searching Political Dialogue*

Taylor Crockett

EPPS 6354: Information Management

6 May 2024

# Introduction

The process of legislation is simultaneously public and private. Elected officials have acclimated to using social media to curate their self-images, stay in touch with their constituents, and advocate for issues (Hoffmann, Suphan and Meckel 2016)—such is the rise of the “celebrity politician” of the 21st century, or one who has adopted communication advances to keep an ever-present hold in citizens’ lives and create competitive advantages. This advent in turn informs a more public manner of legislation that can be followed by those who previously would have lacked outlets to keep up: more and more individuals turn to social media to get their news, past 60% in 2017 (Fletcher and Nielsen 2017).

Because of the nature of public speech on online forums, historically, social media has been a boon for researchers to scrape data on policymakers and perform content analysis: for instance, one study attempted to categorize types of political posts on Instagram,[[1]](#footnote-1) and another study used Facebook text to create political profiles using a network model.[[2]](#footnote-2) However, this type of communication from legislators is intentionally and carefully curated, usually by large teams, to convey certain messages. The likelihood of public stances correlating with actual legislative dialogue is relatively unestablished, which is one of the foremost reasons I began to develop a method to collect and assess data on the content of legislative dialogue in a less-public setting.

In the United States, congressional committee hearings serve as a place for legislators to gather information and facilitate their own decision-making processes. Each committee is composed of a smaller body of legislators who often have experience in the committee’s specific area and who are distributed across parties uniformly according to the current Congress’s party distribution. In a given hearing, by calling witnesses and directing the flow of questioning, each legislator sheds large amounts of light on their own thought processes and intentions. For instance, as I found in a working paper that I’m developing contemporaneously to this project, Republican legislators were significantly more likely to hold hearings on China and to call witnesses who spoke on China than Democrats, especially when Republicans held that hearing’s respective chamber.

The possibilities of using committee hearing data are enormous, firstly because the data will satisfy a compromise between the public-private dimensionality of the legislative process: while committee hearings are most often public with the exception of closed-door hearings entailing classified information, the content of hearings is relatively difficult to access. For instance, to find transcripts of a hearing, one must visit the Government Publishing Office or its website and search for the desired transcript. Even among those who know how to access a hearing by attending, watching or reading, hearing content is relatively unpopular when opposed to floor speeches or interviews because it is heavily procedural and investigative, while other political content is intentionally explosive and flashy. I spoke more on the differences between hearing content and other political dialogue in a previous paper about Diversity, Equity and Inclusion (DEI) policy in the context of congressional committee hearings:

Why is dialogue in committee hearings hypothesized to differ from other political dialogue like press statements and social media posts? First, the dialogue is *actionable*. While political dialogue in a democracy that is meant to be public-facing is meant to be an advertisement, showcasing a policymaker's electability to their potential voters, committee hearing dialogue focuses on the policymaking itself. Hearings are a vital opportunity for legislators to get information that will lead to later decision-making on the floor (Diermeier and Feddersen 2000). Communication to voters is curated, but dialogue in hearings is less so. More importantly, the dialogue is relative to other policymakers rather than directly to the voters: certain legislators grandstand in order to gain leverage within or outside of their party (Park 2021).

Secondly, the dialogue is relatively *insulated*. Though there are some prominent cases of congressional committee hearings attracting public attention,[[3]](#footnote-3) most hearings pass by largely unwatched and unnoticed. In 2021 alone, there were 1,118 published hearing transcripts. To expect the average voter to know the goings-on in committee hearings is to expect the impossible. Therefore, in theory, legislators will engage in committee dialogue more focused on achieving their real policy goals rather than appealing to voters, because chances to make a statement are slim.

Due to these reasons, I see congressional committee hearings as an untapped resource for content analysis research. Furthermore, and more importantly, I see hearings as a vital part of the democratic process that are largely left secretive to the average voter, not because of an inherent secrecy by legislators, but simply because of a difficulty of access. The nature of the data is unwieldy: as accessed, transcripts are uncleaned, unorganized, and untagged. Namely, the transformation of hearing content into an accessible database is intense.

Therefore, a tool to compile all hearing content into an accessible and useful database will first require intense amounts of preprocessing and organization into a schema. Moreover, the data exist in large transcript chunks, but the data is most rich and useful at the person-entity level. Searching for a single hearing and accessing its entire uncleaned transcript is relatively easy, but the overhead of finding every instance of speech in that raw transcript by a single legislator, for instance, is probably enough to turn away anyone who wants to casually access the data as basic information. By parsing each transcript and organizing it at the person-entity level rather than at the hearing-entity level, the accessibility and usefulness of the tool will be multiplied.

The end goal of this project is to create an organized database and a clean, simple, and searchable interface tool to access it.

# Data Methodology

## Collection

Data was retrieved from the U.S. Government Publishing Office, which digitally provides federal documents at its GovInfo website service, a succession of the previous FDsys service. To pull and compile transcripts was a multistep process that involved the GovInfo developer API

## Preprocessing

## SQL

# References

Diermeier, Daniel, and Timothy J. Feddersen. 2000. "Information and Congressional Hearings." *American Journal of Political Science* 44 (1). doi:doi.org/10.2307/2669292.

Fletcher, R., and R.K. Nielsen. 2017. "Are news audiences increasingly fragmented? A cross-national comparative analysis of cross-platform news audience fragmentation and duplication." *Journal of Communication* 476-498. doi:https://doi.org/10.1111/jcom.12315.

Grčić, Klara, Marina Bagić Babac, and Vedran Podobnik. 2017. "Generating Politician Profiles based on Content Analysis." *Journal of Universal Computer Science* 23 (3): 236-255.

Hoffmann, Christian Pieter, Anne Suphan, and Miriam Meckel. 2016. "The impact of use motives on politicians’ social media adoption." *Journal of Information Technology & Politics* 13 (3): 239-256. doi:https://doi.org/10.1080/19331681.2016.1169241.

Koblin, John. 2022. "At Least 20 Million Watched Jan. 6 Hearing." *New York Times.* https://www.nytimes.com/2022/06/10/business/media/jan-6-hearing-ratings.html.

Park, Ju Yeon. 2021. "When Do Politicians Grandstand? Measuring Message Politics in Committee Hearings." *The Journal of Politics* 83 (1).

Peng, Yilang. 2021. "What Makes Politicians’ Instagram Posts Popular? Analyzing Social Media Strategies of Candidates and Office Holders with Computer Vision." *The International Journal of Press/Politics* 26 (1): 143-166. doi:https://doi.org/10.1177/1940161220964769.

1. Peng 2021 [↑](#footnote-ref-1)
2. Grčić, Bagić Babac and Podobnik 2017 [↑](#footnote-ref-2)
3. Koblin 2022 [↑](#footnote-ref-3)