Abstract

While there is much political science research focusing on special interest groups and how they affect decision making on federal policy there is limited similar research focusing on municipal policy. Local interest groups play an important role in decision making in city council meetings as they lobby, provide public comment, and help craft legislation. This research aims to study how local special interest groups are referenced during city council meetings, to measure and understand the impact they have on decision making and public policies. To do so, we use Natural Language Processing (NLP) to process a dataset of city council meeting transcripts, to understand who is participating and how they participate. We will then tie the detected references of special interest groups to legislative outcomes in order to measure the impact that special interest groups have on the local municipal process. This quarter, our team explored the use of Spacy's Named Entity Recognition (NER) model to identify references to special interest groups (SIGs) in city council transcripts. While the model performed well in identifying entities such as PERSON, ORG, it fell short in accurately identifying SIGs. As a result, we pivoted our approach and wanted to train our own Span Categorization model to improve the accuracy of SIG identification. To prepare the dataset for training our model, we utilized Prodigy to annotate 2149 instances of different city council meeting transcripts.

Introduction

The problem of how special interest groups influence policy making in the United States is an important and timely issue. With increasing public concern about the role of money in politics and the influence of powerful interest groups, there is a growing need to better understand the mechanisms through which these groups shape policy outcomes. By using a quantitative approach to analyze the relationship between references to special interest groups in legislative materials and policy outcomes, our research will provide a unique contribution to the existing literature on the topic. While previous research has largely focused on qualitative analyses of interest group influence, our study will use a quantitative approach to identify patterns and trends in the data. By doing so, we hope to gain a deeper understanding of how special interest groups shape public policy in the United States, and who has the power to make those decisions.

Our research aims to generate a dataset that provides valuable insights into the role of special interest groups (SIGs) in the policymaking process. By analyzing local council meeting transcripts, our study seeks to answer key questions about the ways in which SIGs are referenced and the types of SIGs that are most influential in shaping policy outcomes. Through this research, we aim to better understand the mechanisms through which SIGs influence the policymaking process and identify patterns and trends in the data. Generating a comprehensive dataset that answers these important questions will allow us to determine the extent to which special interest groups influence policy outcomes in local municipalities. By understanding how

interest groups are influencing policy outcomes, we can work towards a more equitable and representative political system that truly serves the interests of the American people.

Background and Related Work

Previous studies on local special interest groups have been minimal. Local municipal government activity has been broadly viewed as non-partisan and well-functioning. As a result, the research regarding interest groups has primarily focused on national politics while neglecting the work down at the local level.

More recent work on special interest groups have remedied the absence of data on local interest groups by observing how these groups engage and contribute to public comment during municipal government meetings. These studies have sought to quantitatively measure the responsiveness of public officials to the views expressed in public comment. Responsiveness has been measured by analyzing the transcripts of local municipal government and committee meetings to identify (1) who participates during public comment and (2) the voting outcomes of public policy. Here, public policy is examined as the dependent variable. In his study titled 'Public Comment and Public Policy', Alexander Cahn conducted regression analysis to investigate the existence of a statistically significant relationship between types of public comment and the likelihood of approval.

Largely, current studies have recognized the influence of local interest groups on public policy, but the extent of their effect is still unknown. To build upon these efforts, our group seeks to use natural language processing to examine the presence and activity of interest groups at the local level across multiple municipalities. This approach differs from current research that often focuses on either on the national level or a singular city.

Data and Method

Our initial work in solving the research problem of identifying when special interest groups are most referenced during council meetings involved applying a "Named-Entity Recognition"(NER) model, pre-processing, and analyzing a subset of the dataset. This subset included three months' worth of council meeting data from three cities – Seattle, Oakland, and Louisville. Our first step was to apply spaCy's NER model to recognize references to people, organizations, and special interest groups. We chose to use the "NORP" label in spaCy to identify special interest groups references since "NORP" stands for "Nationalities or religious or political groups". We used the label "PERSON" to identify people references and "ORG" to identify organization references.

We then continued with some data preprocessing on our NER results. Since we were interested in knowing when an entity was recognized during a meeting, we normalized the "sentence-index" to be "percent-progress-in-meeting". To do this, we calculated the percentage

of the meeting where a particular sentence appears, and added a new column called 'percentage' indicating when an entity is recognized in the meeting. This normalization step made it easier for us to make plots later on and allowed us to be able to compare data across different meetings.

After normalization, we performed additional filtering on our data. We created a list of entities that are safe to remove and frequent entities that are not meaningful to our analysis. Additionally, we used different filtering strategies, including substring match and exact match, to ensure that more unique entities such as council member names are given a larger tolerance while more general entities are matched exactly. This allows us to avoid accidentally filtering out important special interest groups that contain commonly occurring words like “county” or “community”. For example, 'Jefferson county farm bureau' would have been erroneously excluded if we only used substring matching for the word “county”.

As a result of the data processing step, we created a cleaned NER results dataset with columns including “session_id”, “sentence_index”, “named_entity_type”, “entity”, “city”, and “percentage”. With the cleaned and filtered dataset, we then created plots using the seaborn and matplotlib libraries to visualize our NER results. The plots we produced include various histograms that show the percentage of recognized entities at a given percentage progress of meetings, a box plot that shows the distribution of recognized entities at a given percentage progress of the meeting, and different bar charts that visualize the top 20 recognized entities across all meetings and its occurrences.

Figure A.3 When SIGs are referenced acrossing all meeetings (by city and by entity type)

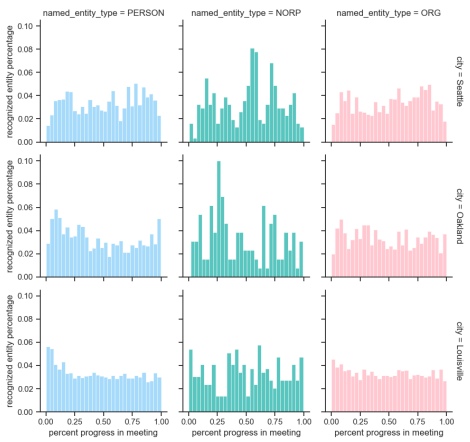

Figure A.3 When SIGs are Referenced Across all Meetings: Figure shows the percentage of recognized entities at a given percentage progress of the meeting using the NER results dataset after filtering out unmeaningful entities. The named-entity types included are PERSON, ORG, and NORP and are plotted individually again for each city. With the filtered NER results, there has been some change in the distributions across the three cities for all entity types.

Figure A.5 Recongized Entities Across Meetings

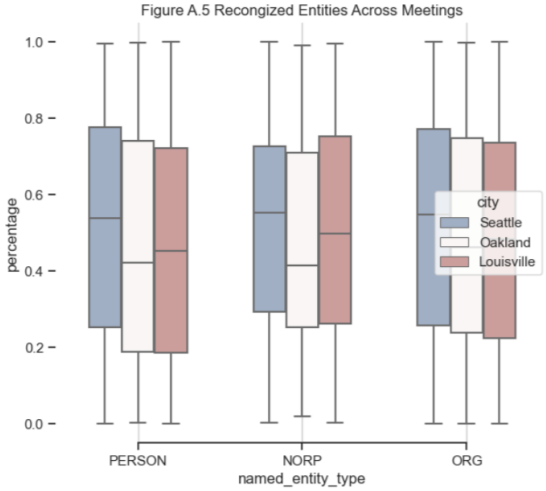

Figure B.4 Top 20 recognized entities across all meetings (for each entity type)

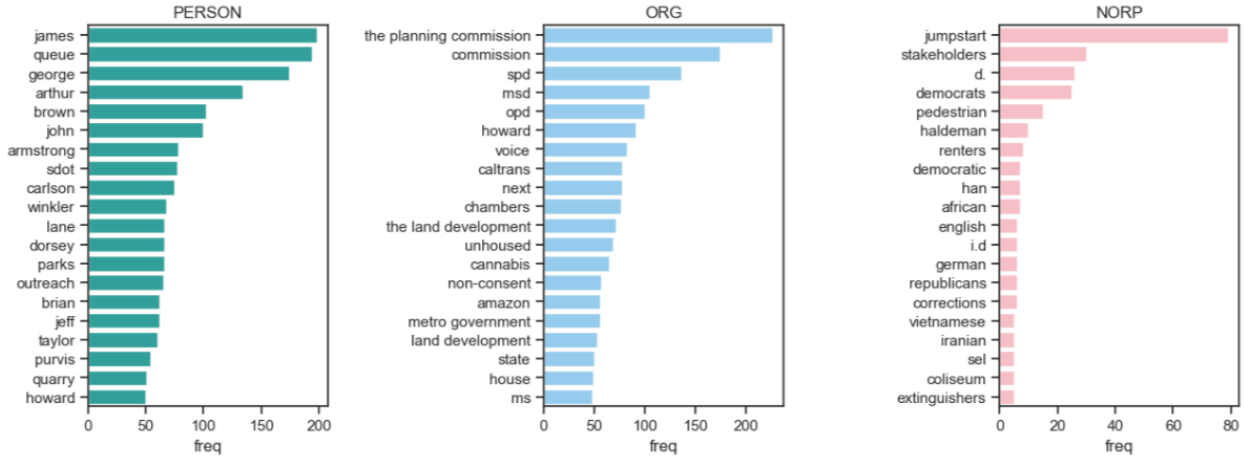

Figure B.4 Top 20 Recognized Entities Across all Meetings: Figure shows the entities with the highest frequency of mentions across all meetings from September and December 2022 using the filtered NER results data for each city. The named entity types of interest, PEOPLE, ORG, and NORP, are plotted individually. For PERSON, the top entities are "james", "queue", "george". For ORG, the top entities are "the planning commission", "commision", and "spd". For NORP, the top three entities are "jumpstart", "stakeholders", and "d,".

What worked well in our initial work is that we were able to use the named entity recognition model on each sentence of each transcript dataset to recognize references, clean and process the dataset, and produce various data visualizations. As a result, we gained a good understanding of the percentage of the types of entities that are recognized, the distribution of recognized entities at a given percentage progress of the meeting, and the most appeared entities across all meetings.

"cdp-affiliations" Dataset

The dataset we compiled is the combination of all 3 of our prodigy annotations over 2 weeks on various city council transcripts across multiple municipalities. It consists of 2149 annotations generated through the use of a Span Categorization model, which allows for labeling potentially overlapping spans of text. Each annotation is a JSON object that documents the sentence or sentence chunk analyzed, along with its tokens, annotation interface used, tag, and timestamp. We chose Prodigy, an annotation tool for NLP, for our annotation on council meeting transcript data because it provides a user-friendly interface for performing the task efficiently. With the "cdp-affiliations" dataset, we can generate more accurate training data to develop a customized model that predicts references to individuals and special interest groups. The resulting model will enable us to make more accurate predictions about the entities we are specifically interested in, which will provide better insights into how special interest groups are referenced during council meetings and their influence on local policy decisions.

Our original research method using SpaCy's Named Entity Recognition (NER) model was unable to resolve the research questions because the entities being recognized were far too general to perform analysis. Despite filtering out an extensive list of words out the transcripts, the model returned words such as "queue", "jumpstart", "d", names of council members, and

other extraneous words. Due to these results, we determined that the general NER models were not equipped to detect the entities relevant to the specific context of municipal meetings.

We determined that customizing and training the NER pipeline to specifically detect local interest groups would result in greater precision and significance. To do so, we have manually annotated the transcripts of municipal meetings to create "training" data for our Span Categorization model. The annotation process presents three challenges: (1) a lack of standard annotation "rules" among groups members creates doubt about when and how entities should be annotated, (2) annotation decisions often require looking at several sentences, and (3) flawed transcripts make it difficult to identify the exact name of the special interest groups.

Our group initially decided that the references that people make to organizations could be represented with 7 labels. Though these labels are specific, we decided to reduce the labels to a number of 2: Individual not Affiliated with an Organization (I-NAG) and Individual Affiliated with an Organization (I-AG). Using the binary method of labeling has allowed for easier, more efficient annotation, while minimizing the nuance between labels to prevent inconsistencies in labeling methods across the team.

Conclusion

This quarter, our group was able to use NER techniques to detect the presence and frequency of references to SIGs in the transcripts. We have also shown that NLP can be a valuable tool in understanding how SIGs are referenced during city council meetings. However, we hope to use the "cdp-affiliations" dataset to train a customized model that would produce more accurate results when detecting local interest groups and other related references. One of our objectives for next quarter is to also come up with more strict annotation criterias to ensure consistency in our annotation work. By using the annotated dataset to train a model to specifically detect local interest groups, we can gain a deeper understanding of their impacts on local policy making from identifying patterns in the council meeting data.

Citation:

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.*

Waskom, M. et al., 2017. *mwaskom/seaborn: v0.8.1 (September 2017)*, Zenodo. Available at: https://doi.org/10.5281/zenodo.883859.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.