Feature Engineering for US State Legislative Hearings: Stance, Affiliation, Engagement and Absentees

Grace, Joshua

Josh.grace@yale.edu
Yale University, United States of America

Khosmood, Foaad

foaad@calpoly.edu
California Polytechnic State University, United States of America

Introduction

In US State government legislatures, most activity occurs in committees made of lawmakers discussing bills. This paper presents systems to extract measures for legislators' engagement and absence during committee meetings and the stance and affiliation of non-lawmakers making public comments. We propose a system to track the affiliation of organizations in public comments and whether the organizational representative supports or opposes the bill under discussion. The model tracking affiliation achieves an F1 score of 0.872, while the support determination has an F1 score of 0.979. Additionally, a metric to compute legislator engagement and absenteeism is also proposed. A list of the most and least engaged legislators over one California legislative session is presented as proof of concept.

This work is a part of a larger project called AI4Reporters, which aims to improve public access to US State Legislative proceedings. Often, non-legislators, such as invited experts and public commenters, are essential to the discourse that can shape legislation. The models presented in this paper can predict the organizations that commenters are affiliated with and their position on the bill with high confidence levels. Additionally, understanding the participation of their elected representatives is vital to the public. Models are presented in this paper, to predict the absentee rates of legislators and their engagement in proceedings, which were developed in consultation with experts on the California state legislature. Code and data used in this work is available at https://github.com/digitaldemocracy/DH2023_Grace_Khosmood.

Public Commenter Organization Affiliation

To extract organization affiliation in public comments, a combination of a Stanford NER model (Finkel et al., 2005), a SpaCy NER model (Honnibal et al., 2020), and hand-crafted rules are used. To train the models, 693 comments from California Legislative hearings are manually tagged with the organizations that speakers stated affiliation to. For each comment, 100 synthetic comments are generated by replacing the original organization with a random organization registered with the California State Legislature (Blakeslee et al., 2015).

To combine the models' outputs, a set of rules is applied. If both models tag an organization, it is added to the output list. Any potential organizations left over are ignored if they are outside the first 12 words or exactly match the speaker's name, a California city, or a county name. These potential organizations are added to the output list if they can be placed in a training sentence in the organization slot and still be recognized by a model as an organization.

To evaluate model performance, 3 models are considered: the Stanford NER, SpaCy NER, and combined model. The models were tested on 193 additional public comments. The metrics are included in Table 1.

Table 1: Performance of Organization Tracker Variants

Model Type	True Positives	False Negatives	False Positives	F1
Stanford NER Model	149	45	47	0.764
SpaCy NER Model	146	51	22	0.800
Combined Model	171	31	19	0.872

To demonstrate the organizational affiliation system, the 5,182 hearings with public comments from the California State Legislature 2017-2018 session were processed with the system to determine the organizations most frequently giving public comments. There were 39,258 total and 15,334 unique organizations referenced in hearings. The ten organizations with affiliates commenting on the most hearings are shown in Table 2.

Table 2: Organizations with Affiliates Commenting on the Most Hearings (2017-2018)

Rank	Organization Name
1	ACLU of California
2	California Labor Federation
3	Western Center on Law and Poverty
4	Sierra Club California
5	California Chamber of Commerce
6	California State Association of Counties
7	League of California Cities
8	California Federation of Teachers
9	California District Attorneys Association
10	State Building and Construction Trades Council

Public Commenter Opinion

When giving public comments, commenters use phrases like 'urge an aye vote', 'support this bill', or 'oppose this bill' to convey their opinion to legislators. Key phrases in these comments can be tracked to model the speaker's position on a bill. These phrases are broken into five sets, strong opposition, strong support, medium opposition, medium support, and weak support, and are listed in Table 3. The number of occurrences of the words from each of the five sets are run through a decision tree classifier to determine if the commenter supports, opposes, or is neutral on the bill.

Position tracking was evaluated using F1 scores. The same dataset of 693 comments used for organizational extraction was used to test and train the opinion classifier. Each public comment in the dataset was manually annotated with the commenter's position on the bill. The system achieved an F1 score of 0.9786 on the separate set of 193 comments.

Table 3: Position Categories and Corresponding Phrase Lines

Category	Phrases
Strong Opposition	"oppose", "opposition", "opposing", "opposed"
Strong Support	"support", "supporting"
Medium Opposition	"no vote", "nay vote"
Medium Support	"aye vote", "yes vote"
Weak Support	"cosponsor"

Legislator Absenteeism

Legislators are considered absent from a hearing if they do not speak during the hearings or vote verbally. At the end of most hearings, the committee secretary will call the legislator's last name to request their vote and will indicate that the legislator voted. If a legislator is absent from the hearing, there will be no voting response to the request. So, the committee secretary's speech is parsed to determine if a given legislator was referenced when voting. If a legislator does not speak during the meeting and is not referenced by the committee secretary, then there is a very high probability of the legislator being absent, as present legislators would be expected to perform one or both of these actions. Committee hearings where no vote is taken and hearings where more than 60% of legislators are absent are considered special sessions and are excluded from the legislators' absent count.

Legislator Engagement

To track legislator engagement, a set of rules were developed. Engagement is modeled as detectable verbal interactions during a hearing. This allowed for text-based processing of a legislator's dialog to track engagement. Four factors are considered to calculate engagement in the proceedings: verbally voting on a bill at the end of a hearing (equation 1), speaking during the committee (equation 2), back-and-forth conversations with non-legislators as described in (Klimashevskaia et al., 2021) (equation 3), and asking questions (equation 4). These variables were chosen after discussions with the chief of staff to a former elected member of the leadership in the California State Senate, Christine Robertson. The authors thank her for her insights. Each variable is multiplied by a constant detailed in Table 4 to scale the value. The final engagement score is calculated by summing the four scores and is detailed in equation 5.

$$vote\_score_p = \alpha * \frac{number\_votes_p}{number\_commitee\_hearings_p}$$
Equation 1: Vote Score Equation

$$speaking\_score_p = \beta * number\_times\_speaking_p$$
Equation 2: Speaking Score Equation

$$back\_and\_forth\_score_p = \gamma * number\_words\_in\_back\_and\_forth_p$$
Equation 3: Back and Forth Score Equation

$$question\_score_p = \delta * number\_times\_questions_p$$
Equation 4: Question Score Equation

$$engagement\_score_p = vote\_score_p + speaking\_score_p + question\_score_p + back\_and\_forth\_score_p$$
Equation 5: Legislator Engagement Score Equation

Table 4: Legislator Engagement Constants

Name	Value	Effect
α	0.5	Multiplier for vote score
β	0.0005	Multiplier for speaking score
γ	0.00005	Multiplier for back and forth score
δ	0.01	Multiplier for question score

The legislator engagement score was calculated for each legislator in the 2017-2018 California State Legislative session. The ten most engaged legislators are listed in Table 5, while the ten least engaged legislators are listed in Table 6.

Table 5: Ten Most Engaged California State Legislators (2017-2018)

Ranking	Legislator Name	Engagement Score	Voting Score	Speaking Score	Back and Forth Score	Question Score
1	Hannah-Beth Jackson	23.621	0.319	3.139	10.293	9.87
2	Mike McGuire	16.448	0.196	2.086	9.036	5.13
3	Benjamin Allen	13.44	0.235	1.79	4.875	6.54
4	Ricardo Lara	12.028	0.138	1.655	5.014	5.22
5	Jim Frazier	11.512	0.37	0.841	3.211	7.09
6	John Moorlach	11.34	0.331	1.082	2.906	7.02
7	Richard Pan	11.115	0.298	1.321	4.086	5.41
8	Nancy Skinner	10.604	0.261	1.758	4.144	4.44
9	Bob Wieckowski	10.553	0.29	1.296	3.887	5.08
10	Jim Beall	10.392	0.201	1.631	3.33	5.23

Table 6: Ten Least Engaged California State Legislators (2017-2018)

Ranking	Legislator Name	Engagement Score	Voting Score	Speaking Score	Back and Forth Score	Question Score
113	Rob Bonta	0.507	0.081	0.071	0.114	0.24
114	Raul Bocanegra	0.486	0.086	0.083	0.077	0.24
115	Kevin Mullin	0.446	0.136	0.075	0.045	0.19
116	Ken Cooley	0.428	0.19	0.088	0.01	0.14
117	Sabrina Cervantes	0.422	0.304	0.023	0.015	0.08
118	Sydney Kamlager-Dove [1]	0.289	0.069	0.02	0.04	0.16
119	Jimmy Gomez	0.276	0.036	0.026	0.054	0.16
120	Wendy Carrillo	0.219	0.086	0.067	0.016	0.05
121	Jesse Gabriel [1]	0.092	0.076	0.005	0.011	0.0
122	Luz Rivas [1]	0.068	0.047	0.004	0.008	0.01

Notes

1. Elected in late 2018, and served only for the last quarter of the legislative session

Bibliography

Blakeslee, Sam / Dekhtyar, Alex / Khosmood, Foaad / Kurfess, Franz / Kuboi, Toshihiro / Poschman, Hans / Prinzivalli, Giovanni / Robertson, Christine / Durst, Skylar (2015): "Digital democracy project: Making government more transparent one video at a time". Digital Humanities 2015

Finkel, Jenny Rose / Grenager, Trond / Manning, Christopher D. (2005): "Incorporating non-local information into information extraction systems by gibbs sampling". In Proceedings of the 43rd Annual Meeting the Association for Computational Linguistics (ACL'05), pages 363–370.

Honnibal, Matthew / Montani, Ines (2020): "Spacy: Industrial-strength natural language processing in python."

Klimashevskaia, Anastasiia / Gadgil, Richa / Gerrity, Thomas / Khosmood, Foaad / Gütl, Christian / Howe, Patrick (2021): "Automatic news article generation from legislative proceedings: A phenom-based approach". In Statistical Language and Speech Processing, pages 15–26, Cham. Springer International Publishing.