Modeling Effective Lawmaking at the State Level by Leveraging Lexical and Contextual Information for Predicting Legislative Floor Action

Vlad Eidelman,[*]*Anastassia Kornilova,[†] Daniel Argyle[‡]

*To whom correspondence should be addressed.

Abstract

Modeling United States Congressional legislation and roll-call votes has received significant attention in previous literature. However, while legislators across 50 state governments and the District of Columbia. propose over 100,000 bills each year, and on average enact over 30% of them, state level analysis has received relatively less attention. This is in part due to the difficulty in obtaining the necessary data. But another reason is that since each state legislature is guided by their own procedures, politics and issues, it is difficult to qualitatively asses the factors that affect the likelihood of a legislative initiative succeeding. Herein, we present several methods for modeling the likelihood of a bill receiving floor action across all 50 states and the District of Columbia. We utilize the lexical content of over 1 million bills, along with contextual legislature and legislator derived features

[*]1201 Pennsylvania Ave NW, Washington, DC 20004; vlad@fiscalnote.com

[†]1201 Pennsylvania Ave NW, Washington, DC 20004; anastassia.kornilova@fiscalnote.com

[‡]1201 Pennsylvania Ave NW, Washington, DC 20004; daniel@fiscalnote.com

to build our predictive models, allowing a comparison of the factors that are important to the lawmaking process. Furthermore, we show that these signals hold complementary predictive power, together achieving an average improvement in accuracy of 18% over state specific baselines.

Introduction

Federal institutions in the United States, like Congress and the Supreme Court, play a significant role in lawmaking, and in many observable ways define our legal system. Thus, legal scholarship has been largely focused on understanding these entities and the role they play in our society. As federal legislative and regulatory data have become more readily available, political scientists and legal scholars have become increasingly quantitative, adopting objective data-driven methods for characterizing political and legal behavior and outcomes. Computationally driven analysis has extended into all areas of law, including analyzing the behavior of Supreme Court Justices (Katz, Bommarito, and Blackman. 2017; Lauderdale and Clark 2014), Congressional legislators (Poole and H. L. Rosenthal 2007; Slapin and Proksch 2008), and Administrative agencies (Livermore, Eidelman, and Grom 2018; Kirilenko, Mankad, and Michailidis 2014). The aim of most of this research is to move away from purely subjective analysis that is limited in its ability to quantitatively measure and empirically explain observable legal phenomena.

Although many issue areas are regulated primarily at the federal level, state governments have significant power over others, with an increasing number of issues being decided at the state or local levels, including emerging industries and technologies such as the gig economy and autonomous vehicles (Hedge 1998). Moreover, the total quantity of state legislative activity dwarfs that of Congress. There are 535 members of Congress who introduce over 10,000 pieces of legislation a session,[1] of which less than 5% is enacted. Similar dynamics exist at the

1. A session is the period of time a legislative body is actively enacting legislation, usually one to two years.

state level, except on a much broader scale. There are over 7,000 state legislators, in aggregate introducing over 100,000 pieces of legislation, with over 30% being enacted.

In order to be enacted, every bill must pass through one or more legislative committees and be considered on the chamber floor, a process we refer to as receiving floor action. This process is one of the most pivotal steps during lawmaking (Rosenthal 1974; Hamm 1980; Francis 1989; Rakoff and Sarner 1975), as on average, only 41% of bills receive floor action, with most legislation languishing in the committees.[2]

Legislative policymaking decisions are extremely complex, and are influenced by a myriad of factors, ranging from the content of the legislation, to legislators' personal characteristics, such as profession, religion, and party and ideological affiliations, to their constituents' demographics, to governor agendas, to interest group activities, and to world events (Canfield-Davis et al. 2010; Hicks and Smith 2009; Talbert and Potoski 2002). While there has been substantial scholarship to understand the possible influences on legislator behavior (Canfield-Davis et al. 2010), it would likely be impossible to first, create an exhaustive list, and second, to obtain data representing each of factors.

In this chapter we explore how machine learning and natural language processing tools can be used to better understand state lawmaking dynamics and the legislative process. We apply them to the problem of predicting the likelihood that over 1 million pieces of legislation will reach the floor in each state across all 50 states and the District of Columbia. As there are many dimensions underlying the content of the legislation, such as the policy area and ideology of the sponsor (Linder et al. 2018), that may affect the likelihood of floor action, we focus on two sets of features — contextual legislature and legislator derived features and text. Specifically, we follow previous literature and derive several established contextual features. These features describe the legislative environment or identity of the legislator, such as committee membership, party affiliation, or con-

2. For comparison, 13.3% of bills receive floor action at the Congressional level.

trolling party, and can be easily computed from publicly available data. We take a computational text analysis approach to legislative texts, and without relying on in-depth manual analysis of the scope, impact, or complexity, process the text automatically using natural language processing technologies to identify salient content. Using several machine learning algorithms, we build predictive models in each state from different subsets of the features, quantitatively modeling the floor action process across all 50 states. We find that despite the complexity, we can fairly accurately predict legislative success and achieve large improvements in accuracy over majority baselines.

Congressional Related Work

Much of the research on analyzing the federal legislature is aimed at understanding legislator preferences through the use of voting patterns. One of the most popular techniques in political science is the application of spatial, or ideal point, models built from voting records (Poole and H. Rosenthal 1985; Poole and H. L. Rosenthal 2007), that is often used to represent unidimensional or multidimensional ideological stances (Clinton, Jackman, and Rivers 2004).

As most of this literature is aimed at building descriptive models with explanatory, not predictive capacity, it presents a few shortcomings. The first shortcoming is that most of these models have been limited to in-sample analysis. In other words, the model is only applicable on the data that was used to construct it. For example, the ideal point models mentioned above can infer ideology from past votes, and predict a legislator's vote for an in-sample bill, but are incapable of making out-of-sample predictions, or in other words, predictions on novel bills.

The second shortcoming is that previous work mostly ignores a fundamental aspect of lawmaking, namely, the text of the laws themselves. So while these models can assess legislator preference, there is no indication of what that preference is based on. Unstructured textual artifacts, such as legislation, floor debates, and committee transcripts, are a much richer representation of the lawmaking process, and the law, than structured artifacts, such as observable votes.

By including this data in our models we hope to achieve a better quantitative understanding of the broader dynamics of legislatures.

In recent years a wide variety of primary and secondary legal data, both structured (e.g., votes) and unstructured (e.g., text), has become increasingly available. Coupled with advances in natural language processing and machine learning this data has enabled the construction of richer statistical models for multidimensional ideal point estimation. For example, researchers in the computer science community have turned their attention to Congressional roll call prediction and created various novel models that account for both the text of the legislation and the voting records to predict out-of-sample votes. Gerrish and Blei (2011) use topic models (Blei, Ng, and Jordan 2003) to construct multidimensional ideal point models, Nguyen et al. (2015) use both the legislative text and floor debates to construct a hierarchical topic model, and Kornilova, Argyle, and Eidelman (2018) use a neural network to create continuous embedding vector representations of both bills and legislators.

While the focus of the above is still on roll call prediction through improvements upon multidimensional ideal point estimation, other work has started to emerge with different, but related aims. Yano, Smith, and Wilkerson (2012) introduced the problem of Congressional bill survival, where the task is to predict whether a Congressional bill will be reported out of the committee to which it was assigned. They utilize a logistic regression model with several different feature sets, combining basic contextual binary features — sponsor and committee — with an n-gram representation of the initial bill text version, showing that the combination leads to the best performing predictive model. Nay (2016) focused on the overall passage problem, predicting whether a Congressional bill would be enacted into law. They take a different approach to modeling the text, using word2vec (Mikolov et al. 2013) to create word embeddings, which are combined to construct a continuous vector space representation of the bills. This representation is used for one predictive model, and stacked in an ensemble with gradient boosting and random forest classifiers trained on a set of contextual features. They also show that the best model combines text and contextual legislature and legislator derived features.

Moving beyond primary legislative texts, floor-debate transcripts have been used in a number of applications. Thomas, Pang, and Lee (2006) use transcripts to predict voting. They frame the task as sentiment analysis, using a SVM classifier trained using the text of the transcripts. In addition to the text, they compute agreement links between different legislator's transcripts, showing that both contribute to vote prediction. Iyyer et al. (2014) use transcripts to build a Recursive Neural Network to detect ideology bias.

There is also an established literature examining broader legislative dynamics, such as measuring legislative effectiveness (Harbridge 2016), evaluating the impact of legislation on stock prices using legislator's constituents (Cohen, Diether, and Malloy 2012), creating cosponsorship networks (Fowler 2006), and examining the role of lobbying (Bertrand et al. 2018; Matthew et al. 2013).

State Related Work

As noted above, while Congress has received much of the attention, many important issues are decided at the state or local level. There are a few reasons for this. As Congress has grown more polarized and less able to act in the recent sessions (Hedge 1998), states have seen increased opportunities to pass legislation that Congress cannot or will not pass or even act on. This has been especially true in emerging industries or technologies, such as the gig economy, or autonomous vehicles. In addition, the federal government gives state governments power over certain areas. For example, the insurance and retail industries are much more heavily regulated at the state and local level.

While there is also an increasing amount of state legislative research, states have received significantly less attention than Congress (Hamm, Hedlund, and Miller 2014). One major reason for this is that quantitative methods require data, and the availability of data for Congress far exceeds that of state legislatures. In fact, Yano, Smith, and Wilkerson (2012) noted "When we consider a larger goal of understanding legislative behavior across many legislative bodies (e.g., states in the U.S., other nations, or international bodies), the challenge of creating and maintaining such reliable, clean, and complete databases seems insurmountable."

A second reason is that modeling and comparative analysis across 50 separate localities poses unique challenges. For instance, spatial models require individuals to have expressed a preference on the same item, and thus are not applicable across different sets of actors expressing preferences on different items, i.e., they would require all state legislators to vote on the same bills. Thus, while there has been scholarship quantifying voting, the role of committees, and other legislative processes, it has been limited in scope, to a few sessions or states, or reliant on survey data (Francis 1989; Rakoff and Sarner 1975; Rosenthal 1974; Hamm 1980). For example, Hamm (1980) constructed a multivariate model using intra- and extra-legislative factors for determining the important variables affecting committee survival in Wisconsin and Texas over three sessions. Rakoff and Sarner (1975) performed a similar analysis for three sessions in New York, while Francis (1989) uses survey responses from over 2000 legislators across 50 states to compare committee performance.[3]

More recently, as different kinds of state data has become more accessible, it has enabled larger comparison studies, for instance the affect of professionalism on legislatures (Squire 2007), the affect of interest groups on legislative activity (Gray and Lowery 1995), the application of spatial models (Shor, Berry, and McCarty 2010; Shor and McCarty 2011), and comparisons of textual similarity (Linder et al. 2018).

The contribution of this chapter is to study state legislative dynamics by evaluating how predictable state lawmaking is, and what factors influence that process. We create a novel task — predicting the likelihood of legislation to receive floor action — and utilize a corpus of over 1 million bills to build computational models of all 50 states and the District of Columbia. We present several baseline models utilizing various features and show that combining the legislative and legislator contextual information with the text of bills consistently provides the best predictions. Our analysis considers various factors and their respective im-

3. Similarly, Shor and McCarty (2011) relies on surveys to derive a scaling for all legislators for ideal point modeling.

portance in the predictive models across the states, showing that although there are some consistent patterns, there are many variations and differences in what affects the likelihood in each state.

Data

There is state-to-state variation in the legislative procedure of how a bill becomes law, but the path is largely similar. Legislation is introduced by one or more members of the legislature in their respective chambers,[4] and assigned to one or more standing subject committees.[5] Committees are made up of a subset of members of their respective chambers, and are chaired by the majority party. Once in committee, legislation is subject to debate and amendment only by the committee members, with the successful outcome being a favorable referral, or a recommendation, to be considered by the full chamber on the floor.

The primary data we use to model floor action was scraped directly from each state legislatures' website. For each state, we downloaded legislation, committee, and legislator pages for all sessions that were publicly accessible. Legislation pages were automatically parsed to determine legislative contextual metadata, which includes bill text versions, sponsors, committee assignments, and the timeline of actions. Legislator pages were parsed to obtain sponsor contextual metadata, which includes party affiliation, committee assignments, and committee roles.

States demarcate legislative status in the timeline of actions differently, so we automatically map and normalize all textual descriptions of legislative actions to a finite set of statuses.[6] These statuses are used to determine whether a piece

4. All legislatures are bicameral, with either a House or Assembly as the lower chamber, and the Senate as the upper chamber, except the District of Columbia and Nebraska, which are unicameral.

5. Depending on the state, other groups can introduce legislation, including legislative committees, legislative delegations, the governor, or non-elected individuals. For the purpose of this work we focus on legislator-sponsored legislation.

6. The normalized statuses include 'introduced', 'assigned to committee', 'reported from committee', and 'passed'.

of legislation survived committee and received a floor action, or consideration on the floor. All bills with a status of either having passed in their introductory chamber or having had a recorded floor vote are treated as positive examples, while any status prior to floor action is considered failed, including legislation that was reported out of committee but not considered on the floor.

Finally, since each state follows their own conventions with regard to classifying the type of legislation, we normalize all legislation across states to two types: resolutions and bills. We use the state-assigned classification to group legislation. Legislation assigned by it's state as type 'appointment', 'resolution', 'joint resolution', 'concurrent resolution', 'joint memorial', 'memorial', 'proclamation', or 'nomination' is mapped to resolution. Legislation assigned a type 'bill', 'amendment', 'urgency', 'appropriation', 'tax levy', or 'constitutional amendment' is mapped to bill.

Figure 1 shows the total number of bills introduced and receiving floor action for each state. In total, our dataset consists of 1.3 million pieces of state legislation, broken into 1 million bills, with 360,000 receiving floor action, at an average rate across states of 41%, and 275,000 resolutions, with 210,000 receiving floor action. On average, we have 10 legislative sessions of data per state. Bills represent substantive legislation with a much lower floor action rate, while resolutions are much more likely to receive floor action, so for the rest of this chapter we focus on bills only, and refer to bills and legislation interchangeably. We include 15 sessions of U.S federal legislation in our data for comparative purposes, with 23,000 of 172,000 bills receiving floor action.

Figure 2 presents the percent of bills receiving floor action. It is interesting to note the difference in difficulty for legislation to receive floor action in different states. For example, in New Jersey and Massachusetts, fewer than 15% of bills reach the floor, whereas 75% do in Colorado and Arkansas.[7]

7. Our average across states, chambers, and sessions is in line with previous single state and session findings; in examining five states Rosenthal (1974) found between 34% and 73% of legislation did not survive committee.

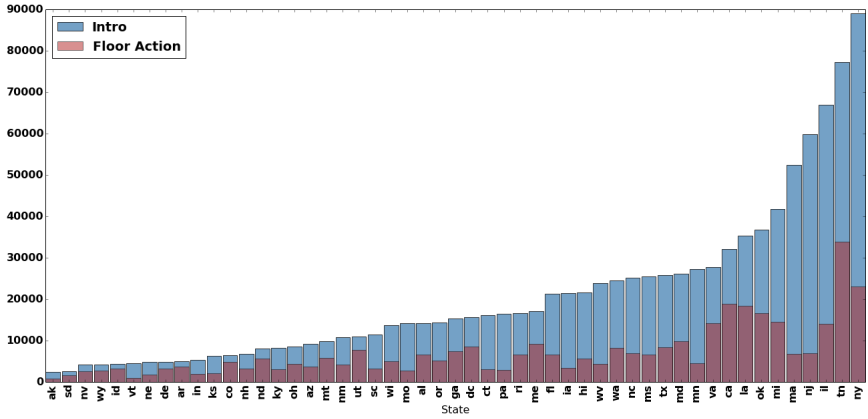


Figure 1. *Number of bills introduced and receiving floor action for each state.*

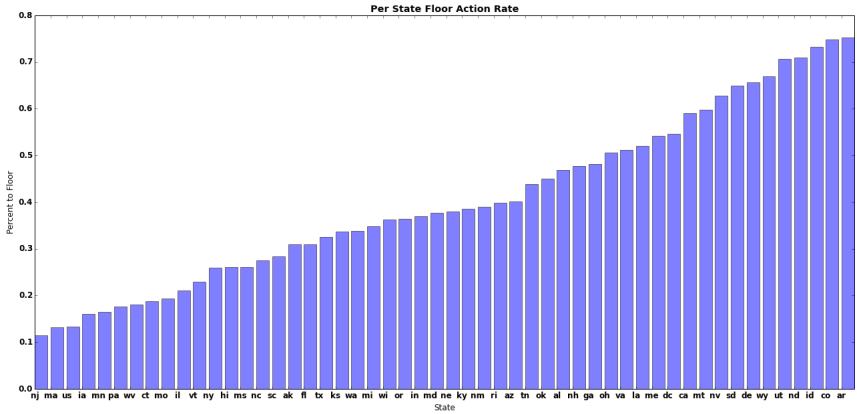


Figure 2. *Percent of bills reaching floor per state.*

Methods

Models

In order to not only be able to predict, but also examine the importance of features to our prediction, we chose three relatively interpretable models for our modeling framework. Formally, let our training data $(\boldsymbol{X}, \boldsymbol{Y})$ consist of n pairs $(\boldsymbol{x}_i, y_i)_{i=1}^n$ where, each $\boldsymbol{x}_i$ is a bill and y_i a binary indicator whether $\boldsymbol{x}_i$ received floor action. Let $\mathbf{f}(\boldsymbol{x}_i)$ be a feature vector representation of $\boldsymbol{x}_i$, and $\mathbf{w}$ the parameter vector indicating the weight of each feature learned by the model. The first two models are linear classifiers — a regularized log-linear model and NBSVM (Wang and Manning 2012) — and the third is non-linear, a tree-based gradient boosted machine (Friedman 2000).

We use the `scikit-learn` (Pedregosa et al. 2011) implementation for the log-linear and gradient boosted models, and implemented NBSVM based on the interpolated version in Wang and Manning (2012).

Hyperparameters in the machine learning algorithms, such as learning rate and regularization strength, have a significant impact on model performance. We use Bayesian hyperparameter optimization (Bergstra et al. 2011) to select the optimal hyperparameters for each model on a held-out development set. We used the tree-structured Parzen Estimator (TPE) algorithm implemented in `hyperopt` for our sequential model-based optimization (Bergstra, Yamins, and Cox 2013). After individually optimizing hyperparameters and training each of the three base models, we use their outputs to train a meta-ensemble model, a regularized conditional log-linear model, forming a linear combination over the three base models' predictions (Breiman 1996).

As the lawmaking process in each state, and even within each chamber, is different, we divide the problem space by state and chamber, building separate models for each subset. Specifically, we consider each of these as separate problems: upper chamber bills, upper chamber resolutions, lower chamber bills, and lower chamber resolutions. It is evident from the prior probability of surviving committee in Table 7 in the Appendix that bills and resolutions receive very different levels of scrutiny. Thus we choose to model them separately. We have 4

types of predictions per state associated with different chamber and legislation types, and each prediction is comprised of 4 model outputs, three from the base models, and one from the meta-ensemble, resulting in 768 models.[8]

Features

As there are many dimensions underlying bills that may affect the likelihood of floor action, we compute and utilize several established contextual legislature and legislator derived features. Previous literature has proposed various factors that may affect legislation, including the content of bills,[9] number of and identity of sponsors, extra-legislative forms of support, timing of introduction, leadership's position, seniority, identity of chairperson of the committee, identity of one's own party, and membership of the dominant faction (Hamm 1980; Rakoff and Sarner 1975; Harbridge 2016; Yano, Smith, and Wilkerson 2012).

In order to quantitatively evaluate these factors and establish a strong baseline from which to measure the affect of text, we include the contextual features shown in Table 1. These indicator features derived from the sponsors, committees, and bills are meant to capture many of the major factors that are proposed in the literature.[10]

To strengthen the representation of legislators in our model beyond the basic features described above, we compute several measures of legislator effectiveness. The effectiveness score is calculated from the sponsoring and cosponsoring activity of each legislator, and meant to represent where they stand in relation to

8. There is only bill type legislation in IN, MA, ME, and WI, and only upper chamber legislation in the District of Columbia and NE so we have (45 states x 4 prediction types + 6 states x 2 prediction types) x 4 models = 768 models.

9. In most previous literature the content is determined via a manual analysis of each bill to establish the scope of impact, the complexity, or the incremental nature.

10. Each count based feature, such as number of sponsors, also spawns a number of discretized features, including ranks, percentiles, and deviations from the mean thereof. We automatically compute companion bills using a cosine-based lexical similarity.

Table 1. *Contextual feature types and descriptions.*

Feature Type	Description
Sponsor	primary and cosponsor(s) identity, primary and cosponsors(s) party affiliation, number of primary and sponsors, number of Republicans, number of Democrats, sponsors bicameral, sponsors bipartisan, sponsor in majority/minority, majority party Republican or Democrat
Committee	identity of assigned committee(s), number of committee assignments, number of sponsors members of the committee, sponsor same party as committee chairman, sponsor role on the committee, referral rate of committee(s)
Bill	chamber, bill type, session, introductory date, companion bill(s) existence, companion(s) current status

other legislators in successfully passing legislation.[11]

Similar to Harbridge (2016), the score we compute for each legislator is a combination of several partial scores, computed for each important stage of the legislative process. Each legislator gets a score for how many bills they sponsored, getting those bills out of committee, getting them to the floor, passing their own chamber, passing the legislature, and getting enacted. The score for each stage is further broken down by how many of those pieces of legislation were substantive, i.e. bills, attempting a meaningful legal change, versus non-substantive (i.e. resolutions). This results in 12 factors for each individual. To compute a score for each legislator's relative performance to the other members

11. This is not a holistic representation of being an effective legislator, as someone may consider themselves effective by not passing anything, or preventing others from doing so. Members may also be highly influential and their support is needed behind the scenes but their names do not appear on the legislation. We can only account for recorded activity. Despite the limitations, we argue this is a fair, if incomplete, assessment of how well the legislator advances their agenda.

in the chamber, we create a weighted combination of that legislator's bills and resolutions, where bills get more weight, and compute the ratio based on the weighted contribution of the other members in the chamber. All the stage scores are then combined into a second weighted combination, where each successive stage in the process gets more weight, to get the final score. Finally, the scores are normalized to 0-10. In addition to using the effectiveness scores directly as features, we further compute and discretize several statistics derived from them, including ranks, percentiles, and deviations from the mean thereof.

To further enrich the bill representation beyond contextual information, we utilize the textual content of the bills. The predominant source of text examined by the NLP community comes from relatively short and topically coherent documents, such as news articles, social media posts and other online resources. Oftentimes the level of analysis will focus on individual sentences, or paragraphs, as opposed to documents, limiting the amount of text that needs to be processed.

Legal text differs from this in a number of ways which make it more challenging. Primarily, the legislation in our collection is comprised of long documents, with an average of 11,000 words, often containing significant amounts of procedural language and pieces of extant statutes. As the length of the document can create challenges for computational solutions to identify the salient points, typical NLP approaches reduce the size of the text being analyzed. For this work we chose to focus on a condensed amount of text, specifically the state provided title and description, that average 17 and 18 words, respectively. Although this is a coarse approximation of the bill content, removing most of the nuances of the scope and impact of the law that a qualitative analysis would focus on, we believe it should capture the substantive aspects of the bill. Both title and description are preprocessed by lowercasing and stemming. We treat each field as a bag-of-words and compute the tf-idf weighting Jurafsky and Martin 2000 for n-grams of size n=1,2,3 on the training data for each prediction task, and select the top 10,000 n-grams from the title and description separately.

While we would like to study the predictability of reaching the floor upon first introduction, bills often change after introduction and are updated with additional information. Thus, we limit our features to those available at the time of first

Table 2. *The five feature settings with contextual and lexical features.*

Condition	Feature Set
`combined`	sponsor, committee, bill, text
`no_txt`	sponsor, committee, bill
`no_txt_spon`	committee, bill
`just_txt`	text
`just_spon`	sponsor

introduction.

Results

In order to clarify the impact that each set of features has on predictive performance, we create five different subsets of features described in Table 2, and train models on each one of them separately.

The first condition, '`combined`' contains all the contextual and text content features. The second condition, '`no_txt`', removes text content, allowing us to study the importance of all contextual features, and by comparing '`combined`' to '`|no_txt|` we can evaluate if text has any complementary information to contextual features. The third condition, '`no_txt_spon`' further removes sponsor features, essentially allowing us to study the importance of committee information. By comparing '`no_txt`' to '`|no_txt_spon|` we can evaluate what sponsors contribute. The fourth and fifth conditions use only sponsor and only text features, respectively, to study the importance of each individually.

All models for a given condition are built from the same training data and feature space. We measure and report several performance metrics of our models using 10-fold cross validation. The baseline model represents guessing the majority class; for some states this means all fail, for others it is all receive floor action, based on the state specific rate.

Although accuracy is informative with respect to how many correct binary decisions the model made, as noted in Bradley (1997) for imbalanced problems such as this, where one class dominates, the baseline accuracy can be very high. As a supplement, it is useful to measure a probabilistic loss, where there is a cost associated with how aligned the model probability was with the correct class. Thus, we move beyond pure predictive performance and consider the actual probability distributions created by our models under different conditions. The log-linear and gradient boosted models are probabilistic, while NBSVM is not, thus we train a probability transformation on top of NBSVM using Platts Scaling (Niculescu-Mizil and Caruana 2005) to obtain probability estimates.

In additional to accuracy, we measure model performance on log-loss and AUROC (area under the receiver operating characteristic curve) (Bradley 1997).[12]

By considering the TP and FP at different values, we can construct a distribution, known as the ROC curve, to visualize the model's performance at various TP and FP thresholds. The area under that curve, AUROC, can be interpreted as the probability that the model will rank a uniformly selected positive instance (floor action) higher than a uniformly selected negative instance (failure), or in other words, the average rank of a positive example. A random model will have a AUROC of 0.5, and a 45-degree diagonal curve, while a perfect model will have an AUROC of 1, and be vertical, then horizontal.

Table 3 shows the average accuracy, *LL*, and AUROC with standard deviations for each of the five conditions on bills. The average baseline accuracy is 68%. The `just_txt` model achieves an accuracy of 73%, outperforming the baseline by 5%, and notably, shows that there is a predictive signal even within the limited amount of text available in the title and descriptions.

To examine where text content is most and least predictive on its own, we disentangle the average performance of the `just_txt` model in Figure 3, showing the per state and chamber pair change from baseline. The states that improve the most over baseline, with 15% improvement or more using only textual fea-

12. For further details see Appendix

Table 3. *Average and standard deviation across states on accuracy, log-loss, AUROC for bills on each feature set.*

Feature Set	Accuracy		Log-Loss		AUROC	
	Ave	Std Dev	Ave	Std Dev	Ave	Std Dev
baseline	0.68	0.1	0.6	0.09	0.5	0
just_txt	0.732	0.09	0.53	0.14	0.7	0.14
just_spon	0.759	0.102	0.48	0.16	0.74	0.15
no_txt_spon	0.81	0.113	0.39	0.18	0.8	0.18
no_txt	0.846	0.098	0.32	0.18	0.82	0.21
combined	**0.859**	0.093	**0.31**	0.17	**0.85**	0.21

tures are Oregon, Oklahoma, Tennessee, District of Columbia, South Carolina, Louisiana (lower), Georgia (lower), and Alabama (lower). On the other hand, text is least predictive in Connecticut, Wyoming, Idaho, New Jersey, Utah (upper), New Hampshire (upper), North Dakota (upper), all underperforming the baseline.

The relatively small improvement over baseline of `just_txt` provides insight into the lawmaking process, raising the possibility that other contextual factors, outside the subject matter of the legislation, such as who the sponsors are and what committee the bill is assigned to, are often more important than the subject of the legislation.

The `just_spon` model achieves an average accuracy of 76%, slightly outperforming `just_txt` with an improvement over baseline of 8%. This further indicates that knowing sponsor related information, without reference to the subject of the legislation, is itself highly predictive. In fact, Figure 4 shows that except for New Hampshire (upper), almost all states achieve gains using sponsor only information, with Oklahoma, Texas, and Ohio achieving gains of 30% or more. The committee information in `no_txt_spon`, which includes the sponsor committee positions, is even more predictive than sponsor and text only, and the addition of sponsors in `no_txt` improves performance by 3.5%.

Including text in the `combined` model further improves performance by

Figure 3. *Change from baseline with text only features.*

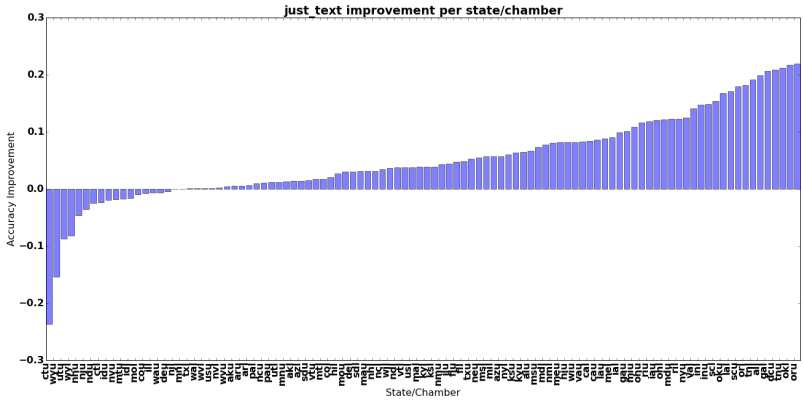
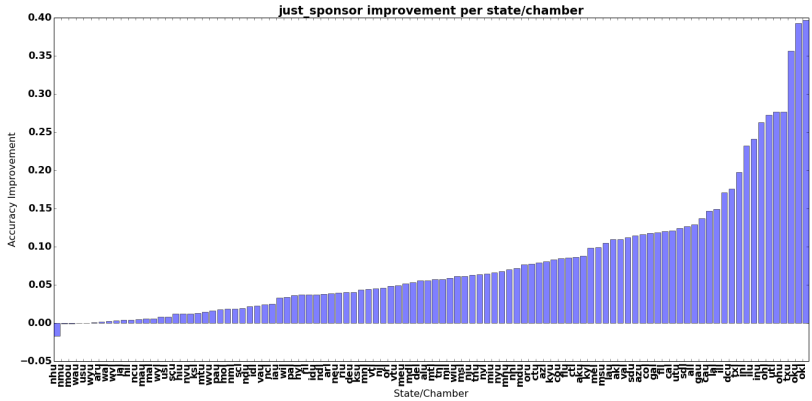


Figure 4. *Change from baseline with sponsor only features.*



1.3% over `no_txt`, and 18% over the majority class baseline, showing the complementary effects of contextual and lexical information, as this model consistently outperforms all others. Figure 5 shows the per state and chamber pair `baseline` and `combined` model performance. The AUROC performance follows a very similar trajectory.

On *LL*, the model performance follows a similar path, with all models showing improvement in probability estimates from the baseline. *LL* almost doubles from the `combined` model's 0.31 to 0.6 on `baseline`. This reinforces that the `combined` model makes very confident correct predictions. Including text in the `combined` improves performance slightly over `no_txt`, while having just sponsors or just text decreases the *LL* to around 0.5.

Table 4. *Average accuracy, log-loss, AUROC for bills using legislative events post introduction.*

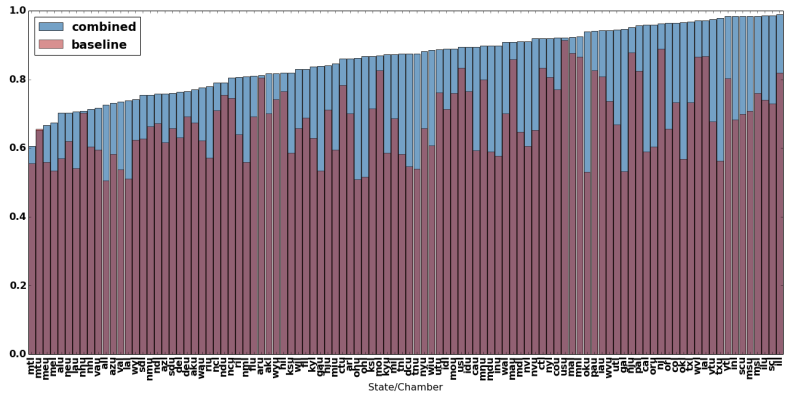
Feature Set	Accuracy		Log-Loss		AUROC	
	Ave	SD	Ave	SD	Ave	SD
`combined`	0.859	0.093	0.31	0.17	0.85	0.21
`combined+act`	**0.94**	0.059	**0.16**	0.12	**0.97**	0.04

Analysis

All contextual and lexical features considered above are available upon the introduction of a bill, or shortly thereafter,[13] thus the evaluation above indicates how well floor action can be predicted from the day of introduction. However, after the bill is introduced, subsequent legislative actions indicate further contextual information about the legislative process. As it is reasonable to assume

13. Some states do not indicate committee assignment immediately, for those we include the first assignment after introduction.

Figure 5. *Prediction accuracy on bills with combined model.*



these actions carry relevant predictive information, we further examine subsequent events in the legislative process in the `combined+act` feature set. We include a binary feature for the occurrence of amendment introduction and outcomes, votes, committee referral outcomes and readings up to the point of floor action. By comparing `combined+act` to `combined` we can examine how important different events in the legislative procedure are to predicting floor action.

Table 4 shows the results of the `combined+act` feature set. Accuracy improves to 0.94, while *LL* drops by half to 0.16, confirming that legislative events occurring up to the point point of floor action carry significant complementary information to other contextual factors and are highly indicative of floor action. While `combined+act` confirms the predictive power of procedural factors outside the legislative text, sponsor, and committee assignment, the `combined` model is arguably the most important result, as it indicates how well we can predict on features that are available upon introduction.

Beyond the predictions, we are interested in identifying the different features that contribute to legislative success across the states. As there are a large number of both models and features, in order to understand the relative predictive

importance of contextual and legislature specific dynamics, we choose several expert-informed factors deemed to be important for floor action, and compare the rank and weight they received in each model.

We first examine the median rank by model weight given to the following features in the `just_spon` condition across all states: bipartisan, sponsor in minority, sponsor in majority, and the number of sponsors. While many of these contextual features are highly ranked, there are many variations and differences across states. The top half of Table 5 shows the top ten states for which each feature was ranked among the top 20. For example, the bipartisan feature is ranked in the top 5 in Missouri, Virginia, Maine, and Mississippi, accounting for up to 6% of the explanatory power. As a comparison, in South Dakota, Hawaii, Minnesota, Wisconsin, and Pennsylvania, bipartisanship ranked lower than 200. Whether the sponsor is in the minority is important in the U.S. Congress, where it is ranked 6^{th}, along with Delaware, Tennessee, and West Virginia. Being in the majority accounts for 10% in Kentucky, and 7% in Wisconsin. This aligns with previous literature, as Wisconsin is known to have a strong party system Hamm 1980, and indeed we find sponsor in majority and in minority features to be ranked 1^{st} and 9^{th}, respectively, while in Texas, which has a much weaker party system, those features are ranked among the lowest of all states.

Similar ranking is presented for committee features in the bottom half of Table 5 in the `no_txt_spon` condition. The committee features play a similarly predictive role, with the sponsors holding membership positions on the committee accounting for over 10% of explanatory power in Delaware, Connecticut, Maine, and South Dakota.

To examine the difference in probability assigned by the models under different conditions, we can examine the probability of floor action our model assigns to legislation that received floor action and did not. We expect the median of the probabilities on legislation that received floor action to be higher than the median of the probabilities on legislation that failed, which is indeed the case. When comparing feature sets, we can compare how much the probability estimates are affected by different features. Taking a representative example where neither contextual nor lexical features dominate — Pennsylvania's lower chamber — the `combined` models median and mean predictions of success on bills receiving action are above 90%, and it has the largest difference between the probability of success assigned to passing bills as opposed to the probability of success assigned to failing bills, which is close to 0%. The `no_txt` model has

Table 5. *Median ranking by weight assigned across states for bipartisan, sponsor in minority, sponsor in majority, and number of sponsor features for sponsor only model, and having a ranking majority of the committee as a sponsor, not being a committee member as a sponsor, and being a member of the committee as a sponsor in the committee model. The top column indicates how many states have that feature ranked within the top 20 weighted features. Top States lists the ten states where each feature was ranked the highest and was one of the first 20 features. Bottom States lists the ten states where each feature was ranked the worst.*

Feature	Median	Top	Top States	Bottom States
Bipartisan	64	11	MO,VA,ME,MS,NC,SC, AK,DE,WA,US,NV	SD,MN,WI,PA,UT,NE, ID,FL,DC,AR
in Minority	24	20	DE,US,WV,TN,IA,WI, AL,ND,MD,MI,RI,NC, MT,OH,PA,MO,SC,CO, WA,NY	CA,IL,HI,TX,NJ,UT, NE,FL,DC,AR
in Majority	23	22	WI,MN,TN,KY,NH,NC, CO,IL,AL,OH,US,WV, GA,MI,MT,IA,ND,SC, DC,MO,PA,NY	AK,TX,MA,VA,NE,NJ, ME,UT,FL,AR
Sponsor	28	20	CO,UT,IL,VT,IN,IA,OR, SD,OH,US,ND,HI,CA, LA,NJ,MT,KS,MA,TN, NE	PA,AZ,WA,WY,NM,NV, VA,MS,MN,AR
Ranking Mbr	24	15	NE,VT,AR,KY,US,GA, OK,ME,NY,OR,SD,IL, MN,DC,WV	KS,MO,MT,OH,PA,RI, TN,UT,VA,WY
No Cmte Mbr	17	23	AR,ME,NE,IL,SD,NC, NV,NM,DE,KY,WV,CO, OK,CT,NY,TX,MD,RI, WI,VT,NJ,GA,MA	HI,ID,IA,KS,MT,OH, PA,TN,UT,WY
Members	6	33	DE,CT,ME,SD,NC,NV, OK,NY,GA,IL,KY,NJ, IN,WI,MD,WA,DC,MI, AR,OR,RI,TX,US,NE, CA,MN,NH,SC,MA,ND, WV,AL,NM	HI,ID,IA,KS,MT,OH,PA, TN,UT,WY

a similar mean, but the probabilities become more distributed on both pass and fail. Removing sponsors significantly affects the distribution, and shifts the mean lower to 70%. `just_spon` and `just_text` both drop the mean to around 40%.

Table 6. *Top and Bottom ranked phrases.*

State	Top Phrases	Bottom Phrases
New Mexico (upper)	day, campus, recognit, month, defin, alcohol, date, recipi, procur, cours, registr plate, revis,	tax credit chang, enmu, residenti, lobbi, statewid, or, abort, safeti, date for, test for, primari care, analysi,
New Mexico (lower)	day, studi, length, citi, of nm, fingerprint, geotherm, fund project, dog, definit, loan for, month,	of game fish, peac, senior citizen, math scienc, transfer of, state fair, self, bachelor, develop tax credit, nmhu, wolf, equip tax,
Pennsylvania (upper)	provid for alloc, creation of board, manufactur or, an appropri to, of applic and, medic examin, fiscal offic, for request for, corpor power, within the general, for the offic,	an act amend, as the tax, known the, act provid, known the tax, wage, act prohibit, citizen, of pennsylvania further, tax, youth, requir the depart,
Pennsylvania (lower)	or the, contract further, and for special, memori highwai, within the general, in game, first class township, whistleblow, emerg telephon, offens of sexual, for promulg,	act amend titl, an act amend, act provid, known, act prohibit, amend the, pennsylvania, an act provid, code of, act establish, an act relat, the constitut,
New York (upper)	fiscal year relat, memori highway, year relat to, implement the health, for retroact real, portion of state, the public protect, implement the public, inc to appli, budget author, program in relat, which are necessari	languag assist, direct the superintend, the develop of, author shall, subsidi, automobil insur, such elect, limit profit, disabl act, polici base, polici to provid
New York (lower)	care insur, applic for real, physic educ, fire district elect, establish credit, to file an, abolit or, hous program, the suspens of, are necessari to, the membership of, relat to hous	appropri, fuel and, numer, school ground, vehicular, incom tax for, prohibit public, tag, senat and assembl, on school, on school, class feloni

Finally, we examine language ranked most and least predictive on the `just_txt`

condition for New Mexico, Pennsylvania, and New York in Table 6.[14] Previous literature has proposed several theories on how content affects legislative passage, including that the more redistributive a policy is perceived, the higher in controversy, or the greater in scope, the lower the passage likelihood Rakoff and Sarner 1975; Hamm 1980. While each state has a unique set of issues that are likely to be taken to the floor, and conversely, to be left in committee, there is also evident overlap. In the top phrases, several states contain budgetary issues, expressed with fiscal and appropriation language, as most states have to pass budgetary measures. We also see commendation and procedural language, which is often less contentious. In the bottom phrases, several states have tax related language, and several education related topics.

While outside the scope of this work, in future work we hope to explore the differing language identified by the model to help identify important questions about the policymaking process in each state, and allow comparison within states of what successful legislation contains, and across states, of how different issues take shape. In addition, as we only included a limited amount of text, we would like to explore how to incorporate the full body text of legislation effectively.

Conclusion

In this chapter we explored the state legislative process by introducing the task of predicting floor action across all 50 states and the District of Columbia. We presented several baseline models and showed that combining contextual information about legislators and legislatures along with bill text consistently provides the best predictions, achieving an accuracy of 86% when predicting which legislation will reach the floor upon first introduction. We further analyzed various factors and their respective importance in the predictive models across the states, gaining a broader understanding of state legislative dynamics. While the factors

14. Addition states are presented in Table 8 in the Appendix.

that influence legislative floor action success are diverse and understandably inconsistent among states, by examining them we can empirically help elucidate the similarities and differences of the policymaking processes.

References

Bergstra, James S., Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. "Algorithms for Hyper-Parameter Optimization." In *Proceedings of NIPS.*

Bergstra, James, Dan Yamins, and David D. Cox. 2013. "Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms." In *Proceedings of the 12th Python in Science Conference.*

Bertrand, Marianne, Matilde Bombardini, Raymond Fisman, and Francesco Trebbi. 2018. *Tax-Exempt Lobbying: Corporate Philanthropy as a Tool for Political Influence.* NBER Working Papers. `http://www.nber.org/papers/w24451`.

Blei, David M., Andrew Ng, and Michael Jordan. 2003. "Latent Dirichlet Allocation." *JMLR* 3.

Bradley, Andrew P. 1997. "The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms." *Pattern Recogn.* (New York, NY, USA) 30, no. 7 (July): 1145–1159.

Breiman, Leo. 1996. "Stacked regressions." *Machine Learning* 24, no. 1 (July): 49–64.

Canfield-Davis, Kathy, Sachin Jain, Don Wattam, Jerry McMurtry, and Mike Johnson. 2010. "Factors of Influence on Legislative Decision Making: A Descriptive Study." *Journal of Legal, Ethical and Regulatory Issues* 13 (2).

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The statistical analysis of roll call data." *American Political Science Review* 98 (02): 355–370.

Cohen, Lauren, Karl B. Diether, and Christopher Malloy. 2012. *Legislating Stock Prices.* NBER Working Papers 18291. August. https://ideas.repec.org/p/nbr/nberwo/18291.html.

Fowler, James H. 2006. "Connecting the Congress: A Study of Cosponsorship Networks." *Political Analysis* 14 (4): 456–487. doi:10.1093/pan/mpl002.

Francis, Wayne L. 1989. *The Legislative Committee Game: A Compartative Analysis of Fifty States.* Columbus: Ohio State University Press.

Friedman, Jerome H. 2000. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29:1189–1232.

Gerrish, Sean, and David M. Blei. 2011. "Predicting Legislative Roll Calls from Text." In *Proceedigns of ICML.*

Gray, Virginia, and David Lowery. 1995. "Interest Representation and Democratic Gridlock." *Legislative Studies Quarterly* 20 (4): 531–552. http://www.jstor.org/stable/440192.

Hamm, Keith E. 1980. "U. S. State Legislative Committee Decisions: Similar Results in Different Settings." *Legislative Studies Quarterly* 5 (1): 31–54. http://www.jstor.org/stable/439440.

Hamm, Keith, Ronald Hedlund, and Nancy Miller. 2014. "State Legislatures." Chap. 13 in *The Oxford Handbook of State and Local Government,* edited by Donald Haider-Markel. Oxford University Press.

Harbridge, Laurel M. 2016. "Legislative Effectiveness in the United States Congress: The Lawmakers. By Craig Volden and Alan E. Wiseman." *The Journal of Politics* 78 (1).

Hedge, David. 1998. *Governance And The Changing American States.* New York: Routledge.

Hicks, William, and Daniel Smith. 2009. “Do Parties Matter? Explaining Legislative Productivity in the American States.” In *The State of the Parties: 2008 and Beyond Conference.*

Iyyer, Mohit, Peter Enns, Jordan L. Boyd-Graber, and Philip Resnik. 2014. “Political Ideology Detection Using Recursive Neural Networks.” In *Proceedings of ACL.*

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 1st. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Katz, Daniel Martin, Michael J. Bommarito, and Josh Blackman. 2017. “A General Approach for Predicting the Behavior of the Supreme Court of the United States.” *PLoS ONE 12.4.*

Kirilenko, Andrei, Shawn Mankad, and George Michailidis. 2014. “Do U.S. Regulators Listen to the Public? Testing the Regulatory Process with the RegRank Algorithm.” *Robert H. Smith School Research Paper.*

Kornilova, Anastassia, Daniel Argyle, and Vlad Eidelman. 2018. “Party Matters: Enhancing Legislative Embeddings with Author Attributes for Vote Prediction.” In *Proceedings of ACL.*

Lauderdale, Benjamin E., and Tom S. Clark. 2014. “Scaling Politically Meaningful Dimensions Using Texts and Votes.” *American Journal of Political Science* 58 (3): 754–771.

Linder, Fridolin, Bruce A. Desmarais, Matthew Burgess, and Eugenia Giraudy. 2018. “Text as Policy: Measuring Policy Similarity Through Bill Text Reuse.” *SSRN.*

Livermore, Michael A., Vladimir Eidelman, and Brian Grom. 2018. “Computationally Assisted Regulatory Participation.” *93 Notre Dame Law Review* 977.

Matthew, Hill, Kelly Wayne, Lockhart Brandon, and Ness Robert. 2013. "Determinants and Effects of Corporate Lobbying." *Financial Management* 42 (4).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed Representations of Words and Phrases and their Compositionality." In *Proceedings of NIPS*.

Nay, John J. 2016. "Predicting and Understanding Law-Making with Machine Learning." *CoRR* abs/1607.02109. arXiv: 1607.02109.

Nguyen, Viet-An, Jordan L. Boyd-Graber, Philip Resnik, and Kristina Miler. 2015. "Tea Party in the House: A Hierarchical Ideal Point Topic Model and Its Application to Republican Legislators in the 112th Congress." In *Proceedings of ACL*.

Niculescu-Mizil, Alexandru, and Rich Caruana. 2005. "Predicting Good Probabilities with Supervised Learning." In *Proceedings of ICML*. http://doi.acm.org/10.1145/1102351.1102430.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *J. Mach. Learn. Res.* 12 (November): 2825–2830.

Poole, Keith T, and Howard Rosenthal. 1985. "A spatial model for legislative roll call analysis." *American Journal of Political Science:* 357–384.

Poole, Keith T, and Howard L Rosenthal. 2007. *Ideology and Congress*. New Brunswick, NJ: Transaction Publishers.

Rakoff, Stuart H., and Ronald Sarner. 1975. "Bill History Analysis: A Probability Model of the State Legislative Process." *Polity* 7 (3): 402–414.

Rosenthal, Alan. 1974. *Legislative Performance in the States: Explorations of Committee Behavior*. New York: Free Press.

Shor, Boris, Christopher Berry, and Nolan McCarty. 2010. "A Bridge to Somewhere: Mapping State and Congressional Ideology on a Cross-institutional Common Space." *Legislative Studies Quarterly* 35 (3): 417–448.

Shor, Boris, and Nolan McCarty. 2011. "The ideological mapping of American legislatures." *American Political Science Review* 105 (03): 530–551.

Slapin, Jonathan B, and Sven-Oliver Proksch. 2008. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52 (3): 705–722.

Squire, Peverill. 2007. "Measuring State Legislative Professionalism: The Squire Index Revisited." *State Politics & Policy Quarterly* 7 (2).

Talbert, Jeffery C., and Matthew Potoski. 2002. "Setting the Legislative Agenda: The Dimensional Structure of Bill Cosponsoring and Floor Voting." *The Journal of Politics* 64 (3).

Thomas, Matt, Bo Pang, and Lillian Lee. 2006. "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts." In *Proceedings of EMNLP.*

Wang, Sida I., and Christopher D. Manning. 2012. "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification." In *ACL (2),* 90–94. The Association for Computer Linguistics.

Yano, Tae, Noah A. Smith, and John D. Wilkerson. 2012. "Textual Predictors of Bill Survival in Congressional Committees." In *Proceedings of NAACL.*

Appendix

The first two models are linear classifiers, where the prediction of floor action, $\hat{y}_i$, is given by $\texttt{sign}(\mathbf{w}^\top \mathbf{f}(\boldsymbol{x}))$. The first is a regularized conditional log-linear model $p_{\mathbf{w}}(y \,|\, \boldsymbol{x})$:

$$p_{\mathbf{w}}(y \,|\, \boldsymbol{x}) = \frac{\exp\{\mathbf{w}^\top \mathbf{f}(\boldsymbol{x})\}}{Z(\boldsymbol{x})} \tag{1}$$

where $Z(\boldsymbol{x})$ is the partition function given by $\sum_y \exp\{\mathbf{w}^\top \mathbf{f}(\boldsymbol{x})\}$. The model optimizes $\mathbf{w}$ according to

$$\min_{\mathbf{w}} \sum_{i}^{n} \log p_{\mathbf{w}}(y_i \,|\, \boldsymbol{x}_i) + \lambda \|\mathbf{w}\| \tag{2}$$

The second model is NBSVM (Wang and Manning 2012), an interpolation between multinomial Naive Bayes and a support vector machine, which optimizes $\mathbf{w}$ according to:

$$\min_{\mathbf{w}} C \sum_{i}^{n} \max(0, 1 - y_i(\mathbf{w}^\top(\mathbf{f}(\boldsymbol{x}_i) \circ \mathbf{r})))^2 + \|\mathbf{w}\|^2 \tag{3}$$

where $\mathbf{r}$ is the log-count ratio of features occurring in positive and negative examples. The third model is non-linear, in the form of a tree-based gradient boosted machine (Friedman 2000), which optimizes $\mathbf{w}$ according to:

$$\min_{\mathbf{w}} \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(\mathbf{t}_k) \tag{4}$$

where K is the number of trees, l is the loss function (typically binomial deviance) and $\hat{y}_i$ is given by $\sum_{k=1}^{K} \mathbf{t}_k(\boldsymbol{x}_i)$ where $\mathbf{t}_k$ is a tree.

Log-loss, LL is defined as:

$$LL = -\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i = \hat{y}_i) \log(p_i) + (1 - \mathbb{1}(y_i = \hat{y}_i) \log(1 - p_i) \tag{5}$$

where $\mathbb{1}(y_i = \hat{y}_i)$ is an binary indicator function equaling 1 if the model prediction $\hat{y}_i$ was correct, and 0 otherwise. LL equals zero for a perfect classifier, and

increases with worse probability estimates. Specifically, LL penalizes models more, the more confident they are in an incorrect classification.

AUROC allows us to measure the relationship between a model's true positive (TP), how many floor action bills were correctly predicted as floor action, and false positive rate (FP) — how many failed bills were predicted as floor action. It is defined by:

$$AUROC = \sum_{i=1}^{N} p(TP)\Delta p(FP) + \frac{1}{2}(\Delta p(TP)\Delta p(FP)) \tag{6}$$

Table 7. *Data statistics for number of bills introduced and receiving floor action for each state.*

State	Bills			Resolutions			
	Floor Action	Introduced	Rate	Floor Action	Introduced	Rate	Sessions
al	6697	14327	0.467	898	1201	0.748	16
ak	781	2527	0.309	372	609	0.611	4
az	3719	9308	0.4	367	868	0.423	24
ar	3809	5076	0.75	156	330	0.473	8
ca	18978	32143	0.59	2141	2554	0.838	17
co	4808	6428	0.748	876	1014	0.864	11
ct	3044	16236	0.187	1258	1816	0.693	8
de	3185	4858	0.656	1262	1386	0.911	6
dc	8515	15593	0.546	11016	18422	0.598	9
fl	6592	21298	0.31	377	1177	0.32	15
ga	7416	15379	0.482	20913	23283	0.898	13
hi	5630	21615	0.26	1256	4196	0.299	5
id	3259	4446	0.733	379	537	0.706	8
il	14106	66926	0.211	19559	22884	0.855	10
in	1958	5291	0.37	-	-	-	4
ia	3434	21457	0.16	874	1569	0.557	7
ks	2123	6324	0.336	889	1039	0.856	6
ky	3149	8185	0.385	4923	6032	0.816	20
la	18346	35277	0.52	13231	14584	0.907	32
me	9268	17095	0.542	-	-	-	8
md	9857	26125	0.377	44	170	0.259	14
ma	6862	52467	0.131	-	-	-	7
mi	14520	41730	0.348	4434	10235	0.433	11

mn	4494	27240	0.165	169	1083	0.156	10
ms	6621	25450	0.26	4577	6234	0.734	21
mo	2736	14143	0.193	224	1244	0.18	8
mt	5910	9905	0.597	761	942	0.808	8
ne	1837	4829	0.38	1358	2422	0.561	6
nv	2614	4163	0.628	276	337	0.819	7
nh	3243	6793	0.477	47	202	0.233	6
nj	6900	59861	0.115	1248	7059	0.177	8
nm	4253	10909	0.39	13	50	0.26	8
ny	23071	89072	0.259	30216	31346	0.964	4
nc	6922	25152	0.275	813	1309	0.621	10
nd	5735	8089	0.709	628	886	0.709	9
oh	4356	8605	0.506	5080	8706	0.584	9
ok	16579	36827	0.45	2601	4004	0.65	10
or	5240	14404	0.364	475	1009	0.471	12
pa	2887	16414	0.176	5049	6001	0.841	5
ri	6596	16584	0.398	3400	3859	0.881	5
sc	3269	11532	0.283	9575	10800	0.887	6
sd	1647	2539	0.649	13	530	0.025	9
tn	33936	77331	0.439	28008	29732	0.942	12
tx	8371	25771	0.325	19321	20607	0.938	9
ut	7816	11072	0.706	436	620	0.703	23
vt	1035	4520	0.229	2247	2353	0.955	4
va	14215	27813	0.511	3678	12562	0.293	24
wa	8317	24578	0.338	1501	2038	0.737	6
wv	4308	23917	0.18	2496	4297	0.581	12
wi	4982	13761	0.362	-	-	-	14
wy	2825	4223	0.669	102	185	0.551	11
us	22973	172921	0.133	6440	31067	0.207	15

Table 8. *Top and bottom ranked phrases for New Jersey, Maryland, California, and Florida.*

State	Top Phrases	Bottom Phrases
New Jersey (upper)	for farmland preserv, preserv trust, green acr fund, acquisit and, mmvv million from, vehicl from, budget for, fund for state, in feder fund, unemploy, for state acquisit, infrastructur trust	retir benefit for, of educ for, school board member, clarifi law, contract and, tax reimburs program, appropri mmvv for, to develop and, tax rate, credit under corpor, certain vehicl

New Jersey (lower)	environment infrastructur, to dissemin, farm to, dmva to, concern certain, and dhs, unsolicit, atm, contract law, link to, manufactur rebat, limit liabil	polit, import, all school, for water, facil to be, respons for, grant program for, relat crime, state administ, from tax, chair, to all
Maryland (upper)	financ the construct, festiv licens, issu the licens, grante provid and, to effect, advisori commiss, an evalu of, that provis of, financ statement, board licens, to borrow, defer	not to, phase, be use as, use as, facil locat in, to own, law petit, or expenditur, trust establish, expend match, and expend match
Maryland (lower)	charl counti alcohol, improv or, to financ the, termin provis relat, counti sale, sanction, alter, counti special tax, montgomeri counti alcohol, report requir repeal, length, licens mc	grant to the, creation state debt, educ fund, state debt baltimor, establish the amount, elimin, disclos to, propos amend, incom tax rate, purpos relat, crimin gang, deced die after
California (upper)	ab, revolv fund, household, intent that, these provis until, restitut, counsel, employe, onli if ab, properti, if ab, would incorpor addit	legisl, cost of, veterinari, enact legisl, to the, law, regul econom, governor, incom tax deduct, hour, motor vehicl recreat, decis
California (lower)	add articl, to amend repeal, bill would incorpor, to add and, budget act of, urgenc statut, make nonsubstant, and make, as bill provid, relat the budget, and of the, ab	would make nonsubstant, enact legisl, make technic nonsubstant, would make technic, unspecifi, code to add, baccalaur degre, salari, fraud prevent, flexibl, of the state, would
Florida (upper)	ogsr, abrog provis relat, grant trust fund, govern act, person inform, to supplement, employ contribut to, legisl audit committe, jac, maintain by the, insur regul, financi inform	senat relat to, senat relat, to, ssb, elder, school, municip that, and legislatur by, that law enforc, provid minimum, admiss to, local law enforc
Florida (lower)	etc, certain propos, re creat, repeal under, to qualifi, boundari, program revis requir, environment permit, counti hospit district, alcohol beverag licens, except under, ranch	hous relat, day, renew energi, provid for alloc, make recommend, for employ of, of damag, from particip, week, catastroph, dhsmv to develop, employ from

Delaware (upper)	uniform, would increas the, amend chapter volum, person convict, relat the delawar, dealer, child support, bureau, violenc, associ, charter chang, for fiscal year	rent, state languag, the content, act regul, certain licens, give local, assembl from, delawar code establish, for citizen, reimburs, propos constitut amend, salari
Delaware (lower)	of the th, tax refund, thi act also, amend of the, this section of, the titl, the act to, of member of, electron transmiss, for in the, and the date, parent guardian	predatori, hour per, relat state employe, unfair practic, communic, open meet, equal the, to the construct, the construct, medicaid, state agenc, relat to prevail