Textual Predictors of Bill Survival in Congressional Committees

Tae Yano **Noah A. Smith**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{taey,nasmith}@cs.cmu.edu

John D. Wilkerson
Department of Political Science
University of Washington
Seattle, WA 98195, USA
jwilker@u.washington.edu

Abstract

A U.S. Congressional bill is a textual artifact that must pass through a series of hurdles to become a law. In this paper, we focus on one of the most precarious and least understood stages in a bill's life: its consideration, behind closed doors, by a Congressional committee. We construct predictive models of whether a bill will survive committee, starting with a strong, novel baseline that uses features of the bill's sponsor and the committee it is referred to. We augment the model with information from the *contents* of bills, comparing different hypotheses about how a committee decides a bill's fate. These models give significant reductions in prediction error and highlight the importance of bill substance in explanations of policy-making and agenda-setting.

1 Introduction

In representative governments, laws result from a complex social process. Central to that process is language. Text data emerging from the process include debates among legislators (Laver et al., 2003; Quinn et al., 2010; Beigman Klebanov et al., 2008), press releases (Grimmer, 2010), accounts of these debates in the press, policy proposals, and laws.

In the work reported here, we seek to exploit text data—specifically, the text of Congressional bills—to understand the lawmaking process. We consider an especially murky part of that process that is difficult to study because it happens largely behind closed doors: the handling of bills by Congressional committees. This early stage of a bill's life is precarious: roughly 85% of bills do not survive committee. By contrast, nearly 90% of bills that are recommended by a committee (i.e., survive the committee and are introduced for debate on the floor) will survive a roll call vote by the legislature. Because filtering by these powerful Congressional committees is both more opaque and more selective than the actions of the legislature as a whole, we believe that text-based models can play a central role in understanding this stage of lawmaking.

This paper's contributions are: (i) We formulate computationally the prediction of which bills will a survive Congressional committee, presenting a (baseline) model based on observable features associated with a bill, the committee(s) it is assigned to, members of that committee, the Congress as a whole, and expert combinations of those features. The task formulation and baseline model are novel. (ii) We propose several extensions of that strong baseline with information derived from the text of a bill. (iii) We validate our models on a hard predictive task: predicting which bills will survive committee. Text is shown to be highly beneficial. (iv) We present a discussion of the predictive features selected by our model and what they suggest about the underlying political process. (v) We release our corpus of over 50,000 bills and associated metadata to the research community for further study.[1]

We give brief background on how bills become U.S. laws in §2. We describe our data in §3. The modeling framework and baseline are then introduced (§4), followed by our text-based models with experiments (§5), then further discussion (§6).

[1] http://www.ark.cs.cmu.edu/bills

2 How Bills Become Laws

In the U.S., federal laws are passed by the U.S. Congress, which consists of two "chambers," the House of Representatives (commonly called the "House") and the Senate. To become law, a bill (i.e., a proposed law) must pass a vote in both chambers and then be signed by the U.S. President. If the President refuses to sign a bill (called a "veto"), it may still become law if both chambers of Congress overrides the veto through a two-thirds majority.

Much less discussed is the process by which bills come into existence. A bill is formally proposed by a member of Congress, known as its sponsor. Once proposed, it is routed to one or more (usually just one) of about twenty subject-specializing committees in each chamber. Unlike floor proceedings, transcripts of the proceedings of Congressional committees are published at the discretion of the committee and are usually publicly unavailable.

Each committee has a chairman (a member of the majority party in the chamber) and is further divided into subcommittees. Collectively a few thousand bills per year are referred to Congress' committees for consideration. Committees then recommend (report) only about 15% for consideration and voting by the full chamber.

The U.S. House is larger (435 voting members compared to 100 in the Senate) and, in recent history, understood to be more polarized than the Senate (McCarty et al., 2006). All of its seats are up for election every two years. A "Congress" often refers to a two-year instantiation of the body with a particular set of legislators (e.g., the 112th Congress convened on January 3, 2011 and adjourns on January 3, 2013). In this paper, we limit our attention to bills referred to committees in the House.

3 Data

We have collected the text of all bills introduced in the U.S. House of Representatives from the 103rd to the 111th Congresses (1/3/1993–1/3/2011). Here we consider only the version of the bill as originally introduced. After introduction, a bill's title and contents can change significantly, which we ignore here.

These bills were downloaded directly from the Library of Congress's Thomas website.[2] Information about the makeup of House committees was obtained from Charles Stewart's resources at MIT,[3] while additional sponsor and bill information (e.g., sponsor party affiliation and bill topic) was obtained from E. Scott Adler and John Wilkerson's Congressional Bills Project at the University of Washington.[4]

Cong.	Maj.	Total Introduced	Survival Rate (%) Total	Rep.	Dem.
103	Dem.	5,311	11.7	3.4	16.2
104	Rep.	4,345	13.7	19.7	6.1
105	Rep.	4,875	13.2	19.0	5.4
106	Rep.	5,682	15.1	20.9	7.0
107	Rep.	5,768	12.1	17.5	5.8
108	Rep.	5,432	14.0	21.0	5.9
109	Rep.	6,437	11.8	16.9	5.1
110	Dem.	7,341	14.5	8.5	18.0
111	Dem.	6,571	12.6	8.1	14.5
Total		51,762	13.2	15.9	10.7

Table 1: Count of introduced bills per Congress, along with survival rate, and breakdown by the bill sponsor's party affiliation. Note that the probability of survival increases by a factor of 2–5 when the sponsor is in the majority party. Horizontal lines delineate presidential administrations (Clinton, Bush, and Obama).

In our corpus, each bill is associated with its title, text, committee referral(s), and a binary value indicating whether or not the committee reported the bill to the chamber. We also extracted metadata, such as sponsor's name, from each bill's summary page provided by the Library of Congress.

There were a total of 51,762 bills in the House during this seventeen-year period, of which 6,828 survived committee and progressed further. See Table 1 for the breakdown by Congress and party.

In this paper, we will consider a primary train-test split of the bills by Congress, with the 103rd–110th Congresses serving as the training dataset and the 111th as the test dataset. This allows us to simulate the task of "forecasting" which bills will survive in a future Congress. In §5.5, we will show that a similar result is obtained on different data splits.

These data are, in principle, "freely available" to the public, but they are not accessible in a uni-

[2] http://thomas.loc.gov/home/thomas.php

[3] http://web.mit.edu/17.251/www/data_page.html

[4] http://congressionalbills.org

fied, structured form. Considerable effort must be expended to align databases from a variety of sources, and significant domain knowledge about the structure of Congress and its operation is required to disambiguate the data. Further exploration of the deeper relationships among the legislators, their roles in past Congresses, their standing with their constituencies, their political campaigns, and so on, will require ongoing effort in joining data from disparate sources.

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.

We view text content—noisy and complex as it is—as an attractive alternative, or at least a complementary information source. Though unstructured, text is made up of features that are relatively easy for humans to interpret, offering a way to not only predict, but also explain legislative outcomes.

4 A Predictive Model

We next consider a modeling framework for predicting bill survival or death in committee. We briefly review logistic regression models (section 4.1), then turn to the non-textual features that form a baseline and a starting point for the use of text (section 4.2).

4.1 Modeling Framework

Our approach to predicting a bill's survival is logistic regression. Specifically, let X be a random variable associated with a bill, and let $\mathbf{f}$ be a feature vector function that encodes observable features of the bill. Let Y be a binary random variable corresponding to bill survival ($Y = 1$) or death ($Y = 0$). Let:

$$p_{\mathbf{w}}(Y = 1 \mid X = x) = \frac{\exp \mathbf{w}^\top \mathbf{f}(x)}{1 + \exp \mathbf{w}^\top \mathbf{f}(x)} \qquad (1)$$

where $\mathbf{w}$ are "weight" parameters associating each feature in the feature vector $\mathbf{f}(x)$ with each outcome. This leads to the predictive rule:

$$\hat{y}(x) = \begin{cases} 1 & \text{if } \mathbf{w}^\top \mathbf{f}(x) > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

We train the model by maximizing log-likelihood plus a a sparsity-inducing log-prior that encourages many weights to go to zero:

$$\max_{\mathbf{w}} \sum_i \log p_{\mathbf{w}}(y_i \mid x_i) - \lambda \|\mathbf{w}\|_1 \qquad (3)$$

where i indexes training examples (specifically, each training instance is a bill referred to a single committee). The second term is an ℓ_1 norm, equivalent to a Laplacian prior on the weights. The value of λ, which controls sparsity, is chosen on a held-out subset of the training data.

Linear models like this one, commonly called "exponential" or "max ent" models, are attractive because they are intelligible. The magnitude of a weight indicates a feature's importance in the prediction, and its sign indicates the direction of the effect.

We note that the ℓ_1 regularizer is not ideal for identifying predictive features. When two features are strongly correlated, it tends to choose one of them to include in the model and eliminate the other, despite the fact that they are both predictive. It is therefore important to remember that a weight of zero does not imply that the corresponding feature is unimportant. We chose to cope with this potential elimination of good features so that our models would be compact and easily interpretable.

4.2 Features

In American politics, the survival or death of many bills can be explained in terms of expertise, entrepreneurship, and procedural control, which are manifest in committee membership, sponsor attributes, and majority party affiliation. We therefore begin with a strong baseline that includes features encoding many expected effects on bill success. These include basic structural features and some interactions.

The basic features are all binary. The value of the random variable X includes information about the bill, its sponsor, and the committee to which the bill is referred. In addition to a bias feature (always equal to 1), we include the following features:

1. For each party p, is the bill's sponsor affiliated with p?
2. Is the bill's sponsor in the same party as the committee chair? Equivalently, is the bill's sponsor in the majority party of the House?
3. Is the bill's sponsor a member of the committee?

4. Is the bill's sponsor a *majority* member of the committee? (This feature conjoins 2 and 3.)
5. Is the bill's sponsor the chairman of the committee?

6. For each House member j, did j sponsor the bill?
7. For each House member j, is the bill sponsored by j and referred to a committee he chairs? (This feature conjoins 5 and 6.)
8. For each House member j, is the bill sponsored by j and is j in the same party as the committee chair? (This feature conjoins 2 and 6.)
9. For each state s, is the bill's sponsor from s?
10. For each month m, is the bill introduced during m?

11. For $v \in \{1, 2\}$, is the bill introduced during the vth year of the (two-year) Congress?

The features above were engineered in preliminary model development, before text was incorporated.[5]

4.3 Experiment

Performance. Considering the 111th Congress as a test set (6,571 instances), a most-frequent-class predictor (i.e., a constant prediction that no bill will survive committee) achieves an error rate of 12.6% (more details in Table 3). A model trained on the 103rd–110th Congresses (45,191 bills) contains 3,731 instantiated features above achieved 11.8% error (again, see Table 3).

Discussion. When inspecting linear models, considering feature weights can be misleading, since (even with regularization) large weights often correspond to small effects in the training data. Our methodology for inspecting models is therefore as follows: we calculate the *impact* of each feature on the final decision for class y, defined for feature j as

$$\frac{w_j}{N} \sum_{i=1}^{N} f_j(x_i) \quad (4)$$

where i indexes test examples (of which there are N). Impact is the average effect of a feature on the model's score for class y. Note that it is not affected by the true label for an example. Impact is additive, which allows us to measure and compare the influence of sets of features *within a model* on model predictions. Impact is not, however, directly comparable *across* models.

[5] One surprisingly detrimental feature, omitted here, was the identity of the committee. Bill success rates vary greatly across committees (e.g., Appropriations recommends about half of bills, while Ways and Means only 7%). We suspect that this feature simply has poor generalization ability across Congresses. (In §5.2 we will consider preferences of *individuals* on committees, based on text, which appears to benefit predictive performance.)

Bill Survival	
sponsor is in the majority party (2)	0.525
sponsor is in the majority party and on the committee (4)	0.233
sponsor is a Democrat (1)	0.135
sponsor is on the committee (3)	0.108
bill introduced in year 1 (11)	0.098
sponsor is the referred committee's chair (5)	0.073
sponsor is a Republican (1)	0.069
Bill Death	
bill's sponsor is from NY (9)	-0.036
sponsor is Ron Paul (Rep., TX) (6)	-0.023
bill introduced in December (10)	-0.018
sponsor is Bob Filner (Dem., CA) (6)	-0.013

Table 2: Baseline model: high-impact features associated with each outcome and their impact scores (eq. 4).

The highest impact features are shown in Table 2. Unsurprisingly, the model's predictions are strongly influenced (toward survival) when a bill is sponsored by someone who is on the committee and/or in the majority party. Feature 2, the sponsor being on the committee, accounted for nearly 27% of all (absolute) impact, followed by the member-specific features (6–8, 19%), the sponsor being in the majority and on the committee (4, 12%), and the party of the sponsor (1, 10%).

We note that impact as a tool for interpreting models has some drawbacks. If a large portion of bills in the test set happen to have a particular feature, that feature may have a high impact score for the dominant class (death). This probably explains the high impact of "sponsor is a Democrat" (Table 2); Democrats led the 111th Congress, and introduced more bills, most of which died.

5 Adding Text

We turn next to the use of text data to augment the predictive power of our baseline model. We will propose three ways of using the title and/or text of a bill to create features. From a computational perspective, each approach merely augments the baseline model with features that may reduce predictive

errors—our measure of the success of the hypothesis. From a political science perspective, each proposal corresponds to a different explanation of how committees come to decisions.

5.1 Functional Bill Categories

An important insight from political science is that bills can be categorized in general ways that are related to their likelihood of success. In their study on legislative success, Adler and Wilkerson (2005) distinguish Congressional bills into several categories that capture bills that are on the extremes in terms of the importance and/or urgency of the issue addressed. We expect to find that distinguishing bills by their substance will reduce prediction errors.

- bills addressing **trivial** issues, such as those naming a federal building or facility or coining commemorative medals;
- bills that make **technical** changes to existing laws, usually at the request of the executive agency responsible for its implementation;
- bills addressing **recurring** issues, such as annual appropriations or more sporadic reauthorizations of expiring federal programs or laws; and
- bills addressing **important**, urgent issues, such as bills introduced in response to the 9/11 terrorist attacks or a sharp spike in oil prices.

Adler and Wilkerson (2005) annotated House bills for the 101st–105th Congresses using the above categories (all other bills were deemed to be "discretionary"). Out of this set we use the portion that overlaps with our bill collection (103rd–105th). Of 14,528 bills, 1,580 were labeled as trivial, 119 as technical, 972 as recurring, and 1,508 as important. Our hypothesis is that these categories can help explain which bills survive committees.

To categorize the bills in the other Congresses of our dataset, we trained binary logistic regression models to label bills with each of the three most frequent bill types above (trivial, recurring, and important) based on unigram features of the body of bill text. (There is some overlap among categories in the annotated data, so we opted for three binary classifiers rather than multi-class.) In a ten-fold cross-validated experiment, this model averaged 83% accuracy across the prediction tasks. We used the manually annotated labels for the bills in the 103rd–105th Congresses; for other bills, we calculated each model's probability that the bill belonged to the target category.[6] These values were used to define binary indicators for each classifier's probability regions: $[0, 0.3)$; $[0.3, 0.4)$; $[0.4, 0.5)$; $[0.5, 1.0]$. For each of the three labels, we included two classifiers trained with different hyperparameter settings, giving a total of 24 additional features. All baseline features were retained.

Performance. Including functional category features reduces the prediction error slightly but significantly relative to the baseline (just over 1% relative error reduction)—see Table 3.[7]

Discussion. Considering the model's weights, the log-odds are most strongly influenced toward bill success by bills that seem "important" according to the classifiers. 55% of this model's features had non-zero impact on test-set predictions; compare this to only 36% of the baseline model's features.[8] Further, the category features accounted for 66% of the total (absolute) impact of all features. Taken altogether, these observations suggest that bill category features are a more compact substitute for many of the baseline features,[9] but that they do not offer much additional predictive information beyond the baseline (error is only slightly reduced). It is also possible that our categories do not perfectly capture the perceptions of committees making decisions about bills. Refinement of the categories within the pre-

[6] In preliminary experiments, we used the 103rd–105th data to measure the effect of automatic vs. manual categories. Though the particulars of the earlier model and the smaller dataset size make controlled comparison impossible, we note that gold-standard annotations achieved 1–2% lower absolute error across cross-validation folds.

[7] We note that preliminary investigations conjoining the bill category features with baseline features did not show any gains. Prior work by Adler and Wilkerson (2012) suggests that bill category interacts with the sponsor's identity, but does not consider bill success prediction; we leave a more careful exploration of this interaction in our framework to future work.

[8] Note that ℓ_1-regularized models make global decisions about which features to include, so the new features influence which baseline features get non-zero weights. Comparing the absolute number of features in the final selected models is not meaningful, since it depends on the hyperparameter λ, which is tuned separately for each model.

[9] This substitutability is unsurprising in some scenarios; e.g., successful reauthorization bills are often sponsored by committee leadership.

	Model	Error (%)	False +	False –	True +	# Feats.	Size	Effective
	most frequent class	12.6	0	828	0	–	–	–
§4.2	baseline (no text)	11.8	69	709	119	3,731	1,284	460
§5.1	bill categories	11.7	52	716	112	3,755	274	152
§5.2	proxy vote, chair only	10.8	111	596	232	3,780	1,111	425
	proxy vote, majority	11.3	134	606	222	3,777	526	254
	proxy vote, whole committee	10.9	123	596	232	3,777	1,131	433
	proxy vote, all three	10.9	110	606	222	3,872	305	178
§5.3	unigram & bigram	9.8	106	541	287	28,246	199	194
§5.4	full model (all of the above)	9.6	120	514	314	28,411	1,096	1,069

Table 3: Key experimental results; models were trained on the 103rd–110th Congresses and tested on the 111th. Baseline features are included in each model listed below the baseline. "# Feats." is the total number of features available to the model; "Size" is the number of features with non-zero weights in the final selected sparse model; "Effective" is the number of features with non-zero impact (eq. 4) on test data. Each model's improvement over the baseline is significant (McNemar's test, $p < 0.0001$ except bill categories, for which $p < 0.065$).

dictive framework we have laid out here is left to future research.

5.2 Textual Proxy Votes

We next consider a different view of text: as a means of profiling the preferences and agendas of legislators. Our hypothesis here is that committees operate similarly to the legislature as a whole: when a bill comes to a committee for consideration, members of the committee vote on whether it will survive. Of course, deliberation and compromise may take place before such a vote; our simple model does not attempt to account for such complex processes, instead merely positing a hidden roll call vote.

Although the actions of legislators on committees are hidden, their voting behavior on the floor is observed. Roll call data is frequently used in political science to estimate *spatial* models of legislators and legislation (Poole and Rosenthal, 1985; Poole and Rosenthal, 1991; Jackman, 2001; Clinton et al., 2004). These models help visualize politics in terms of intuitive, low-dimensional spaces which often correspond closely to our intuitions about "left" and "right" in American politics. Recently, Gerrish and Blei (2011) showed how such models could naturally be augmented with models of text. Such models are based on *observed* voting; it is left to future work to reduce the dimensionality of *hidden* votes within the survival prediction model here.

Our approach is to construct a *proxy vote*; an estimate of a roll call vote by members of the committee on the bill. We consider three variants, each based on the same estimate of the individual committee members' votes:

- Only the committee chairman's vote matters.
- Only majority-party committee members vote.
- All committee members vote.

We will compare these three versions of the proxy vote feature experimentally, but abstractly they can all be defined the same way. Let $\mathcal{C}$ denote the set of committee members who can vote on a bill x. Then the proxy vote equals:

$$\frac{1}{|\mathcal{C}|} \sum_{j \in \mathcal{C}} \mathbb{E}[V_{j,x}] \quad (5)$$

(If x is referred to more than one committee, we average the above feature across committees.) We treat the vote by representative j on bill x as a binary random variable $V_{j,x}$ corresponding to a vote for (1) or against (0) the bill. We do not observe $V_{j,x}$; instead we estimate its expected value, which will be between 0 and 1. Note that, by linearity of expectation, the sum in equation 5 is the expected value of the number of committee members who "voted" for the bill; dividing by $|\mathcal{C}|$ gives a value that, if our estimates are correct, should be close to 1 when the bill is likely to be favored by the committee and 0 when it is likely to be disfavored.

To estimate $\mathbb{E}[V_{j,x}]$, we use a simple probabilistic model of $V_{j,x}$ given the bill x and the past voting record of representative j.[10] Let $\mathcal{R}_j$ be a set of

[10] We note that the observable roll call votes on the floor of

bills that representative j has publicly voted on, on the floor of the House, in the past.[11] For $x \in \mathcal{R}_j$, let $V_{j,x}$ be 1 if j voted for the bill and 0 if j voted against it. Further, define a similarity measure between bills; here we use cosine similarity of two bills' tfidf vectors.[12] We denote by $sim(x, x')$ the similarity of bills x and x'.

The probabilistic model is as follows. First, the representative selects a bill he has voted on previously; he is likely to choose a bill that is similar to x. More formally, given representative j and bill x, randomly choose a bill X' from $\mathcal{R}_j$ according to:

$$p(X' = x' \mid j, x) = \frac{\exp sim(x,x')}{\sum_{x'' \in \mathcal{R}_j} \exp sim(x,x'')} \quad (6)$$

An attractive property of this distribution is that it has no parameters to estimate; it is defined entirely by the text of bills in $\mathcal{R}_j$. Second, the representative votes on x identically to how he voted on X'. Formally, let $V_{j,x} = V_{j,x'}$, which is observed.

The above model gives a closed form for the expectation of $V_{j,x}$:

$$\mathbb{E}[V_{j,x}] = \sum_{x' \in \mathcal{R}_j} p(X' = x' \mid j, x) \cdot V_{j,x'} \quad (7)$$

In addition to the proxy vote score in eq. 5, we calculate a similar expected vote based on "nay" votes, and consider a second score that is the ratio of the "yea" proxy vote to the "nay" proxy vote. Both of these scores are continuous values; we quantize them into bins, giving 141 features.[13]

Performance. Models built using the baseline features plus, in turn, each of the three variations of the proxy vote feature ($\mathcal{C}$ defined to include the chair only, majority party members, or the full committee), and *all* three sets of proxy vote features, were compared—see Table 3. All three models showed improvement over the baseline. Using the chairman-only committee (followed closely by whole committee and all three) turned out to be the best performing among them, with a 8% relative error reduction.

Discussion. Nearly 58% of the features in the combined model had non-zero impact at test time, and 38% of total absolute impact was due to these features. Comparing the performance of these four models suggests that, as is widely believed in political science, the preferences of the committee chair are a major factor in which bills survive.

5.3 Direct Use of Content: Bag of Words

Our third hypothesis is that committees make collective decisions by considering the contents of bills directly. A sensible starting point is to treat our model as a document classifier and incorporate standard features of the text *directly* into the model, rather than deriving functional categories or proxy votes from the text.[14] Perhaps unsurprisingly, this approach will perform better than the previous two.

Following Pang and Lee (2004), who used word and bigram features to model an author's sentiment, and Kogan et al. (2009), who used word and bigram features to directly predict a future outcome, we incorporate binary features for the presence or absence of terms in the body and (separately) in the title of the bill. We include unigram features for the body and unigram and bigram features for the title.[15] The result is 28,246 features, of which 24,515 are lexical.

Performance. Combined with baseline features, word and bigram features led to nearly 18% relative error reduction compared to the baseline and 9% relative to the best model above (Table 3). The model is very small (under 200 features), and 98% of the features in the model impacted test-time predictions. The model's gain over the baseline is not sensitive to the score threshold; see Figure 1.

A key finding is that the bag of words model out-

the U.S. House consist of a very different sample of bills than those we consider in this study; indeed, votes on the floor correspond to bills that *survived* committee. We leave attempts to characterize and control for this bias to future work.

[11] To simplify matters, we use all bills from the training period that j has voted on. For future predictions (on the test set), these are all in the past, but in the training set they may include bills that come later than a given training example.

[12] We first eliminated punctutation and numbers from the texts, then removed unigrams which occured in more than 75% or less than 0.05% of the training documents. Tfidf scores were calculated based on the result.

[13] We discretized the continuous values by 0.01 increment for proxy vote score, and 0.1 increment for proxy vote rate scores. We further combined outlier bins (one for exremely large values, one for extremely small values).

[14] The models from §5.1 and §5.2 can be understood from a machine learning perspective as task-specific dimensionality reduction methods on the words.

[15] Punctuation marks are removed from the text, and numbers are collapsed into single indicator. We filtered terms appearing in fewer than 0.5% and more than 30% of training documents.

Bill Survival				Bill Death			
Contents		Title		Contents		Title	
resources	0.112	title as	0.052	percent	-0.074	internal	-0.058
ms	0.056	other purposes	0.041	revenue	-0.061	the internal	0.024
authorization	0.053	for other	0.028	speaker	-0.050	revenue	-0.022
information	0.049	amended by	0.017	security	-0.037	prohibit	-0.020
authorize	0.030	of the	0.017	energy	-0.037	internal revenue	-0.019
march	0.029	for the	0.014	make	-0.030	the social	-0.018
amounts	0.027	public	0.012	require	-0.029	amend title	-0.016
its	0.026	extend	0.011	human	-0.029	to provide	-0.015
administration	0.026	designate the	0.010	concerned	-0.029	establish	-0.015
texas	0.024	as amended	0.009	department	-0.027	SYMBOL to	-0.014
interior	0.023	located	0.009	receive	-0.025	duty on	-0.013
judiciary	0.021	relief	0.009	armed	-0.024	revenue code	-0.013

Table 4: Full model: text terms with highest impact (eq. 4). Impact scores are not comparable across models, so for comparison, the impacts for the features from Table 2 here are, respectively: 0.534, 0.181, 10^{-4}, 0.196, 0.123, 0.063, 0.053; -0.011, 0, 0.003, 0.

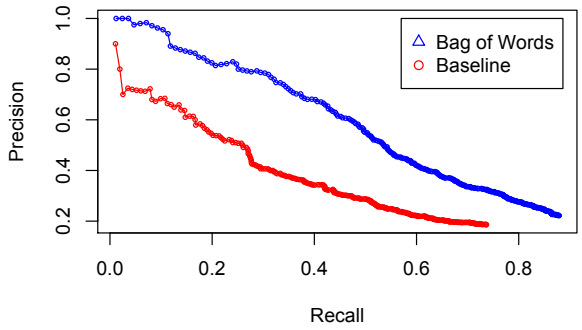

Figure 1: Precision-recall curve (survival is the target class) comparing the bag of words model to the baseline.

performs the bill categories and proxy vote models. This suggests that there is more information in the text contents than either the functional categories or similarity to past bills.[16]

5.4 Full Model

Finally, we considered a model using all three kinds of text features. Shown in Table 3, this reduces error only 2% relative to the bag of words model. This leads us to believe that direct use of text captures most of what functional bill category and proxy vote features capture about bill success.

Table 4 shows the terms with greatest impact. When predicting bills to survive, the model seems to focus on explanations for minor legislation. For example, *interior* and *resources* may indicate non-controversial local land transfer bills. In titles, *designate* and *located* have to do with naming federal buildings (e.g., post offices).

As for bills that die, the model appears to have captured two related facts about proposed legislation. One is that legislators often sponsor bills to express support or concern about an issue with little expectation that the bill will become a law. If such "position-taking" accounts for many of the bills proposed, then we would expect features with high impact toward failure predictions to relate to such issues. This would explain the terms *energy*, *security*, and *human* (if used in the context of human rights or human cloning). The second fact is that some bills die because committees ultimately bundle their contents into bigger bills. There are many such bills relating to tax policy (leading to the terms contained in the trigram *Internal Revenue Service*, the American tax collection agency) and *Social* Security policy (a collection of social welfare and social insurance programs), for example.[17]

[16] We also experimented with dimensionality reduction with latent Dirichlet allocation (Blei et al., 2003). We used the topic posteriors as features in lieu of words during training and testing. The symmetric Dirichlet hyperparameter was fixed at 0.1, and we explored 10–200 topics. Although this offered speedups in training time, the performance was consistently worse than the bag of words model, for each number of topics.

[17] The term *speaker* likely refers to the first ten bill numbers, which are "reserved for the speaker," which actually implies that no bill was introduced. Our process for marking bills that survive (based on committee recommendation data) leaves these unmarked, hence they "died" in our gold-standard data. The experiments revealed this uninteresting anomaly.

	Model	Error (%) 109th	110th
	most frequent class	11.8	14.5
§4.2	baseline (no text)	11.1	13.9
§5.1	bill categories	10.9	13.6
§5.2	proxy vote, all three	9.9	12.7
§5.3	unigram & bigram	8.9	10.6
§5.4	full model	8.9	10.9

Table 5: Replicated results on two different data splits. Columns are marked by the test-set Congress. See §5.5.

5.5 Replication

To avoid drawing conclusions based on a single, possibly idiosyncratic Congress, we repeated the experiment using the 109th and 110th Congresses as test datasets, training only on bills prior to the test set. The error patterns are similar to the primary split; see Table 5.

6 Discussion

From a political science perspective, our experimental results using text underscore the importance of considering the substance of policy proposals (here, bills) when attempting to explain their progress. An important research direction in political science, one in which NLP must play a role, is how different types of issues are managed in legislatures. Our results also suggest that political considerations may induce lawmakers to sponsor certain types of bills with no real expectation of seeing them enacted into law.

Considerable recent work has modeled text alongside data about social behavior. This includes predictive settings (Kogan et al., 2009; Lerman et al., 2008), various kinds of sentiment and opinion analysis (Thomas et al., 2006; Monroe et al., 2008; O'Connor et al., 2010; Das et al., 2009), and exploratory models (Steyvers and Griffiths, 2007). In political science specifically, the "text as data" movement (Grimmer and Stewart, 2012; O'Connor et al., 2011) has leveraged tools from NLP in quantitative research. For example, Grimmer (2010) and Quinn et al. (2006) used topic models to study, respectively, Supreme Court proceedings and Senate speeches. Closest to this work, Gerrish and Blei (2011) combined topic models with spatial roll call models to predict votes in the legislature from text alone. Their best results, however, came from a text regression model quite similar to our direct text model.

7 Conclusions

We presented a novel task: predicting whether a Congressional bill will be recommended by a committee. We introduced a strong, expert-informed baseline that uses basic social features, then demonstrated substantial improvents on the task using text in a variety of ways. Comparison leads to insights about American lawmaking. The data are available to the research community.

Acknowledgments

We thank the anonymous reviewers, David Bamman, Justin Grimmer, Michael Heilman, Brendan O'Connor, Dani Yogatama, and other members of the ARK research group for helpful feedback. This research was supported by DARPA grant N10AP20042.

References

E. Scott Adler and John Wilkerson. 2005. The scope and urgency of legislation: Reconsidering bill success in the house of representatives. Paper presented at the annual meeting of the American Political Science Association.

E. Scott Adler and John Wilkerson. 2012. *Congress and the Politics of Problem Solving*. Cambridge University Press, London.

Beata Beigman Klebanov, Daniel Diermeier, and Eyal Beigman. 2008. Lexical cohesion analysis of political speech. *Political Analysis*, 16(4):447–463.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Joshua Clinton, Simon Jackman, and Doug Rivers. 2004. The statistical analysis of roll-call data. *American Political Science Review*, 98(2):355–370.

Pradipto Das, Rohini Srihari, and Smruthi Mukund. 2009. Discovering voter preferences in blogs using mixtures of topic models. In *Proceedings of the Third Workshop on Analytics for Noisy Unstructured Text Data*.

Sean Gerrish and David Blei. 2011. Predicting legislative roll calls from text. In *Proc. of ICML*.

Justin Grimmer and Brandon Stewart. 2012. Text as data: The promise and pitfalls of automatic content analysis methods for political documents. `http://www.stanford.edu/~jgrimmer/tad2.pdf`.

Justin Grimmer. 2010. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, 18(1):1–35.

Simon Jackman. 2001. Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking. *Political Analysis*, 9(3):227–241.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proc. of NAACL*.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

Kevin Lerman, Ari Gilder, Mark Dredze, and Fernando Pereira. 2008. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proc. of COLING*.

Nolan McCarty, Howard Rosenthal, and Keith T. Poole. 2006. *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press.

Burt Monroe, Michael Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proc. of ICWSM*.

Brendan O'Connor, David Bamman, and Noah A. Smith. 2011. Computational text analysis for social science: Model complexity and assumptions. In *Proc. of the NIPS Workshop on Comptuational Social Science and the Wisdom of Crowds*.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of ACL*.

Keith T. Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384.

Keith T. Poole and Howard Rosenthal. 1991. Patterns of congressional voting. *American Journal of Political Science*, 35(1):228–278.

Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2006. An automated method of topic-coding legislative speech over time with application to the 105th–108th U.S. Senate. Paper presented at the meeting of the Midwest Political Science Association.

Kevin M. Quinn, Burt L Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum.

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