Adding the Magic Words: The Importance of Legal Details in Successful Legislation*

Matthew J. Denny†
Department of Political Science
Penn State University

April 1, 2019

Abstract

All pieces of legislation live on a spectrum between pure position-taking stunts with no chance of becoming law, and serious lawmaking efforts that are almost sure to advance for institutional or procedural reasons. The position taking bills are sometimes easy to spot, such as the dozens of Republican attempts to repeal the Affordable Care Act, but have so far eluded efforts at large scale classification. Yet the seriousness of the lawmaking effort behind a piece of legislation is critical information for those seeking to understand why some bills advance through the legislative process while others languish in committee. In this paper, I introduce an approach to assessing how serious of a lawmaking effort a bill is, by looking for clues in the legal and technical language used in that bill. I hypothesize that increased use of legal and technical terminology, and the inclusion of certain technical provisions in a piece of legislation provide signals about the quality of the bill or "seriousness" of the lawmaking effort it represents. To assess my hypothesis, I develop and apply a novel statistical model for detecting boilerplate legal terminology in a corpus of legislation. I then apply this method to the text of over twenty years of U.S. congressional bills. Consistent with my hypotheses, I find that the amount of legal and technical detail contained in a bill is predictive of whether it will advance out of committee, and whether it will eventually become law.

*First version: March 20, 2018. This version: April 1, 2019. This research was supported by the National Science Foundation under IGERT Grant DGE-1144860. This paper is the second chapter of my dissertation. Software implementing the measurement technique described in this paper is available: github.com/matthewjdenny/SpeedReader

†mdenny@psu.edu; 203 Pond Lab, Pennsylvania State University, University Park, PA 16802

1 Introduction

The overwhelming majority of bills introduced in the U.S. Congress never get much consideration from the chamber. Of the approximately 86,000 non-minor, non-appropriations bills introduced in a House or Senate committee between 1993 and 2016, less than 10% made it out of committee, and only about 3% became law. As a result, the conventional wisdom suggests that Congress is broadly unproductive (Mayhew 1991).

Scholars have long recognized these challenges and have devoted significant attention to understanding why some bills succeed and others fail. Some have argued that constraints on the legislative agenda mean that certain bills, such as reauthorizations or leadership bills will be much more likely to become law because they are institutionally advantaged (e.g. Woon 2009). Others have shown that support from outside interest groups can help push a bill through the legislative process, while opposition can keep a policy idea from making it out of committee (e.g. Hall and Wayman 1990; Hall and Deardorff 2006; Grossmann and Pyle 2013). Yet another strain of research has demonstrated the importance of characteristics of the coalition of legislators involved in drafting and advancing a bill, such as securing bipartisan support, to its chances of advancing (e.g. Matthews 1960; Thomas and Grofman 1992; Anderson et al. 2003; Miquel and Snyder 2006; Hasecke and Mycoff 2007; Desmarais et al. 2013). Finally, a large body of research has explored the relationship between the size of a bill's co-sponsoring coalition, and its chances for success (see, for example: Campbell 1982; Krehbiel 1995; Kessler and Krehbiel 1996; Wilson and Young 1997; Harward and Moffett 2010; Craig et al. 2018).

Yet there is another perspective that suggests that many legislators simply are not focussed on the legislation they introduce actually becoming law, and are instead primarily interested in legislation as a position taking opportunity (Matthews 1960). This is exemplified by the now famous quote from of Senator Carl Hayden's advice to new members of Congress:

> There are two kinds of Congressmen—show horses and work horses. If you want to get your name in the papers, be a show horse. If you want to gain the respect of your colleagues, keep quiet and be a work horse (Matthews 1960, p. 94).

Extending this perspective beyond legislators to bills themselves, we can also find examples of purely position-taking legislation being introduced by legislators who also introduce legislation that is intended to become law. This poses a fundamental problem for all studies of lawmaking: how are we supposed to understand the determinants of legislative success when a (potentially) large number of bills are not intended to succeed from the get-go?

In this study, I begin to lay the theoretical and methodological foundation for understanding the seriousness of the lawmaking effort encapsulated in a bill by examining the *text of the bill itself*. I follow a simple intuition, that legislators will devote more time, effort, and staffing resources towards drafting bills that are intended to become law, and relatively less resources towards drafting legislation that is intended primarily as a position taking instrument. I operationalize this intuition as the amount of legal and technical boilerplate language (relatively standardized legal jargon meant to "glue" together and clarify the substantive policy ideas in a bill) included in the substantive provisions of a bill— what I refer to as the amount of *legal detail* in those provisions, and develop a new statistical method to identify this language.

To bolster this analysis, I also explore the use of non-substantive provisions in legislation, such as *definitions*, *findings*, and *technical corrections*. These provisions provide context for interpreting and implementing the substantive policies contained in a piece of legislation. I argue that these two types of legislative text (legal details and non-substantive provisions) are important because they indicate that legal and staffing resources were devoted to drafting the bill—and by extension that it was intended as a more serious lawmaking effort. As a result, *I expect that the initial versions of bills containing more of both types of legal and technical content will be more likely to advance out of committee, and ultimately to become law*.

Measuring the amount of legal and technical language and non-substantive provisions in a bill is a challenging endeavor. Existing empirical investigations of legislative text that sought to (remove) legal and technical language and non-substantive provisions have relied on hand curated lists of terms and ad-hoc rules for identifying this language. Such an approach is insufficient for my purposes because I seek to precisely identify the universe of legal and technical language and non-substantive provisions in a bill, and not just those terms that interfere with an analysis of substantive language. To this end, I introduce new approaches to identifying both types of language in legislation. Specifically, I extend a concept from information theory called *mutual information*, and use it to find the statistical signature of legal language in legislation. To identify non-substantive provisions in bills, I use the legislation parser from Denny (2018), and develop a typology of non-substantive sections. Both methods are also subjected to human validation efforts, which indicate that the computational methods I employ bear similarity to human judgements.

To assess my theory that the inclusion of legal details in legislation signals the seriousness of the lawmaking effort contained in that piece of legislation, I analyze a corpus of approximately 86,000 non-minor, non-appropriations bills introduced in the U.S. Congress between 1993 and 2016. I find that the amount of legal detail included in the substantive provisions of a piece of legislation is a strong positive predictor of a bill seeing any action beyond committee (the first major hurdle a bill must clear in the U.S. Congress), and of it ultimately becoming law. I also find that the initial versions of bills that advance out of committee (and those that eventually become law) tend to contain a greater number of several types of non-substantive provisions than those that do not, on average. Importantly, I find that the main effect of both types of legal and technical language is on the probability a bill advances out of committee, indicating that bill "quality" is most important as a signal early in the legislative process. Taken together, I believe my findings provide evidence that bill quality, operationalized as the amount of legal and technical detail included in legislation, is an important factor associated with bill advancement. By extension, I argue that the amount of legal and technical detail included in legislation is therefore a reasonable proxy for the seriousness

of the lawmaking effort represented by a bill.

This study has broad methodological and theoretical implications. The method (and accompanying software) I develop for identifying legal and technical jargon in legislation is broadly applicable to settings where scholars have document-level metadata, and want to identify domain stopwords[1] in their corpus. For example, scholars often apply topic models (see, for example, Grimmer and Stewart 2013; Roberts et al. 2014) to try and understand the topical themes in a collection of political texts, but domain stopwords can make the output uninterpretable. Theoretically, my approach to characterizing the amount of legal and technical detail included in legislation represents a first step towards accounting for variation in bill quality for scholars studying legislative success. My findings also build on the growing body of work around the textual signature of unorthodox lawmaking (see, for example, Adler et al. 2003; Wilkerson et al. 2015; Casas et al. 2018), and productivity of Congress (Mayhew 1991; Binder 1999, 2015). I do so by highlighting another dimension (legal and technical content) on which successful lawmaking efforts differ from unsuccessful ones, emphasizing the need to rethink traditional measures of legislative productivity and effectiveness.

2 Theory and Hypotheses

> "I love these members, they get up and say, 'Read the bill,' " said Rep. John Conyers. "What good is reading the bill if it's a thousand pages and you don't have two days and two lawyers to find out what it means after you read the bill?"– John Conyers (D-MI, qtd. in Blumenthal (2009)).

Rep. John Conyers (D-MI) was famously quoted as saying that most members of Congress do not (carefully) read the bills that they vote on. And, if they did, he claimed, most legislators would be unlikely to understand the language without the help of a team of lawyers.[2] Most serious legislative proposals are challenging to read because they are composed of complicated, technically

[1] Very common words such as "the" and "if" that tend to obscure patterns in substantively interesting words.
[2] https://sunlightfoundation.com/2009/07/27/rep-conyers-dont-read-the-bill/

written provisions that seek to clarify and codify the intentions of the authors. While this technical language makes bills difficult to parse, it is incredibly important. After all, bureaucrats, judges, and other government officials tasked with implementing the legislation must be able to determine precisely what they are required to do. Thus, far from purely satisfying technocratic legislative drafting standards, this language can be critical to ensure that a law is implemented as its authors intended and that it can survive legal challenges.

Put another way, words matter, even if they seem unrelated to the actual substance of a policy proposal. For example, a key provision of the 900-page Affordable Care Act was challenged throughout the federal judicial hierarchy, including the Supreme Court, over a four-word drafting error.[3] At the same, bills that do not contain any of the requisite legal and technical content to integrate with existing laws are generally dismissed as position taking stunts. A recent example was a one-sentence bill introduced by Rep. Matt Gaetz (R-FL) that would have abolished the EPA, effective December 31, 2018.[4] This bill was generally derided in the press for ignoring all of the complex legal issues that would arise from abolishing the EPA, making it completely impractical for Congress to enact, even for those legislators who generally agreed with its sentiment. In a similar vein, Clark (2009) shows that many legislators introduce overly broad legislation to curb the federal judiciary with little intention that these bills become law.

Surprisingly, despite the fact that the text of legislation is what must be interpreted and implemented as a bill becomes law (Strokoff 1996), scholars have devoted relatively little attention to understanding lawmaking in terms of the bill text itself. While some have focused on the content of legislation in a particular issue area (e.g. Huber and Shipan 2002) and others have taken a bird's eye view of the topical diversity of the legislative agenda (e.g. Mayhew 1974, 1991; Binder 1999), our understanding of lawmaking is still mostly divorced from the words that eventually become

[3] https://www.nytimes.com/2015/05/26/us/politics/contested-words-in-affordable-care-act-may-have-been-left-by-mistake.html

[4] https://www.cnbc.com/2017/02/15/freshman-republican-congressman-reveals-bill-to-abolish-the-epa.html

law. Furthermore, where scholars have used text as data methods to study lawmaking, they tend to remove legal and technical language as a preprocessing step, treating the very language that bureaucrats and judges will actually seek to interpret as a hindrance to their analyses (Yano et al. 2012; Wilkerson et al. 2015; Casas et al. 2018).

While the legal and technical minutiae of legislation are certainly not headline-grabbing, legislative drafting experts recognize how critical they are to a bill actually encapsulating the policy ideas it contains. This is because such language is required for a bill to interface with the existing body of law, and to meet the formatting requirements for legislation set out by the Office of the Legislative Counsel (Strokoff 1996). As Strokoff puts it: "Attorneys are charged with taking the idea of any Member or committee of the House of Representatives requesting the services of the (Office of the Legislative Counsel) and transforming it into legislative language or, as one of my clients used to say, 'the magic words.'" If the inclusion of this legal detail in legislation is *necessary* for a bill to become law, then the quality of these details should be important indicator as to whether a bill was intended advance through the legislative process (and whether it actually will).

Put another way, it seems reasonable that the inclusion of legal detail in a bill may serve as an indicator of bill "quality". Because including this detail requires the staffing and legal resources of the authoring coalition, I expect that legislators will devote more of these resources towards bills they are more heavily invested in becoming law. For example, the authors of the "Healthy Forests Restoration Act of 2003" (a bill that became law in 2003), spent almost four pages defining what constitutes a community that is "at risk" for wildfires, for the purposes of funding under the bill.[5] This attention to detail beginning with the introduced version of the bill could be seen as an indication that it was a "serious" lawmaking effort, and worthy of further consideration beyond committee. Contrast this with the one-sentence position taking bill (to terminate the EPA) introduced by Rep. Matt Gaetz (R-FL) in early 2018, that he likely could have written by himself, in

[5] https://www.congress.gov/bill/108th-congress/house-bill/1904

ten minutes, with minimal aide from his legal staffers.[6] What becomes clear by reading the text of as-introduced versions of successful bills is that they tend to contain lots of legal detail, supporting the idea that legal detail is an indicator of quality.

Another possible explanation for this devotion of resources towards including legal detail in the early versions of successful bills is that the bill's authors may believe the bill is likely to advance for institutional or procedural reasons. For example, when a bill is sponsored by a committee chair, party leadership, or is a reauthorization bill, members of Congress may deem it effective to devote their limited staffing resources towards that bill. One prominent example of such a bill was the USA PATRIOT Act.[7] The introduced version of this bill spanned 341 pages and contained deliberately broad, yet complex legal language granting security agencies wide ranging powers to pursue individuals they deemed to be threats to the United States. This represented a significant outlay of staff resources, but the bill's authors had strong reasons to believe that it would become law, due to the recent 9/11 terrorist attacks driving demand for a response from Congress.

While it is difficult to empirically distinguish between a quality-signaling, and a beliefs-based explanation for the inclusion of legal detail in successful legislation, the end results should be the same. In either case, we should expect that the early versions of these more serious lawmaking efforts will contain more legal detail than the early versions of bills primarily intended as postion taking vehicles, for example. Regardless of the direction of causality (legal detail $\longrightarrow$ success, beliefs about success $\longrightarrow$ legal detail), I still expect that legal detail in the text of a bill should be an important predictor of whether it will advance out of committee (see serious consideration in Congress), and whether it will ultimately become law. In formulating my hypothesis, I focus more specifically on the substantive policy provisions of serious (successful) lawmaking efforts (those provisions that actually spell out what the bill does, as opposed to justifying why the bill is needed, for example) as containing more legal detail than the substantive policy provisions of less serious

[6] https://www.congress.gov/bill/115th-congress/house-bill/861
[7] https://www.congress.gov/bill/107th-congress/house-bill/3162/text/ih

bills.

Hypothesis 1 *Bills containing more legal detail in their initial versions will be more likely to advance out of committee, and to eventually become law.*

In addition to the use of legal detail to clarify the meaning of substantive policy proposals, a bill can also include a number of non-substantive provisions that clarify the intent and implementation of a policy. Most successful bills contain a number of these non-substantive provisions, which typically serve to either clarify the meaning of terms or concepts in the substantive provisions of the bill, or provide some justification for why the policies included in the bill are a good idea. To give a sense of the types of provisions contained in legislation, Table 1 provides descriptive statistics of the types of sections found in the bills I analyze in this study. These counts are drawn from the replication data provided by Denny (2018), who develops an algorithm for segmenting U.S. Congressional bills into their constituent sections.

While the bulk of the provision in these bills are substantive provisions or titles (e.g. "TITLE III: Improving The Nutritional Value of School Lunches ..."), the bills I examine in this study also contain a wide variety of non-substantive provisions. In particular, I focus on the inclusion of seven of these types of non-substantive provisions as potential indicators of bill quality. These provisions are highlighted in blue in Table 1, and fall into two broad groups: provisions that clarify or specify what other provisions in the bill (or existing laws) do, and provisions that clarify the views of a bills' authors. Before discussing these provisions, note that I do not consider four additional types of non-substantive provisions (*Front Matter, Short Titles, Tables of Contents, Effective Dates*), because they are typically short, and highly standardized across bills.

Authorizations of Appropriations, Conforming Amendments, Definitions, and *Technical Corrections* all serve to clarify the interpretation of other provisions in the bill (or existing laws). For example, *Authorizations of Appropriations* specify the dollar amounts to be allocated to pay for

Type	# of Provisions	# of Bills	Substantive?
Authorization of Appropriations	6,363	5,485	No
Conforming Amendments	3,166	2,326	No
Definitions	17,283	14,331	No
Effective Date	8,860	7,349	No
Findings	18,349	17,234	No
Front Matter	85,941	85,941	No
Purposes	4,000	3,533	No
Sense of House/Senate/Congress	2,499	1,974	No
Short Title	66,571	63,445	No
Substantive Provision	372,729	85,945	Yes
Substantive Title	104	93	Yes
Technical Correction	618	497	No
Table of Contents	1,118	1,117	No
Combined	587,601	85,949	

Table 1: Counts of provisions by type. **# of Provisions** indicates the number of provisions of a given type, while **# of Bills** records the number of unique bills containing a provision of that type. **Substantive?** indicates whether each provision type was considered to be a substantive policy provision. Provision types highlighted in blue are the non-substantive provisions I focus on for the purpose of my analysis.

programs described in the substantive provisions of a bill[8], while *Definitions* sections typically clarify the meaning of terms that are used across numerous substantive provisions in a bill.[9] Similarly, *Conforming Amendments*, and *Technical Corrections* alter everything from formatting and spelling errors in other bills or laws, to changing the content of their substantive provisions. What these types of provisions share in common is that they interface with other substantive provisions (both in the same bill, and in other bills), and serve to clarify and or specify the meaning of these provisions.

Findings, *Purposes*, and *Sense of House/Senate/Congress* provisions are non-binding, and do not

[8]Note that Denny (2018) classifies unusually long *Authorization of Appropriations* sections as substantive provisions because they typically mix a specification of how much is to be spent with a description of *how* it can be spent.

[9]Note that substantive provisions may also precisely define terms within the provision, but that these definitions are typically less germane to the rest of the substantive provisions in the bill.

directly interface with the substantive provisions of a bill. Yet I argue that they still have an important political purpose in justifying the policy proposals included in a bill, and expressing the views of those involved in supporting the bill. *Findings* provisions typically report the conclusions of academic, government, or industry research on the issues addressed in a bill. For example, a bill seeking to address opioid addition might include a findings section that describes the scope of the public health problem, and what researchers have determined to be effective solutions. Similarly, *Purposes* provisions typically spell out in plain English what a bill's authors intend to accomplish with that legislation. Finally, *Sense of House/Senate/Congress* offer a platform for Members of Congress to register their opinions on issues (usually) related to the policy content of a bill. In particular, these provisions can be added strategically to a bill as a (non-binding) form of political compromise.

A prime example of the use of *Sense of House/Senate/Congress* provisions as a mechanism for (soft) political compromise was the Keystone XL Pipeline Approval Act (114-S-1), that was introduced by Sen. John Hoeven (R-ND) in 2015.[10] This controversial bill was eventually vetoed by President Obama, but would have allowed for the construction of a pipeline connecting Canada's tar sands with the Gulf of Mexico. The bill was supported by all Republicans in the Senate, and a number of Democrats in oil-producing states, but was controversial due to the environmental impact of burning fossil fuels, and the impact of the pipeline on the communities it was supposed to pass through. As an apparent compromise, the bill contains two non-binding *Sense of the Senate* provisions, one of which "Expresses the sense of the Senate that climate change is real and not a hoax", and the other of which encourages Congress to pass an excise tax on oil derived from tar sands. Thus, adding provisions of this type may be seen as a form of political compromise, and should be positively associated with bill advancement.

To summarize, My argument is that the inclusion of these non-substantive provisions can serve

[10] https://www.congress.gov/bill/114th-congress/senate-bill/1

as a mechanism to clarify and constrain the scope of policy proposals, and as a non-binding mechanism for political compromise. I therefore expect that the inclusion of any of the seven types of provisions discussed above is another indicator of the seriousness of a lawmaking effort, and the likelihood a bill will advance out of committee and eventually become law. Here again, it could be that the bill's authors include these non-substantive provisions to head off opposition, or because they believe that a bill is likely to advance for institutional or procedural reasons, and therefore put more effort into ensuring its smooth passage.

Hypothesis 2 *Bills containing more non-substantive provisions in their initial versions will be more likely to advance out of committee, and to eventually become law.*

3 Data and Methods

In order to assess the hypotheses laid out in the previous section, I needed to: (1) develop a measure of the amount of legal detail included in the substantive provisions of a bill; and (2) to count the number of non-substantive sections in that bill. Fortunately, I was able to rely on the labels assigned by Denny (2018) as a basis for (2), so my main methodological challenge was to measure the amount of legal detail included in the substantive provisions of a bill. Before I discuss my measurement approach, I begin by discussing the corpus of congressional bills I use, and how I preprocessed them.

The bill text data for this study were collected by Handler et al. (2016); Denny (2018) and include all versions (introduced, reported, etc.) of all bills introduced during the 103rd-114th sessions of Congress (1993-2016). These text data were also linked to bill-level metadata compiled by the Congressional Bills Project (Adler and Wilkerson 2012). Using the algorithm developed by Denny (2018), each bill was also broken up into its constituent provisions (sections) and each section was tagged for its type (front matter, substantive provision, definitions, purposes, etc.), so that I could determine which types of non-substantive sections each bill contained. The full dataset includes

almost a million individual sections from 117,910 versions of 92,660 unique bills, containing over 300 million words.

Before analyzing these data, I decided to exclude four types of bills from my analysis. The first type were "minor bills", as coded by the Congressional Bills Project (Adler and Wilkerson 2012). These bills typically deal with issues like renaming post offices or creating commemorative coins, and are different enough from the majority of policy proposals that I feel they should not be considered together. Additionally, because these bills are less controversial, they tend to become law at a higher rate than non-minor bills. I also exclude all private bills, because they also tend to deal with more minor issues (such as the relief of an individual), and are not a good comparison to legislation like the Affordable Care Act. Next, I follow Casas et al. (2018) in excluding all bills referred to the House or Senate Appropriations Committees from my analysis because some of these bills are essentially must-pass legislation (to keep the government running), and therefore are unlikely to be subject to the same dynamics in terms of their chances for success. Finally, I exclude bills that were not originally introduced in either a House or Senate committee (for example, Joint Resolutions), as I seek to compare bills with similar starting points. Therefore, out of an initial corpus of 92,660 bills, I am left with 85,949 after applying these filters.

After applying the filters discussed above, I was left with the full text of 85,949 bills that had been split into their respective provisions by Denny (2018). My next step was to preprocessing my data into a document-term matrix (counts of each unique term in each provision), because the method I developed (described below) for identifying the legal and technical language in substantive provisions operates on this matrix. The standard approach to forming a document-term matrix is to use individual words (unigrams) as the terms (Grimmer and Stewart 2013). However, unigrams can often have an ambiguous meaning in political texts. For example, the term "section" could refer to "section 22(a)" of a bill, or "section 8 housing".

To address this issue, I chose to use phrases instead of unigrams as the terms in my document term-matrix. A phrase is a coherent multi-word expression such as "wildlife preservation", or "prohibit gun sales" that provides additional context for each individual word by keeping them in sequence. In addition, much of the legal and technical language I found while reading legislation tended to come in the form of multiword expressions (e.g. "establish a committee", "subparagraph (b)"). I extracted phrases from the text of each bill section using the phrasemachine R package (Handler et al. 2016). These phrases were limited to a maximum length of 3 constituent words, as this length produced highly interpretable phrases while limiting the number of unique terms to several million.[11] This process resulted in a document term matrix containing more than 200 million terms, with about 5.8 million unique terms.

3.1 Identifying Legal Detail in Legislation

At a high level, we can think of bills as being composed of two types of language: policy language (what the bill does), and legal details (the legal and technical "glue" that holds the policy language together and clarifies its meaning).[12] To a human coder, legal details are relatively easily distinguishable from policy language. But human coding is not readily scalable to a corpus containing tens of thousands of documents. I take a computational approach to identifying legal details in legislative texts based on their statistical signature. The method I develop runs in only a few minutes on a standard laptop, allowing me to quickly and accurately classify these terms.

This approach is based off of the following intuition. In the U.S. Congress, we expect that Democrats and Republicans will have different policy goals, so policy language should be used differently by members of each party. What I expect to remain relatively constant (between parties, within an issue area), is the use legal and technical language (legal details). If a member of

[11] Phrases were extracted using the "PhrasesNoCoord" grammar included in phrasemachine. I also limited the n-gram size to three in order to avoid too much double counting of terms in the vocabulary (e.g. "red car", "shiny red car", "new shiny red car").

[12] See Denny et al. (2015) for a more in-depth discussion.

Congress wants to place a number of restrictions on how money may be spent by an agency, or lay out a detailed rule-making process, or spell out the specifics of a grant program, their staff, or lawyers working for the Office of the Legislative Counsel will use standard legal and technical language to do so (Strokoff 1996). This uniformity is reinforced by the standard training most law students receive in crafting the legal language that goes into legislation (see, for example Eskridge et al. 2014).

To formulate my claim more precisely, I expect that: on average, in a given issue area, Democrats and Republicans will use legal and technical language in a similar way. Put another way, we would say that if we are given information about the use of a legal or technical word or phrase across documents, it will not help us distinguish which party introduced it. For example, if we are given information about the number of times the phrase "described in section" occurs in each document in a corpus of Congressional bills, we are no closer to knowing what a bill that has a high or low count of this phrase is about, or which party introduced it. Building off of this intuition, I take a statistical approach to identifying which terms give us no meaningful information about the partisanship of a bill's authors.

3.2 Using Contingency Tables to Classify Boilerplate Terms

The most common representation of text data in social science applications is as a document-term matrix, where each row represents a document, and each column represents a unique term in the vocabulary (Grimmer and Stewart 2013). Entries in this matrix then record the count of term j in document i. The rows of this matrix (documents) can then be collapsed over various combinations of metadata attributes to form a *contingency table*. Thus, I started with the document term matrix described above, and used it (along with the Congressional Bills Project metadata) to create a contingency table where each row recorded the counts of terms in bills sponsored by members of a given party, in a given minor issue area (using the Policy Agendas Project minor topic codes). More specifically, in this study I formed a contingency table recording the number of times each

Category	"striking paragraph"	"opioid addiction"	"nuclear power"	"Affordable Care Act"	...
Democrat, Nuclear Energy	1,023	0	123	0	...
Republican, Nuclear Energy	826	0	415	0	...
Democrat, Health Insurance	956	124	0	36	...
Republican, Health Insurance	1,452	141	0	285	...
...	...	...	...	...	...

Table 2: Each row in this example contingency tabel records the count of each unique term aggregated across all bills in a particular category (e.g. Democrat sponsored bills, about nuclear energy policy).

term was used in Democrat and Republican sponsored bills in each of about 240 minor issue areas (e.g. "Health: Drug Industry" or "Environment: Drinking Water") as coded by the Congressional Bills Project. Thus the full contingency table contained approximately 480 rows (240 issues $\times$ 2 parties), and about 5 million columns (the number of unique terms in the corpus). A small example contingency table is illustrated in Table 2.

Intuitively, some columns (vocabulary terms) of this contingency table will give us more information about which row we are in. In the example in Table 2, the phrase "Affordable Care Act" is used much more frequently in Republican sponsored healthcare legislation than in Democrat sponsored legislation on the same issue. Thus, if we came across the term "Affordable Care Act" in a bill section, it would be reasonable to guess that the bill section came from a Republican sponsored bill about healthcare.

If we instead came across the phrase "striking paragraph" in a bill section, it would be much more difficult to make an educated guess about which category the bill section belonged in, because "striking paragraph" is used relatively similarly by both Democrats and Republicans. Therefore, to identify legal and technical terms, I needed to identify terms that were not associated with any particular party, within an issue area.

Category	"striking paragraph"	"opioid addiction"	"nuclear power"	"Affordable Care Act"	...
Democrat, Health Insurance	0.10	0.01	0.00	0.01	...
Republican, Health Insurance	0.15	0.01	0.00	0.03	...

Table 3: An example joint distribution over parties and terms.

3.3 Term Contributions to Mutual Information

The approach I take to identifying legal and technical terms is based on the average contribution each term makes to a quantity called the *mutual information* (Shannon 1948) between terms and categories (Democrat and Republican bills, about a given issue). For a review of related approaches, see Appendix A. In practice, this involves considering the relative number of times each term is used by members of each party, in bills from each issue area. More concretely, I pulled out pairs of rows from the contingency table described in the previous section recording the count of terms in bills sponsored by Democrats and Republicans, respectively, on a given issue area. For example, on such pair of rows might record term counts from Democrat and Republican legislation about health insurance. Finally, each cell in the resulting $2 \times \sim$5 million contingency table was divided by its sum, forming a joint distribution over terms and categories:

Mutual information is a measure of the degree of statistical association between two discrete random variables. In other words, it is a number that characterizes how much information the columns of a joint distribution (vocabulary) give us about which row we are in (class), and vice versa. More formally, the mutual information of two discrete random variables C and V is defined as:

$$I(C;V) = \sum_{c \in C} \sum_{v \in V} p(c,v) \log \left(\frac{p(c,v)}{p(c)\,p(v)} \right) \tag{1}$$

In this equation, p(c,v) is the joint probability of observing c and v, and p(c) and p(v) are the marginal probabilities of observing and c and v respectively (Cover and Thomas 1991). Looking at the form of this equation, we see that the mutual information of a joint distribution will increase

Distribution 1

	"pursuant to section"	"fiscal year"
Democrat	0.25	0.25
Republican	0.25	0.25

I(C;V) = 4 (.25 × log(1)) = 0

Distribution 2

	"repeal Obamacare"	"carbon tax"
Democrat	0.00	0.50
Republican	0.50	0.00

I(C;V) = 2 (.5 × log(2)) = 0.693

Table 4: Example mutual information calculations for two joint distributions

as we include more terms that distinguish between categories, and decrease as we include more terms that make it harder to distinguish between categories. This result is not intuitive, so I provide several illustrations below.

Table 4 provides examples of mutual information calculated on two toy joint distributions. Distribution 1 places uniform probability on all entries. If we calculate the mutual information of this distribution, we see that each term will involve multiplying by the log of 1 (which equals zero), so the mutual information will be zero. This makes intuitive sense, because if we come across a term in a document belonging to one of these categories, we can do no better than chance at guessing the category, because knowing the term was in the document gives us no new information. Now, if we consider Distribution 2, we see that the terms distinguish perfectly between categories. This is reflected in the higher mutual information for this joint distribution.

With this intuition in mind, we can now illustrate the core property of mutual information on which I rely to identify terms that make it harder to distinguish between categories. Distribution 1 in Table 5 includes three terms, two of which seem to distinguish between categories while one ("section") does not. Distribution 2 in Table 5 was based on the same raw term counts, but with the "section" column removed. We can see that when we remove "section" from the vocabulary in this toy example, the mutual information of the joint distribution increases substantially. This again makes intuitive sense, because we have removed a lot of information (terms that occurred

	Distribution 1				Distribution 2		
	"section"	**"birth control"**	**"insurance"**			**"birth control"**	**"insurance"**
Democrat	0.36	0.08	0.00		**Democrat**	0.33	0.00
Republican	0.20	0.02	0.14		**Republican**	0.08	0.58
	I(C;V) = 0.11					I(C;V) = 0.428	

Table 5: Example mutual information scores for a distribution with and without the first column included.

with high relative frequency) which made it harder to tell from a randomly selected term which category we were in. Thus we could say that in this toy example, the term "section" made a *mutual information contribution* of 0.11 - 0.428 = -0.317.

I use these mutual information contributions for each term as a way to quantify how much information they give us about whether a Democrat or Republican wrote a particular bill (in a given issue area). In order to identify legal and technical terms I simply average the mutual information contributions of a term in each issue area by repeating the process of calculating its mutual information contribution for each issue area. I refer to this average over mutual information contributions as a terms' Average Contribution to Mutual Information (ACMI). A formal definition of the ACMI for a given term is provided in Appendix B.

Finally, following from the intuition laid out earlier, if a term has a negative ACMI, I classify it as a legal/technical term.

Definition 1 *A* legal/technical *term is a term whose ACMI is less than zero.*

To some readers, the ACMI approach described will seem quite similar to many other methods for calculating statistical associations on contingency tables. Why not use a χ^2 test statistic, or any one of dozens of other measures of statistical association, instead of ACMI? The primary difference between ACMI and more standard test statistics is that ACMI is self-consciously relative. In other

words, the "zero contribution threshold" changes for each corpus as it only depends on the corpus itself. This allows me to avoid having to select a statistical significance threshold for each dataset, and ensures that the threshold I do select (0) is always theoretically motivated.

4 Measurement and Validation

As discussed earlier, to form the contingency table necessary to identify legal and technical terms via my ACMI method, I used the partisanship of the sponsor of a bill, and the bills' "minor topic label" (Adler and Wilkerson 2012) to form categories. There are approximately 240 minor topic labels in the Congressional Bills Project metadata, covering relatively fine-grained topics such as "nuclear power" or "national parks" legislation. For each of these topics, I formed a contingency table consisting of counts of each term in substantive bill sections sponsored by members of each party. Having prepared the data for analysis, I then proceeded to calculate ACMI contributions for each term in the vocabulary.[13]. This process resulted in about 287,000 (3%) terms with a negative ACMI. Table 6 contains some example terms with a negative ACMI.

{pursuant to section}, {period beginning}, {provided in subsection}, {described in paragraph}, {accordance with section}, {the secretary}, {for purposes}, {united states code}, {notwithstanding}, {amendment made}, {act shall be}, {such regulations}, {effective date}, {shall take effect}, {date of enactment}, {striking subsection}, {inserting after paragraph}

Table 6: Example terms with a negative ACMI.

One well-known feature of information theoretic quantities like PMI and ACMI is that they tend to perform poorly for terms that appear infrequently in the corpus (O'Connor 2014). Visual inspection of the terms with a negative ACMI revealed terms that only appeared in bills concentrated in five or fewer minor topics were overwhelmingly not legal or technical terms. Discarding these terms left approximately 15,000 terms remaining. Visual inspection of the remaining terms occasionally revealed phrases like "social security" or "healthcare premiums", but generally indicated

[13] Details of my implementation are provided in Appendix C).

good performance in identifying legal and technical language.

In order to better assess the accuracy of my method, I performed a human coding validation task. The validation task consisted of hand coding all terms with a negative ACMI, that appeared in bills about at least 100 minor topics (about 7,800 terms). For each term, I marked it as a false positive if it gave any indication of being policy language. This criterion is likely to be overly strict, as some terms like "health" were used over a million times, and could have easily been written by a lawyer as part of a boilerplate passage. Additionally, I coded the top 1,000 most frequently appearing terms not coded as legal or technical for false negatives. The results of this coding exercise are provided in Table 7.

	Legal and Technical Terms	Substantive Terms
Negative ACMI (7,800 terms)	77.4%	22.6%
	(157,629,373)	(23,165,756)
Positive ACMI (1,000 terms)	13.8%	86.2%
	(751,747)	(15,991,770)

Table 7: Confusion matrix for human coding validation results. Total count of all terms in category is provided in parentheses.

My validation coding results indicated a false positive rate of approximately 23% and a false negative rate of 13.8%. However, it is difficult to characterize the quality of these results without assessing how they affect the final distribution over legal and technical terms across substantive provisions in legislation. I therefore created counts of legal and technical terms in each substantive provision using all terms I hand coded as legal or technical, all terms with a negative ACMI that appeared in at least 5 issue areas, and all terms with a negative ACMI (287,000 terms). The correlations among these counts are provided in Table 8.

The striking result from Table 8 is that the false positives largely do not matter for the overall distribution of legal and technical terms across substantive provisions. Including all terms with a negative ACMI essentially adds more to the count of legal and technical terms in all substantive

	Hand Coded	5 Issue Areas	Negative ACMI
Hand Coded	1		
5 Issue Areas	0.997	1	
Negative ACMI	0.993	0.997	1

Table 8: Correlations among bill section legal and technical term counts using different cut-offs for identifying these terms.

provisions, but in proportion to those terms I hand coded as being legal or technical terms. Therefore, I took the approach of treating all terms with a negative ACMI score as legal or technical terms. As further evidence that this decision was not consequential, the results I present in the next section are not sensitive to the use of any of the three ways of forming legal and technical term counts illustrated above.

Having successfully formed counts of legal and technical terms in all substantive provisions of all bills in my corpus, my final task was to prepare my data for analysis. As I wanted to assess the relationship between the inclusion of legal details and non-substantive sections and the chances of success for these bills, I needed to aggregate my measurements from the provision to the bill level. To do so, I calculated the average number of legal and technical terms in the substantive provisions of each bill, as well as the average proportion of terms in those substantive provisions that were legal or technical. I chose to operationalize the concept of legal details as the average *proportion* of legal and technical terms in the substantive provisions of a bill in order to account for the fact that bills about some issues might simply require more words to describe their policy proposals than others. To operationalize the use of non-substantive provisions in a bill, I created an indicator variable that recorded whether each of the seven types of non-substantive provisions I identified earlier in the paper were present in a bill. I chose to use an indicator instead of a count so that these variables would not simply be conflated will bill length. Furthermore, despite noting similarities among some of these non-substantive provisions, I wanted to empirically explore the relative importance of the inclusion of each type of provision in a bill.

5 Analysis

In this section, I present the results of two regression analyses aimed at assessing the two hypotheses laid out earlier in the paper. To reiterate, my first hypothesis is that bills containing more legal detail in their initial versions will be more likely to advance out of committee, and to eventually become law. My second hypothesis is that bills containing more non-substantive provisions in their initial versions will be more likely to advance out of committee, and to eventually become law. To assess both of these hypotheses jointly, I estimated two logistic regression models at the bill-level, with an indicator for whether each bill advanced out of committee (became law) as the outcome. By selecting a regression modelling approach, I was also able to control for a number of standard predictors of bill success, and well as controlling for issue area, and session of Congress.

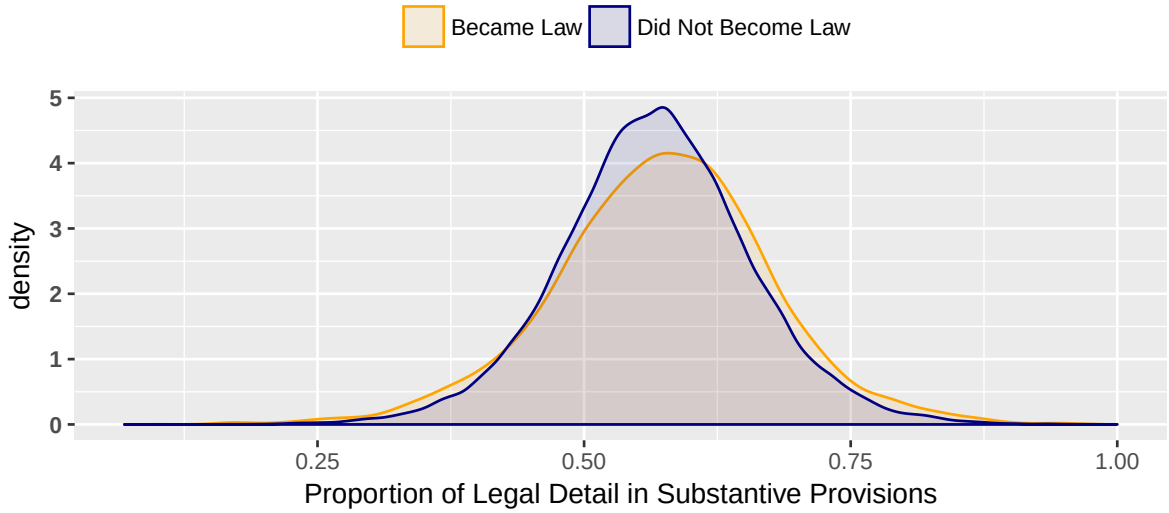

Figure 1: Distributions of the average proportion of boilerplate terms in substantive provisions of the versions as introduced of bills that eventually became law, and those that did not.

To assess Hypothesis 1, I included a variable in the model that recorded the average proportion of legal and technical language (legal detail) in the substantive provisions of a bill. Figure 1 depicts the distributions over the average proportion of legal and technical terms in the substantive provisions of the introduced versions of bills that eventually became law, and those that did not. To assess Hypothesis 2, I included dummy variables for the presence of each of the seven types of

non-substantive sections identified earlier in the introduced version of a bill. To get a sense of how frequently these sections appear, Figure 2 displays the proportion of successful and unsuccessful bills containing provisions of each type.

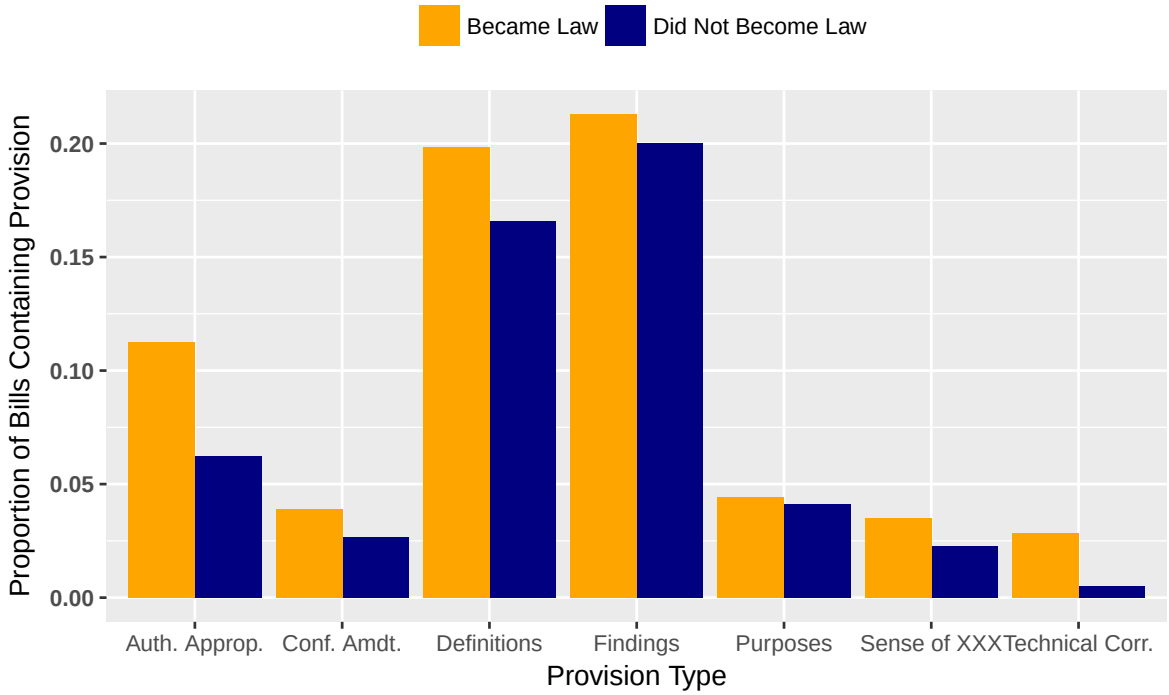

Figure 2: The proportion of introduced version of bills that eventually became law (or did not) containing at least one of each of seven types of non-substantive provisions.

Following Casas et al. (2018), I include a number of control variables in my regression model. These include fixed effects for the "major topic area" of a bill (approximately 20 policy areas from the Congressional Bills Project), and the session of Congress the bill was introduced (as a categorical variable). I also included controls for the gender and the ideological extremeness (absolute value of first dimension DW-NOMINATE score) of the sponsor, as well as whether the sponsor was a member of the majority party. Additionally, I included controls for whether the bill was a leadership bill (the first 10 numbered bills in the House each session), a revenue bill, a reauthorization bill, and an indicator for whether the bill was introduced in the Senate. I also included controls for the log of the number of cosponsors of the bill, and the log of the number of sections

in the bill. Finally, I included controls for whether the sponsor was a member of the committee to which the bill was referred, or if they were the chair or ranking member of either the committee or subcommittee of referral. In particular, I expect the parameter estimates for these committee variables to be positive, as being a chair or ranking member of the referring committee is a significant institutional advantage in advancing legislation.

I began by estimating the model, and confirming that the model AIC is improved when including the non-substantive section dummies and proportion of legal detail variables. I then generated marginal effects on the relative likelihood that a bill advances out of committee or becomes law for all covariates (excluding the session of Congress and major topic fixed effects for readability). These marginal effects[14] (with 95% confidence intervals) are depicted in Figure 3 for passage out of committee, and in Figure 4 for passage into law.

As we can see from Figures 3 and 4 respectively, the marginal effects of the control variables tend to be in the expected direction. For example, a bill's sponsor being a member, ranking member, or chair of the committee to which a bill was referred are all associated with an increased probability of that bill advancing out of committee and ultimately succeeding. Similarly, leadership and reauthorization bills are much more likely to be successful, and the number of cosponsors is also positively associated with bill success. Interestingly, the effect for the log of the number of sections (provisions) in a bill is not significantly different from one, indicating that longer (potentially omnibus or otherwise collaborative) bills are no more likely to become law than shorter bills, controlling for the other variables in the model.

With regards to Hypothesis 1, Figures 3 and 4 indicate that increasing the average proportion of legal detail in the substantive sections of a bill from its minimum observed value (about 6%) to its maximum observed value (99%) is associated with roughly a doubling in the relative likelihood

[14]Marginal effects were calculated by changing the value of a categorical covariate from its reference category, and from changing the value of nominal variables from their minimum to maximum values.

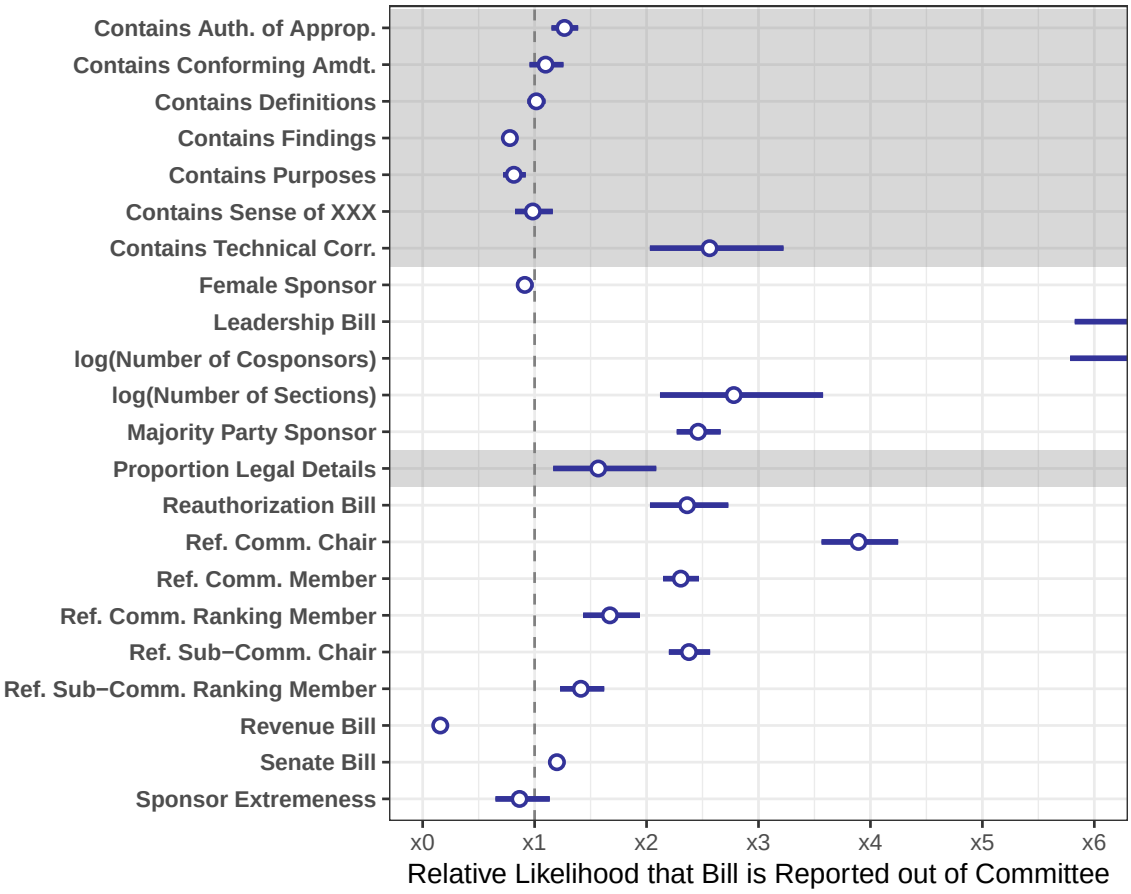

Figure 3: Marginal effects (with 95% confidence intervals) of covariates on the relative likelihood of a bill being reported out of committee. Variables of interest are highlight in gray.

of that bill advancing out of committee, and of eventually become law, respectively. These parameter estimates are consistent with my hypothesis that containing more legal detail should be more likely to advance out of committee and ultimately succeed in Congress.

As for Hypothesis 2, the results are more mixed. We see that the inclusion of *Authorization of Appropriations* and *Technical Corrections* provisions are associated with an increased relative likelihood that a bill is reported out of committee, and that it eventually becomes law. It makes intuitive sense that the inclusion of *Technical Corrections* provisions in a bill would be positively associated with its chances for success if these tend to clarify or correct relatively minor points in existing laws. Conversely, we see that bills containing *Findings*, and *Purposes* sections are relatively less

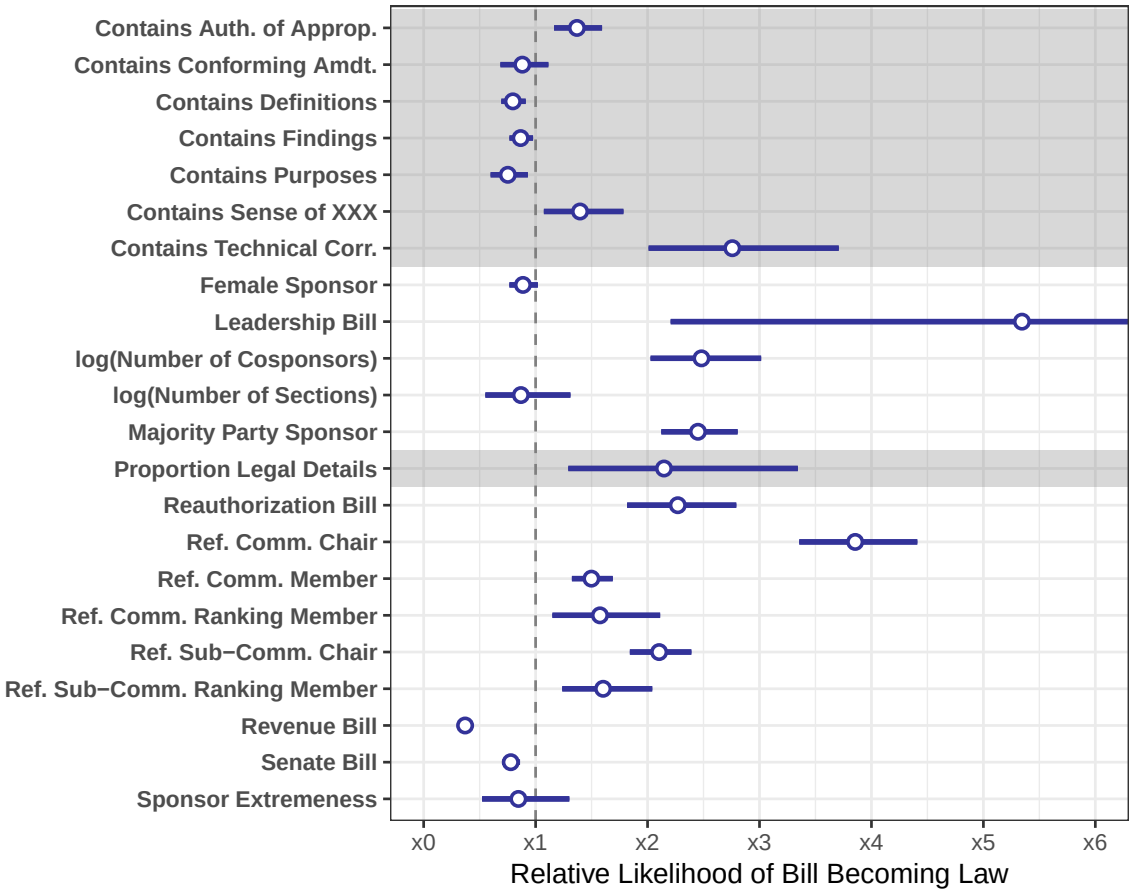

Figure 4: Marginal effects (with 95% confidence intervals) of covariates on the relative likelihood of a bill becoming law when it is first introduced. Variables of interest are highlight in gray.

likely to advance out of committee, or to become law. It may be the case that including *Findings*, and *Purposes* sections in a bill signals a position taking effort (because these provisions tend to be written in plain English), thus explaining their negative association with the relative probability of bill success. Overall, the results suggest that the inclusion of certain non-substantive provisions in legislation is predictive of bill advancement, but that these effects are not uniform. Future work should explore the role that these provisions play in greater detail.

While the regression results in Figures 3 and 4 tell a mostly consistent story about the association between increased legal and technical language content in legislation and bill advancement, they do not help us to understand where in the legislative process this language is most strongly

associated with success. This is because the effects for bills that become law could be predominantly driven by the effects on passage our of committee, or vice versa. To understand how legal language relates to bill advancement later in the legislative process, I replicated the analyses in Figures 3 and 4 but now only using the text of the version of bills that were reported out of committee (and thus only bills that were reported out of committee), and predicting whether those bills would become law. The marginal effects for this regression are reported in Figure 5. As we can see, most marginal effects are no longer different from zero, including those for legal details and the inclusion of most non-substantive provisions. This suggests that the primary impact of legal language comes in clearing the first hurdle of advancing out of committee. This is consistent with my theory that the inclusion of legal details serves as an early indicator fo bill quality.

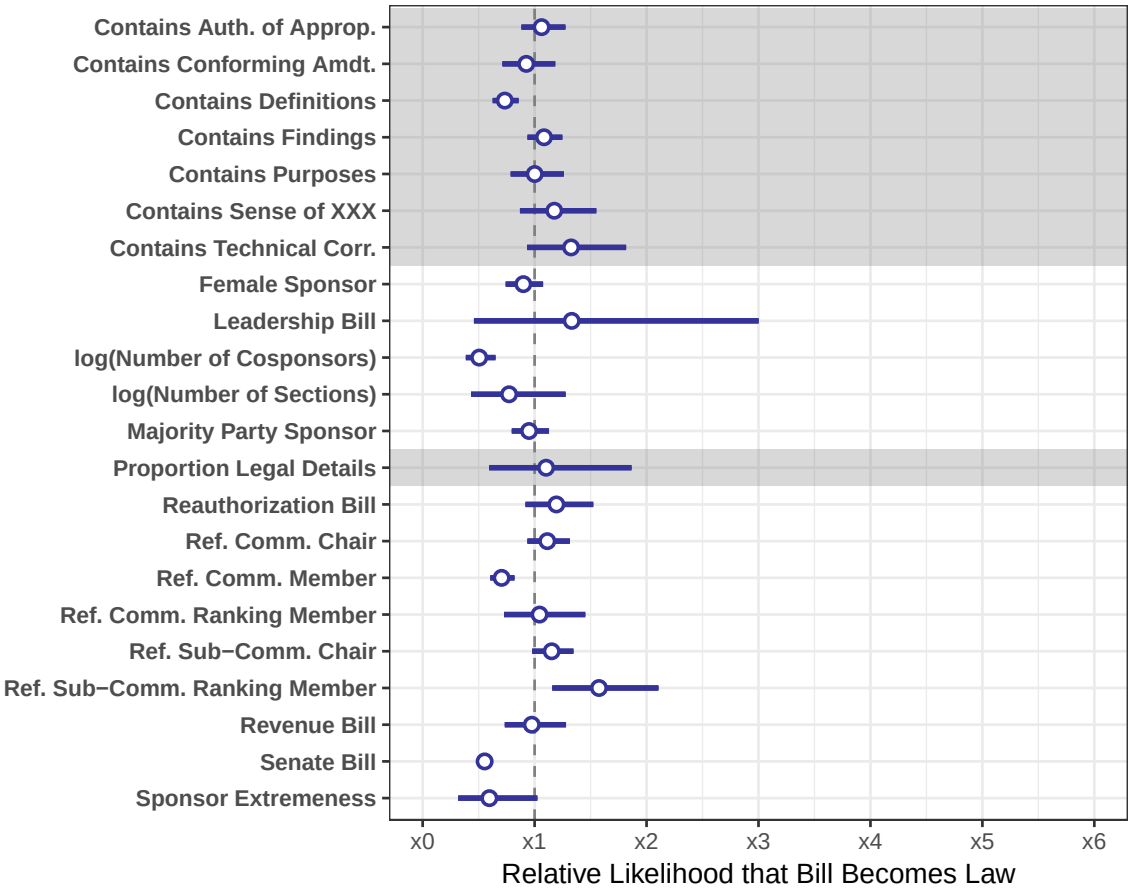

Figure 5: Marginal effects (with 95% confidence intervals) of covariates on the relative likelihood of a bill becoming law, conditional on it having been reported out of committee. Variables of interest are highlight in gray.

6 Discussion

Members of Congress introduce legislation for a variety of reasons beyond policy-making, and likely have varying expectations about the chances for success of each bill. Some bills may be almost certain to advance for institutional and procedural reasons, while others may be introduced as pure position taking stunts. While a large body of research has sought to understand why some bills succeed where others fail, scholars have lacked a way to characterize bill "quality", and to account for it in their analyses of legislative success. In this study, I focus on the inclusion of legal and technical details in legislation, along with a number of non-substantive provisions as indicators that a bill is intended as a genuine lawmaking effort. I hypothesize that bills containing more legal detail in their substantive provisions will be more likely to advance out of committee, and ultimately to become law. I argue that this is a result of legislators recognizing that bills containing more details represent more serious lawmaking efforts, and are thus worthy of further consideration.

To assess this hypothesis, I developed and employed text as data methods to identify and measure the use of legal and technical language, and non-substantive provisions in a large corpus of U.S. Congressional bills introduced between 1993 and 2016. I then estimated a logistic regression model predicting bill success, and controlling for a wide variety of factors associated with bill success that had been previously identified in the literature. My results indicated support for my first hypothesis that the inclusion of a greater proportion of legal detail in legislation is associated with an increased likelihood of that bill advancing out of committee, and ultimately of becoming law. As for my second hypothesis that bills containing several different types of non-substantive provisions will also be more likely to become law, the results are more mixed. I find that the inclusion of several types of non-substantive provisions in a bill are positively associated with its chances of advancing, but that the inclusion of some non-substantive provisions may signal a position taking effort.

One of the main limitations of my approach is that I cannot parse out whether the inclusion of this legal and non-substantive content in legislation *causes* a bill to be successful, or whether it is *consequence* of the bill's authors recognizing that it is likely to be successful for institution or procedural reasons, leading them to include this content in response. While my theory about the *importance* of this content does not rest on the direction of causality, it would be valuable to better understand the causal mechanism. Additionally, an intuitive consequence of my argument that genuine policymaking efforts will contain more legal detail is that pure position-taking bills should waste almost no time on the inclusion of such details. Yet, I find no sharp difference across successful and unsuccessful legislation, suggesting either that many position taking bills include at least some legal details, or that my measurement approach does not capture some important dimensions of bill quality. Furthermore, I do not account for "hitchhiker" bills (bills that are included in their entirety as provisions of other bills) or the reuse of specific provisions across multiple bills in my analysis, and this omission may lead me to underestimate the differences in the use of legal details across legislation.

My theory and findings also leave many open questions that should be explored in future research. For example, are most of the legal details included in legislation in the form of longer boilerplate passages, or do these details tend to be more dispersed through the text of its substantive provisions? And are there some staffers, outside interest groups, or individual legislators that are "better" at including these details in legislation than others? Furthermore, how do these legal details change as a bill moves through the legislative process? Are there some committees that have a particular propensity to add or remove these legal details as part of the committee markup process? Moving on to the inclusion of non-substantive provisions in legislation, more theoretical attention could be devoted to understanding the difference between these types of provisions in terms of their relationship to the success of a bill. For example, does the inclusion of *Findings* or *Purposes* provisions indicate that a piece of legislation was merely a platform for legislators to take positions?

My findings, along with the method I develop, also have important implications beyond the scope of this study. For example, the method I develop for identifying legal and technical language in legislation can be applied much more broadly to identify domain stopwords in political texts. While I am explicitly interested in studying how these terms are used, the method I develop can just as easily be applied to remove these terms before conducting analyses of substantive language. My approach is also easily scalable to millions of documents containing billions of words, and is implemented in an `R` package.[15] Substantively, my findings have several important implications for the study of lawmaking. Most importantly, I argue that future studies of lawmaking need to account for bill quality in assessing legislative success, and my results indicate that legislators do take at least one dimension of bill quality (the inclusion of legal detail) into account when deciding which legislation deserves further consideration. Finally, in this study, I have sought to provide evidence that the actual *words* in legislation matter. To the degree that I have been successful in this endeavor, this should spur legislative scholars to look more deeply into the text of legislation in order to better understand the lawmaking process.

[15] https://github.com/matthewjdenny/SpeedReader

References

Adler, E. Scott and John Wilkerson. Congressional Bills Project: (1980-2012), 2012. http://www.congressionalbills.org/index.html.

Adler, Scott E., T. J. Feeley, and John D. Wilkerson. Bill sponsorship activity and success in Congress: Why we should change the way we study legislative effectiveness. In *Annual Meetings of the Midwest Political Science Association*, 2003.

Anderson, William D, Janet M Box-Steffensmeier, and Valeria Sinclair-Chapman. The keys to legislative success in the US House of Representatives. *Legislative Studies Quarterly*, 28(3): 357–386, 2003.

Battiti, Roberto. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. In *IEEE Transactions on Neural Networks*, volume 5, pages 537–550, 1994.

Binder, Sarah. The Dysfunctional Congress. *Annual Review of Political Science*, 18(1):85–101, 2015. http://www.annualreviews.org/doi/abs/10.1146/annurev-polisci-110813-032156.

Binder, Sarah A. The Dynamics of Legislative Gridlock, 1947-96. *American Political Science Review*, 93(3):519–533, 1999. http://www.jstor.org/stable/10.2307/2585572.

Blumenthal, Paul. Rep. Conyers: Don't Read the Bill, jul 2009. https://sunlightfoundation.com/2009/07/27/rep-conyers-dont-read-the-bill/.

Bouma, Gerlof. Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of German Society for Computational Linguistics (GSCL 2009)*, pages 31–40, 2009. https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf.

Campbell, James E. Cosponsoring legislation in the US Congress. *Legislative Studies Quarterly*, 7(3):415–422, 1982. http://www.jstor.org/stable/10.2307/439366.

Casas, Andreu, Matthew J. Denny, and John Wilkerson. More Effective Than We Thought: Accounting for Legislative Hitchhikers Reveals a More Inclusive and Productive Lawmaking Process. 2018. https://ssrn.com/abstract=3098325.

Church, Kenneth Ward and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29, 1990. http://www.aclweb.org/anthology/J90-1003.

Clark, Tom S. The Separation of Powers, Court-Curbing and Judicial Legitimacy. *American Journal of Political Science*, 53(4):971–989, 2009. http://www.jstor.org/stable/20647961.

Cover, Thomas M. and Joy A. Thomas. *Elements of Information Theory*. John Wiley and Sons, 1991. http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471241954.html.

Craig, Alison W., Janet M. Box-Steffensmeier, and Dino P. Christenson. Cue-Taking in Congress: Interest Group Signals from Dear Colleague Letters. *American Journal of Political Science*, 2018. http://visionsinmethodology.org/wp-content/uploads/sites/4/2016/05/Craig-Crafting-a-Broad-Appeal.pdf.

Denny, Matthew J. More than Control: Partisan Differences in the Use of Statutory Constraints on the Bureaucracy. 2018. https://ssrn.com/abstract=3154577.

Denny, Matthew J., Brendan O. Connor, and Hanna Wallach. A Little Bit of NLP Goes A Long Way: Finding Meaning in Legislative Texts with Phrase Extraction. In *Midwest Political Science Association Anual Meeting*, 2015.

Desmarais, Bruce A., Vincent G. Moscardelli, Brian F. Schaffner, and Michael S. Kowal. Meaningful Collaboration and Legislative Success in the U.S. Senate. 2013.

Eskridge, William N., Abbe R. Gluck, and Victoria F. Nourse. *Statutes, Regulation, and Interpretation: Legislation and Administration in the Republic of Statutes*. West Academic Publishing, St. Paul, MN, 2014. https://searchworks.stanford.edu/view/10745254.

Estévez, Pablo A., Michel Tesmer, Claudio A. Perez, and Jacek M. Zurada. Normalized Mutual Information Feature Selection. *IEEE Transactions on Neural Networks*, 20(2):189–201, 2009. http://ieeexplore.ieee.org/document/4749258/.

Fano, Robert M. *Transmission of Information: A Statistical Theory of Communications*. M.I.T. Press, 1961. https://mitpress.mit.edu/books/transmission-information.

Fleuret, Francois. Fast Binary Feature Selection with Conditional Mutual Information. *Journal of Machine Learning Research*, 5:1531–1555, 2004. http://dl.acm.org/citation.cfm?id=1044711.

Grimmer, Justin and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, jan 2013. http://pan.oxfordjournals.org/cgi/doi/10.1093/pan/mps028.

Grossmann, Matt and Kurt Pyle. Lobbying and congressional bill advancement. *Interest Groups & Advocacy*, 2(1):91–111, 2013. http://link.springer.com/10.1057/iga.2012.18.

Guyon, Isabelle and André Elisseeff. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, 3(3):1157–1182, 2003. http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf.

Hall, Richard L and Alan V Deardorff. Lobbying as Legislative Subsidy. *American Political Science Review*, 100(1):69–84, 2006. http://www.jstor.org/stable/27644332.

Hall, Richard L. and Frank W. Wayman. Buying Time: Moneyed Interests and the Mobilization of Bias in Congressional Committees. *The American Political Science Review*, 84(3):797–820, 1990. http://www.jstor.org/stable/10.2307/1962767.

Handler, Abram, Matthew J. Denny, Hanna Wallach, and Brendan O'Connor. Bag of What? Simple Noun Phrase Extraction for Text Analysis. In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science at the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016. https://brenocon.com/handler2016phrases.pdf.

Harward, Brian M. and Kenneth W. Moffett. The calculus of cosponsorship in the US Senate. *Legislative Studies Quarterly*, 35(1):117–143, 2010. http://onlinelibrary.wiley.com/doi/10.3162/036298010790821950/abstract.

Hasecke, EB and JD Mycoff. Party Loyalty and Legislative Success: Are Loyal Majority Party Members More Successful in the U.S. House of Representatives? *Political Research Quarterly*, 60(4):607–617, 2007. http://prq.sagepub.com/content/60/4/607.short.

Huber, John D. and Charles R. Shipan. *Deliberate Discretion? The Institutional Foundations of Bureaucratic Autonomy*. Cambridge University Press, 2002. https://doi.org/10.1017/CBO9780511804915.

Kessler, Daniel and Keith Krehbiel. Dynamics of Cosponsorship. *American Political Science Review*, 90(3):555–566, 1996. http://www.jstor.org/stable/10.2307/2082608.

Krehbiel, Keith. Cosponsors and Wafflers from A to Z. *American Journal of Political Science*, 39(4):906–923, 1995. http://faculty-gsb.stanford.edu/krehbiel/MyPDFs/96JoPCosponsorsandWafflers.pdf.

Matthews, Donald R. *United States Senators and Their World*. University of North Carolina Press, 1960.

Mayhew, David R. *Congress: The Electoral Connection*, volume 26. Yale University Press, 1974.

Mayhew, David R. *Divided We Govern*. Yale University, 1991.

Miquel, Gerard Padro I. and James M. Snyder. Legislative Effectiveness and Legislative Careers. *Legislative Studies Quarterly*, 31(3):347–381, 2006. http://onlinelibrary.wiley.com/doi/10.3162/036298006X201841/abstract.

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

O'Connor, B. MITEXTEXPLORER: Linked brushing and mutual information for exploratory text data analysis. In *ACL 2014 Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014. http://aclweb.org/anthology/W14-3101.

O'Connor, Brendan, David Bamman, and Noah A. Smith. Computational Text Analysis for Social Science: Model Assumptions and Complexity. In *Second Workshop on Computational Social Science and Wisdom of the Crowds (NIPS 2011)*, 2011. http://repository.cmu.edu/lti/212/.

Peng, Hanchuan, Fuhui Long, and Chris Ding. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005. http://ieeexplore.ieee.org/document/1453511.

Recchia, Gabriel and Michael N. Jones. More data trumps smarter algorithms: comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, 41(3): 647–656, 2009. https://www.researchgate.net/publication/26656397.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082, 2014.

Role, François and Mohamed Nadif. Handling the Impact of Low Frequency Events on Co-Occurrence Based Measures of Word Similarity: A Case Study of Pointwise Mutual Information. In *KDIR 2011 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, number September 2016, pages 226–231, 2011. http://www.scopus.com/inward/record.url?eid=2-s2.0-84862174370{&}partnerID=tZOtx3y1.

Schneider, Karl-Michael. Weighted Average Pointwise Mutual Information for Feature Selection in Text Categorization. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 252–263, 2005. http://dl.acm.org/citation.cfm?id=2101264.

Shannon, Claude E. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 1948. http://dl.acm.org/citation.cfm?id=584093.

Strokoff, Sandra. How Our Laws Are Made: A Ghost Writer's View. *The Philadelphia Lawer, Philadelphia Bar Association Quarterly Magazine*, 59(2), 1996. https://legcounsel.house.gov/HOLC/Before{_}Drafting/Ghost{_}Writer.html.

Thomas, Scott J. and Bernard Grofman. Determinants of Legislative Success in House Committees. *Public Choice*, 74(2):233–243, 1992. http://link.springer.com/article/10.1007/BF00140770.

Wang, Gang, Frederick H. Lochovsky, and Qiang Yang. Feature Selection with Conditional Mutual Information MaxiMin in Text Categorization. In *Proceedings of the Thirteenth ACM conference on Information and knowledge management - CIKM '04*, pages 342–349, 2004.

Wilkerson, John, David Smith, and Nicholas Stramp. Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach. *American Journal of Political Science*, 59(4):943–956, 2015.

Wilson, Rick K. and Cheryl D. Young. Cosponsorship in the U. S. Congress. *Legislative Studies Quarterly*, 22(1):25–43, 1997. http://www.jstor.org/stable/10.2307/440289.

Woon, Jonathan. Issue Attention and Legislative Proposals in the U.S. Senate. *Legislative Studies Quarterly*, 34(1):29–54, 2009. http://www.jstor.org/stable/20680226.

Yang, Howard Hua and John Moody. Feature Selection Based on Joint Mutual Information. In *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, pages 22–25, 1999. https://www.icsi.berkeley.edu/icsi/node/2182.

Yang, Yiming and Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning*, pages 412–420, 1997. http://dl.acm.org/citation.cfm?id=657137.

Yano, Tae, Noah a Smith, and John D Wilkerson. Textual Predictors of Bill Survival in Congressional Committees. *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 793–802, 2012.

A Term-Category Associations

The task of determining term-category associations has a long history, and methods for determining these associations have been profitably applied across a number of fields. For example, *pointwise mutual information*, or PMI (see, for example: Fano 1961; Church and Hanks 1990; Cover and Thomas 1991) is a commonly used statistic for looking at the relationships between terms and categories in a corpus of documents (O'Connor et al. 2011; O'Connor 2014). PMI is calculated on a joint distribution, which can be easily formed from a contingency table by dividing each cell in the contingency table by the sum of the table. PMI tells us how much information knowing the particular value of a categorical variable (e.g. the probability of a particular vocabulary term appearing in a given category) gives us about which category we are in, and vice versa. For given values of discrete random variables $c \in C$ and $v \in V$, the PMI of these two values is defined as:

$$PMI(c;v) = \log\left(\frac{p(c,v)}{p(c)\,p(v)}\right) \tag{2}$$

PMI is often used to find term *collocations*, or terms that tend to appear together (Bouma 2009; Recchia and Jones 2009; Role and Nadif 2011), but can also be used to find words that are highly associated with a given category label (O'Connor 2014). PMI, along with mutual information (the expected value of PMI over the entire joint distribution), have also been used extensively for feature selection[16] in machine learning applications (see, for example: Battiti 1994; Yang and Pedersen 1997; Yang and Moody 1999; Guyon and Elisseeff 2003; Fleuret 2004; Wang et al. 2004; Peng et al. 2005; Estévez et al. 2009).

The task of feature selection is deeply related to the problem of identifying legal and technical language in legislation, except that the focus is placed primarily on finding terms that strongly *distinguish* between categories (policy terms). In the most directly relevant work, Schneider (2005) introduce a weighted average PMI based approach to finding the optimal set of text features to distinguish between a set of categories. However, this approach is not ideal in the current application. The reason is that this method looks for terms that distinguish between all categories simultaneously, but in my application, I want to find words that distinguish between Democrat and Republican sponsored legislation *on average* (controlling for issue area). Thus, this approach will likely include some substantive policy terms such as "increase spending" with legal and technical terms if they are used similarly across issue areas.

Turning to the political science literature, the "Fightin' Words" feature selection method of Monroe et al. (2008) represents another promising approach to identifying legal and technical terms in legislation. Monroe et al. introduce a method for finding terms that statistically distinguish between documents written by members of different parties, which could also be used to find legal and technical terms (those that do not distinguish between parties, on average). However, because the Fightin' Words feature selection method was specifically designed to find terms that *do* distinguish between categories, it does not produce a natural cutoff for identifying terms that actively make it *harder* to distinguish between categories. I build off of the intuition laid out in Monroe et al. (2008) to develop a statistical approach that is specifically tailored to identifying terms that

[16] Feature selection involves picking features that produce good classification accuracy.

make it harder to distinguish between categories.

B ACMI Contributions

The mutual information of two discrete random variables C and V is defined as:

$$I(C;V) = \sum_{c \in C} \sum_{v \in V} p(c,v) \log \left(\frac{p(c,v)}{p(c)\,p(v)} \right) \tag{3}$$

In this application, we are working with a large number of joint distributions formed by normalizing the counts of terms in Democrat and Republican bills, in a particular issue area. Our goal is to determine how much each term contributes to the mutual information of these conditional joint distributions, on average. Thus, we end up calculating the mutual information contributions of each term in these 2 $\times$ V conditional joint distributions, and then averaging these contributions across all conditional joint distributions to get that term's ACMI. Below, I formalize the method for calculating ACMI.

For reasons of notational sanity, let T be the full contingency table described in Section 3.3, and let T_k, ($k \in K$) be the k'th Democrat/Republican row pair (on a given issue area) in this contingency table. So for example, one row pair could be: Democrat bills about nuclear energy policy; and Republican bills about nuclear energy (two rows from the full contingency table). Furthermore, let J_k be the conditional joint distribution implied by dividing all of the cells in T_k by the sum of T_k (normalizing). Now, let $J_k^{(-v)}$ be a new conditional joint distribution generated by removing the v'th column from J_k and then re-normalizing. This new conditional joint distribution would have V-1 columns, and still sum to 1. Additionally, let $\sum(T_k)$ be the sum of term counts in the k'th Democrat/Republican row pair, and $\sum(T)$ be the sum of term counts in the full contingency table. Finally, we will denote the mutual information of the k'th conditional joint distribution as $I(J_k)$.

Adopting the notation described above, the mutual information contribution of term v in conditional joint distribution J_k is just $I(J_k) - I\left(J_k^{(-v)}\right)$. To get the full ACMI for term v, we then just take the weighted average over its mutual information contributions in each J_k as follows:

$$\text{ACMI}\ (v) = \sum_{k \in K} \frac{\sum(T_k)}{\sum(T)} \left[I(J_k) - I\left(J_k^{(-v)}\right) \right] \tag{4}$$

Here, the weight we give to the term mutual information contributions we calculate for each conditional joint distribution is just the proportion of all terms in the full contingency table. We do this type of averaging because we want to rely more on what we learn from category pairs that are associated with lots of tokens (documents), as opposed to those that are associated with only a few.

C Efficient Optimization: Decomposing Mutual Information

To speed up the ACMI calculations, I decomposed the mutual information contribution of each term in the vocabulary. We start with the general form of mutual information, where $c \in C$

(categories) and $v \in V$ (vocabulary terms) are discrete random variables with joint probability $p_{c,v}(c,v)$ and marginal probabilities $p_c(c)$ and $p_v(v)$, respectively:

$$I(C;V) = \sum_{v \in V} \sum_{c \in C} p_{c,v}(c,v) \log \left(\frac{p_{c,v}(c,v)}{p_c(c)\, p_v(v)} \right) \tag{5}$$

In the present application, the joint distributions will have two rows (Democrat and Republican term counts) and a large number of columns (terms in the vocabulary). What we want to find is the difference in mutual information if we remove one column (indexed by v) from the matrix:

$$\Delta(I)_{-v} = I(C;V_{-v}) - I(C;V) \tag{6}$$

If $\Delta(I)_{-v}$ is positive, then term v makes and negative contribution to mutual information in this distribution, and if $\Delta(I)_{-v}$ is negative, then it makes a positive contribution to mutual information in this distribution. We begin with the direct effect DE_v of removing v, which is the removal of the following two terms from the sum:

$$DE_v = p_{c,v}(c_1,v) \log \left(\frac{p_{c,v}(c_1,v)}{p_c(c_1)\, p_v(v)} \right) + p_{c,v}(c_2,v) \log \left(\frac{p_{c,v}(c_2,v)}{p_c(c_2)\, p_v(v)} \right) \tag{7}$$

These values can be cached in a straightforward way by recording their sum in a vector and simply subtracting that sum from the total mutual information with that column included. The challenging calculation comes in through the indirect effects IE_v on the other values in the sum. $p_{c,v}(c,v)$ will be affected through the denominator (we are removing some number of words from the corpus), so the denominator in each case will need to be multiplied by

$$D = \frac{\sum(C;V)}{\sum(C;V) - \sum v} \tag{8}$$

For those terms outside of the log, this effect is common across all terms, so we can multiply the whole sum by D. We see that $p_{c,v}(c,v)$ also enters inside of the log, where we need to do the same multiplication. Fortunately, we can separate the log of a product into the sum of logs as:

$$\log(xy) = \log(x) + \log(y) \tag{9}$$

so we can just take the sum of $\log(D)$ over the non-zero terms in $I(C;V_{-v})$. Finally, for the $p_c(c_i)$ terms, we will need to perform a similar multiplication:

$$D_c = \frac{\sum(c_i;V)}{\sum(c_i;V) - (c_i;V_v)} \tag{10}$$

Thus we just need to keep track of the number of non-zero entries in the first and second rows (NZ_1, NZ_2), and we can use these counts to make a similar log addition adjustment. The critical point here is that the only entries we have to care about are the non-zero entries. We can therefore take advantage of sparsity in the category term distributions (most terms have zero count in a given distribution). Now we can use this decomposition of the effect of removing term v from the

vocabulary to efficiently calculate $I(C;V_{-v})$, which then gives us $\Delta(I)_{-v}$ (our objective).

$$I(C;V_{-v}) = D \times \left[I(C;V) - D + \sum_{i} \sum_{I(C_i;V_{-v})\neq 0} \log(D_c) \right] \tag{11}$$

Using an $\{i, j, v\}$ sparse matrix representation of the document term matrix (where i denotes the row index, j denotes the column index, and v denotes the non-zero value at i, j) makes this computation very efficient, facilitating fast vocabulary partitioning. After implementing this method, I compared the results of this method to the naive approach and confirmed that it yields identical results.