# Measuring the Policy Content of Congressional and American State Legislation Using Machine Learning

Ethan Dee, Ph.D.* and Alex Garlick, Ph.D.†

November 26, 2024

**Abstract**

We use a machine learning model based on the transformer architecture to replicate and expand the Comparative Agenda Project's coverage of American legislatures. Our model is jointly trained on pre-coded Congressional and Pennsylvania legislation and it compares favorably to extant supervised machine learning models. Using Pennsylvania as a keystone allows us to bridge the national and state legislative contexts, and produce 1.687 million estimates of the leading policy in legislative documents from Congress and the 50 state legislatures since about 2009. Validations show the model agrees with human-coders on the vast majority of policy assignments, and the disagreements are based more on inconsistencies in the codebook's logic than random error. We discuss the challenges with applying a model like this to the study of legislative institutions.

This draft was prepared for the 2024 Artificial Intelligence and the Study of Political Institutions Conference at the University of Southern California, Los Angeles, California. Please do not cite or circulate this version. The output of the data described herein are available here: https://osf.io/e2unp/?view_only=e0302a513e7e4cc9999feab2045d47b3.

## 1 Background & Summary

Observing legislative agendas has allowed scholars of public policy or political institutions to answer important questions about the behavior of American political elites, such as why do they pay sporadic attention to certain issues (Baumgartner and Jones 2002), why members of Congress are polarized (Lee 2009, Lapinski 2013), and if government officials represent the wishes the many or a privileged few (Gilens and Page 2014, Barberá et al. 2019). The largest project to propagate data on policy agendas of US national institutions is the Comparative Agendas Project (CAP, née Policy Agendas Project), originally produced by Frank Baumgartner and Bryan Jones. CAP includes the Congressional Bills Project (Adler and Wilkerson 2015) to observe the policies legislated by the US Congress, as well the policy content of a wide range of political documents generated by parties, the media and public officials, such as the sentences uttered by the president in the annual State of the Union.

While the project has expanded internationally (e.g. Širinić and Čakar 2019), there's little coverage of US sub-national institutions, restricting the ability of the project to address questions of federalism or state/local politics. To be fair, this is a limitation of American politics beyond CAP (Anzia 2019). An exception being the Pennsylvania Policy Database Project (PAPDP, McLaughlin et al. 2010) which produced a state-specific code book, as state legislators and Members of Congress attend to different issues. The codebook accounts for how members of

*Amazon, Inc.

†Assistant Professor at the University of Vermont; Department of Political Science; alex.garlick@uvm.edu

Congress spend time on foreign embassies and diplomats, while state legislators are worried about fire stations and police officers. PAPDP employed hand-coders to measure the policy content of the Pennsylvania legislature from 1979-2010, but this raises two concerns. First, hand-coding has a constant returns to scale and there's simply too many bills introduced in all 50 states on an annual basis for any team to keep up. Second, national and state agendas are not automatically analogous, shown by the PAPDP both to trimming and expanding the list of polices to properly cover the body's work.

This article draws on recent advances in machine learning (ML) to overcome these difficulties and put the state legislatures and Congress in a common space. Recent versions of the CBP have used machine learning approaches to code Congressional bills (Hillard, Purpura, and Wilkerson 2008, Collingwood and Wilkerson 2012), successfully replicating the work of hand-coders. But as a point of reference, the CBP and PAPDP strove for 90% inter-coder reliability at the major topic-level. Our first step is to use a newer generation of natural language processing tools to replicate the ML-generated codes in the CBP. Next, we use the PAPDP's codes of the Pennsylvania legislation as a bridge between the federal and state contexts. After verifying the accuracy of its Pennsylvania estimates, we ensure our approach is not overfit on this one state by training and validating the Illinois legislative record, a state which creates the second most amount of legislation behind only New York. Having calibrated the model for the states, we then code the remaining 48 states from 2009-2023, which altogether is 1.591 million state legislative documents (including bills and resolutions).

Technically, there are two major innovations that differentiate our model from the extant ML approaches to coding Congressional legislation. The previous ML models on Congress used a "bag of words" model. These researchers would first pre-process text (reducing punctuation, changing capitalization, word-stemming and/or lemmatiziation), and then consider bills as unordered groups of the words. This can lose important context, such that the numerical equivalent of "this bill is about jails not hospitals" would be identical to "this bill is about hospitals not jails." Instead, we use the word-piece embedding approach, which is built off the intuition that "You shall know a word by the company it keeps" (Firth 1957). Word embeddings have been used in major consumer products like the Google search engine, and have been described in great detail (Wu et al. 2016), but a demonstration of their ability to track context is that the model would consider $king - man + woman = queen$. The next major departure is to lean on the transformer architecture (Vaswani et al. 2017), which allows a word's embedding to vary depending on the words that co-occur with it. A major innovation in transformer models is "self-attention," where the model weighs the importance of words in the input sequence as they pertain to a focus word, generated by its training data. Our methods section details how we adapt the transformer model to American legislative data, in particular how we leverage situations where hand-coders disagreed on bills, even with the same title.

We demonstrate the internal and external validity of these estimates with a number of tests. Compared to the extant bag of words models, we document a minor, but tangible improvement (See Appendices 5.1.1 and 5.1.2). We then test the results in three ways: first, within the Congressional setting, i.e. on Congressional sessions that are temporally out-of-sample with respect to the training data. Second, we create a series of "synthetic" bills, or documents featuring terms which we have a prior belief on where they should be assigned, such as "coronavirus" or the abbreviation "UVM." The model correctly places "coronavirus" in health, and "UVM" in education, even though neither was in the training corpus. Finally, we conducted another out-of-domain test, by assigning the subtopic descriptions for the CAP master codebooks. These are overwhelmingly assigned to the correct category, and the exceptions reveal actual disagreements in contemporary politics, as the model coded "tax administration, enforcement and auditing" as macroeconomics, but the codebook slates it under "Government Operations." So our algorithm joins a list of political philosophers dating back to John Locke or George Harrison who have pondered the nature of taxes. The state legislative output has fewer options for comparison, but we show that Illinois bills are overwhelmingly referred to the expected committee of jurisdiction.

In general, we follow the convention of the CAP to assign a single "leading" policy to each bill. Jones himself has noted that multiple codes may more faithfully represent the policy topics

of the underlying legislation (see our earlier our discussion of taxes, which are a matter of government operations but clearly affect the macroeconomy), but stresses that single-codes are necessary for maintaining "time-series consistency" (Jones 2016). But only including a single code would waste important information generated by the model, so we report "confidence scores," with which we verify that it is calibrated in the sense that higher confidence denotes higher probability of a true positive. A researcher looking to use a calibrated model as a companion or replacement for hand-coder efforts can, as is done in Collingwood and Wilkerson 2012, combine the model's high-confidence predictions and hand-coders' intervention for low-confidence predictions to produce the final output dataset. Or, they can inspect topic conflation and assess whether the codes or bills themselves should be re-evaluated. Also, the model's confidence scores can be used to rank-order its predictions, forming statements about its "top-K" predictive accuracy, and identifying instances of split confidence wherein a bill might more accurately be reflected as multi- as opposed to single-topic. This can allow a researcher looking for a broad sweep of a single policy area to detect bills where a topic is a significant, if not leading, consideration.

A great deal of public attention has been recently drawn toward the use of Large Language Models (LLMs), including those at the heart of our ML model, for generative AI, with models such as Chat-GPT and Gemini. However, the use of proprietary models for research has also raised concerns, including about equity and reproducibility (Palmer, Smith, and Spirling 2024). This paper shows a tremendous opportunity in fine-tuning a component of the popular models, the transformer architectures, for the applied task of supervised multi-class classification task of assigning bills to policy areas. Compared to earlier "bag of words" ML models, the transformer architecture eliminates much of the pre-processing, which can make it simpler to replicate. Also the models perform well with far less training data, allowing researchers to implement automated methods with a far lower initial labor investment. Future work could amend this model to consider politician rhetoric, interest group witness testimony, or any number of political texts.

Our next section covers the technical methods of the method, and each major conceptual step going from Congress to Pennsylvania to Illinois to the remaining 48 states. Then we have an extensive validations section, showing both the internal and external validity of the model. We conclude with a discussion of the data records and offer usage suggestions to other researchers using these data.

# 2 Methods

The Comparative Agendas Project (CAP) offers researchers unparalleled observation of elite political behavior within American national institutions, and increasingly in European countries, but poses two challenges for replication or extension. First, as previously mentioned, CAP requires coders to identify a single "leading" topic for whatever they are sorting. Second, its hand-coders demonstrate a 90% agreement rate. If our model perfectly emulated the hand-coder behavior, our results would be limited to 90% accuracy. In this section we will detail how our model is built and how it addresses these two CAP-specific concerns, as well as building a bridge across the American federalism system to code state legislative data as well. The next section contains our internal and external validation exercises; this section includes a number of tests taken to fit the model to inform how it was optimized for the American federal context.

## 2.1 Technical approach

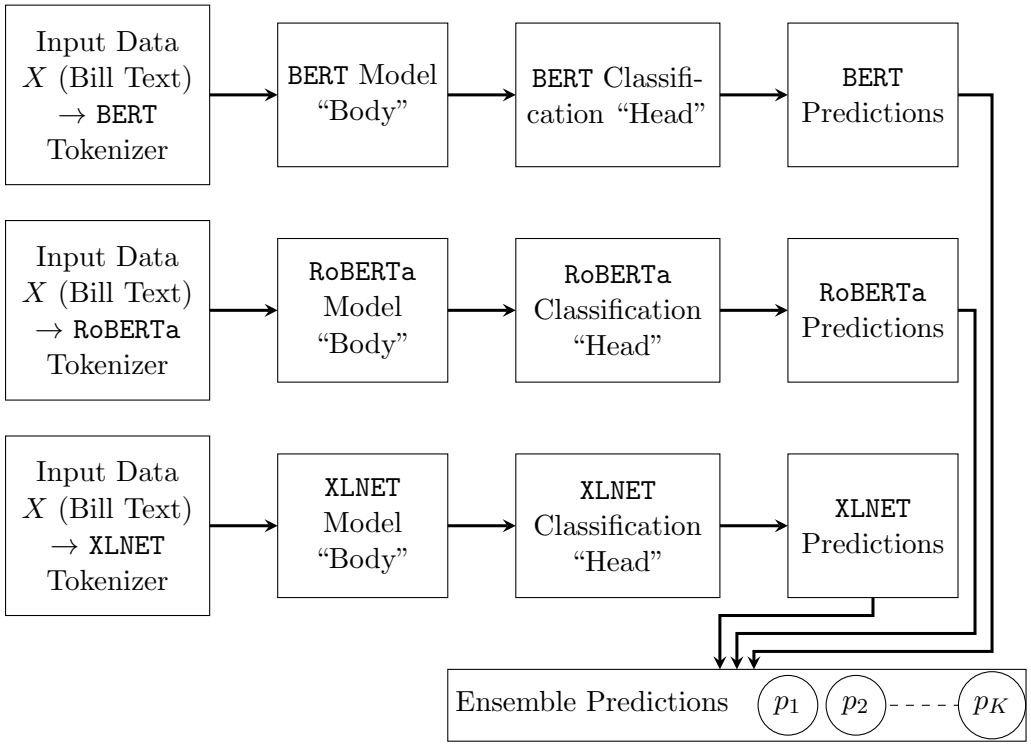
The introduction described the conceptual advancement from "bag of words" models to models which come from the family of transformer-based deep neural networks,[1] which use contextual word-piece embedding representations of documents. Specifically, we use an ensemble model

1. For a thorough discussion of the inner workings of the transformer architecture, see: http://nlp.seas.harvard.edu/annotated-transformer/

encompassing: (1) BERT, (2) RoBERTa, which extends BERT's masked language modeling pre-training task and does away with its next sentence prediction task. (3) `XLNET` (Yang et al. 2020), which uses a technique called "permutation language modeling" and is meant to excel at capturing long-range dependencies. `XLNET` also allows more than 512 word-piece tokens on the input sequence, unlike RoBERTa and BERT allowing us to accommodate bills with unusually long titles and descriptions. We scraped the bill titles and descriptions from legiscan.

Figure 1 presents the transformer-based ensemble model, which combines three common transformer-based architectures - BERT, RoBERTa, and `XLNET` - to classify legislation into policy areas.

Figure 1: Ensemble Model Architecture




Notes: In a multi-class classification task, $\sum_{k \in K} p_k = 1$, and the maximum $p_{k,k \in K}$ is taken to be the model's prediction for the bill's topic, $\hat{y}$. In a multi-label classification task, the probabilities need not sum to 1. For either classification task, the 1024-dimensional document embedding generated by the RoBERTa-large model "body" forms the input for the classification "head," which is itself a deep neural network. The model output is technically the output layer of the classification head, but the head is decomposed in this way to illustrate the specific form of the output.

To explain its inner workings, we will focus on RoBERTa, but the discussion is nearly identical for all three architectures. The RoBERTa model forms the "body" of a larger deep neural network designed to predict the topic(s) to which a bill attends. The input data (the bill's title or summary) is converted to numerical token IDs reflecting RoBERTa's word-pieces ("tokenization"). In the "large" version of the RoBERTa model that we utilize each word-piece is represented in the network by a 1024-dimensional contextual embedding. Eventually, RoBERTa generates a document-level embedding - an impression of the document as a whole. The model's classification "head" consists of several layers of a fully-connected, feed-forward neural network, which takes as input the model body's vector.[2] The classification head's output is a set of $K$ logits - one for each topic - denoting the pseudo-probability the model assigns to the presence of each topic $k \in K$. In the case of multi-class classification, the maximum of those

2. For RoBERTa-large, this vector has a dimensionality of 1024.

logits is taken to be the model's predicted topic for the bill.

The ensemble model has three sets of inputs - the input data $X$ fed through the tokenizers specific to BERT, RoBERTa, and `XLNET` - and four sets of outputs, three of which are the individual underlying transformer-based architectures' topic probabilities for the document, and the fourth is a simple feedforward neural network which concatenates the three models' topic probabilities to generate a fourth set of topic probabilities. This fourth component effectively creates a "meta-model" of the three constituent models, the value-added from which is granting the model the ability to learn the topic-specific strengths and weaknesses of each constituent model, and prioritize the input from each in generating its predictions accordingly. For example, if the BERT model appears to severely under-perform with a particular topic $k_A \in K$, the meta-model can down-weight BERT's input regarding this topic and defer to RoBERTa and `XLNet`.

## 2.2 CAP Specifications

The above model was trained on the Congressional Bills Project (CBP) hand-coded data. The CBP is an offshoot of Comparative Agendas Project[3] (CAP) itself a culmination of projects adhering closely to the Policy Agendas Project (PAP, Baumgartner and Jones 2002) codebook. It constructs a classification system defining 21 "major topics" denoting broad policy areas, shown in Table 1, the 220 "subtopics" minor topic codes nested inside the major topics are beyond the scope of this paper. The codebook has evolved over time, as the "Culture" and "Immigration" major topics used to be subtopics of "Education" and "Labor," respectively. In the bottom row we include codes necessary to later bridge to the state legislative context: "Local Government" and "Private Bills."

Table 1: CAP Major Topics

| | | |
|---|---|---|
| (0100) Macroeconomics | (0200) Civil Rights | (0300) Health |
| (0400) Agriculture | (0500) Labor | (0600) Education |
| (0700) Environment | (0800) Energy | (0900) Immigration |
| (1000) Transportation | (1200) Law and Crime | (1300) Social Welfare |
| (1400) Housing | (1500) Domestic Commerce | (1600) Defense |
| (1700) Technology | (1800) Foreign Trade | (1900) International Affairs |
| (2000) Government Operations | (2100) Public Lands | (2300) Culture |
| (2400) *Local Government Ops.* | (9999) *Private bills* | |

*Emphasis on codes included to build a state-federal common space.*

We took care to address the challenges that can arise when hand-coding data. Namely, the hand-coded data are occasionally measured with error, with the project "[striving] for 90% interannotator reliability at the major topic level, and 80% at the subtopic level during the training process."[4] Given the inherent complexity in deciding on the "leading" policy area to which a bill attends, this is an impressive rating, as the authors report most discrepancies reflect disagreements about a bill's primary topic. But this measurement error complicates the task of training a supervised machine learning model on these data. Trained to emulate the hand-coders of the CBP, the model attains an $F_1$ score of roughly 90%.

For the model training procedure, we kept bills with duplicate bill titles grouped together in either the training or validation data. The model's training data comprise 85% of all bills hand-coded by the CBP, and it is evaluated primarily for its $F_1$ score on the validation data, both across topics and overall. The partitioning of bills into the training and validation sets is done at-random. The model trains until validation loss no longer decreases.

3. https://www.comparativeagendas.net/
4. http://www.congressionalbills.org/codebooks.html

Table 2 presents the model's performance on the validation data (46,062 bills). All topics achieve an $F_1$ score of 81% or more, apart from "Culture," which is significantly under-sampled and a relatively new addition to the Congressional Bills Project dataset. Several topics clear the 90% inter-coder reliability threshold expected of hand-coders. The model achieves a global micro- and macro-average $F_1$ score of 90% and 87%, respectively.[5] The model is essentially a perfect classifier with respect to the "Private Bills" topic. While there are multiple topics that fall beneath the 90% threshold expected of the CBP hand-coders, the "gap" between the model's per-topic $F_1$ score and 100% is strongly correlated with the per-topic hand-coder inconsistency, with a correlation of roughly $-0.93$.

Table 2: Model Performance on Validation Data (CBP, 1947-2017)

| Topic | Precision | Recall | $F_1$ Score | Support | Inconsistency* |
|---|---|---|---|---|---|
| Agriculture | 0.91 | 0.89 | 0.90 | 1547 | 8.5% |
| Civil Rights | 0.72 | 0.91 | 0.81 | 839 | 20.2% |
| Culture | 0.61 | 0.88 | 0.72 | 32 | 22.9% |
| Defense | 0.91 | 0.84 | 0.87 | 3341 | 10.2% |
| Domestic Commerce | 0.86 | 0.84 | 0.85 | 2355 | 13.7% |
| Education | 0.89 | 0.92 | 0.91 | 1512 | 9.7% |
| Energy | 0.89 | 0.92 | 0.91 | 1296 | 9.1% |
| Environment | 0.87 | 0.88 | 0.87 | 1363 | 12.1% |
| Foreign Trade | 0.92 | 0.93 | 0.93 | 1956 | 8.4% |
| Government Operations | 0.91 | 0.83 | 0.87 | 4887 | 12.7% |
| Health | 0.91 | 0.93 | 0.92 | 2741 | 9.0% |
| Housing | 0.81 | 0.89 | 0.85 | 833 | 15.9% |
| Immigration | 0.88 | 0.93 | 0.90 | 577 | 11.3% |
| International Affairs | 0.78 | 0.85 | 0.81 | 852 | 16.7% |
| Labor | 0.85 | 0.88 | 0.86 | 1493 | 17.3% |
| Law and Crime | 0.88 | 0.87 | 0.88 | 2168 | 13.6% |
| Macroeconomics | 0.81 | 0.81 | 0.81 | 1749 | 21.5% |
| Private Bills | 0.99 | 0.99 | 0.99 | 8041 | 1.7% |
| Public Lands | 0.92 | 0.90 | 0.91 | 3860 | 10.5% |
| Social Welfare | 0.87 | 0.89 | 0.88 | 1773 | 15.7% |
| Technology | 0.84 | 0.91 | 0.87 | 686 | 11.1% |
| Transportation | 0.87 | 0.91 | 0.89 | 2161 | 10.4% |
| Micro Average | 0.90 | 0.90 | 0.90 | 46062 | Correlation with $F_1$ |
| Macro Average | 0.86 | 0.89 | 0.87 | 46062 | $-0.9280$ |

Notes: "Micro Average" computed across all topics denotes that all true positives and false positives are counted globally, ignoring which topic they came from. This aggregation strategy means each topic contributes to the global average proportional to the number of bills in the support. A "macro averaging" strategy first computes the per-topic value, and then takes a simple average of the each topic's value. Each topic thereby contributes equally to the global average. *"Inconsistency" denotes the hand-coder disagreement rates for each topic.

Given the challenges presented by taking the hand-coded data to be the ground truth, it is an open question whether a global model performance close to 100% is achievable, or even desirable. If the hand-coded data are known to be inconsistent at a rate of 10%, and a model reproduces these data perfectly (achieving an $F_1$ score of 100%), the model is, itself, inconsistent at a rate of 10%. On the other hand, an accuracy of 90% suggests an interval bounding its "true"

5. The macro-average is above 90% when the "Culture" topic is excluded.

accuracy - where hand-coder inconsistency does not exist - between $[80\%, 100\%]$. The question that determines where in this interval the model's true accuracy may exist is whether the model-versus-hand-coder disagreements lead the researcher to side with the hand-coder or with the model.

To explain these discrepancies, we examine the subset of bills with identical bill titles. Of the 523,841 Congressional bills and resolutions measured by the CBP, spanning 1947-2017, hand-coders read 361,747 distinct titles. Table 3 counts the number of times a bill title is repeated; approximately 54.9% of observations concern a uniquely-titled bill, with 18.2% of bill titles appearing twice in the data. The fact that the same title appears more than once presents an opportunity to directly observe the CBP's inter-coder reliability. If a bill title appears more than once, it means that title is coded by hand-coders more than once, and thus every "copy" of a bill title reflects repeated hand-coder efforts to classify it.

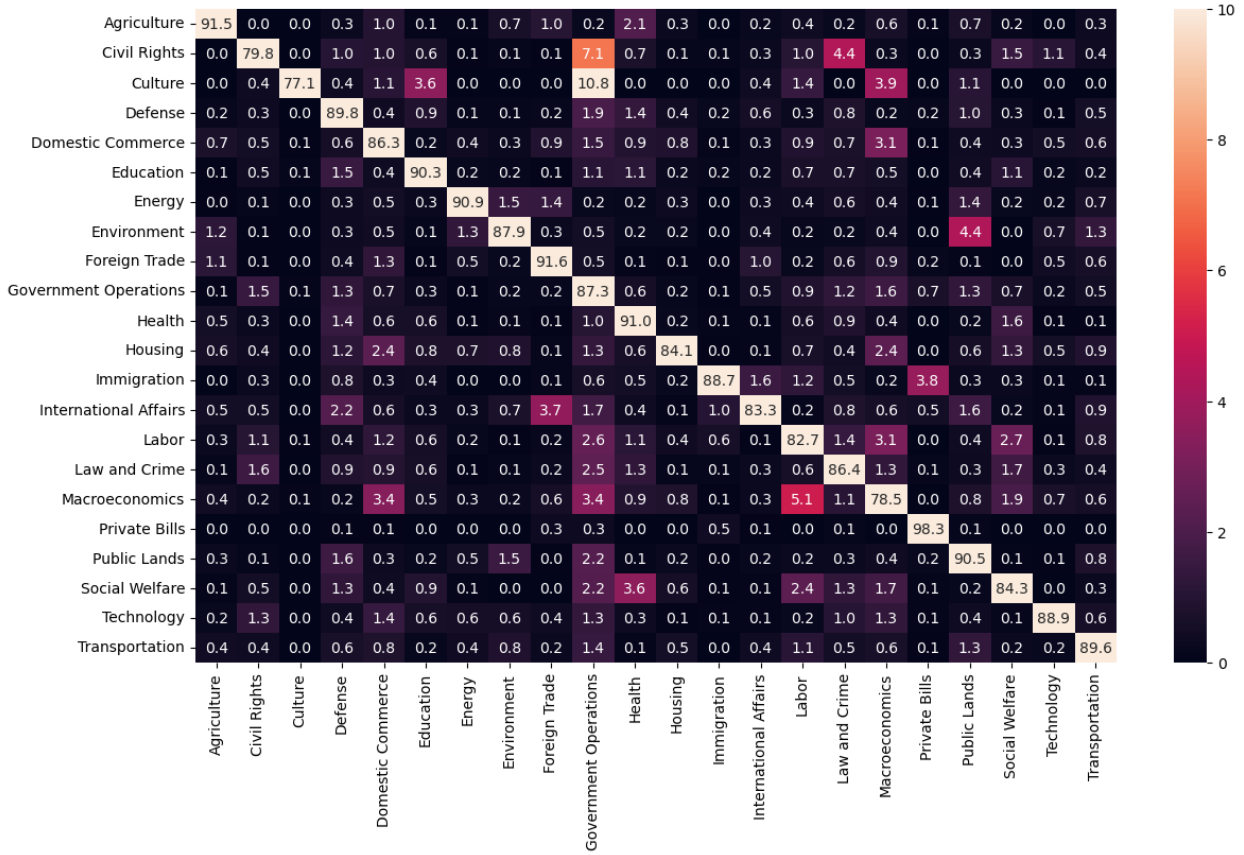
Bills with identical titles show that code inconsistencies reflect legitimate policy disagreements between hand-coders rather error, as code inconsistencies do not appear to be randomly distributed across topics. Figure 2 is a heat map of the frequency with which a duplicated bill title receives conflicting topic assignments, and shows, for example, that row *Macroeconomics*, column *Labor* = 5.1%, policy areas with considerable overlap. This exercise even reveals potential faults of the codebook, as "Macroeconomics" and "Government Operations" are often conflated with other topics, suggesting they serve as "pooling" topics for other subject areas.

Table 3: Title Repetition in Congressional Bills, 1947-2017 ($n = 523,841$)

| Title Frequency | # of Obs |
|---|---|
| 1 | 287,459 (54.9%) |
| 2 | 47,785 (18.2%) |
| 3 | 13,050 (7.5%) |
| 4 | 5,277 (4.0%) |
| 5 | 2,553 (2.4%) |
| 6 | 1,469 (1.7%) |
| 7+ | 4,154 (11.3%) |

Notes: In parentheses is the percentage of observations that involve a bill whose title occurs with a frequency denoted by the row. Only *exact* title repetitions are counted, e.g. "To extend the Renegotiation Act of 1951" and "*An Act* to extend the Renegotiation Act of 1951" are treated as unique titles. In total, there are 361,747 unique titles spread across 523,841 bills and resolutions from 1947-2017.

Figure 2: Hand-Coder Topic Assignment Consistency for Distinct, Repeated Bill Titles




Notes: For code computing hand-coder inconsistency, see Appendix **??**. There are 218,828 bills in the CBP dataset with non-unique or "duplicate" titles (titles which appear more than once), comprising 69,153 distinct titles altogether. A cell denotes the percentage of instances of duplicated bill titles which are coded as *topic* = *row* that are also observed as *topic* = *column*. For example, of the bills whose titles are observed more than once and coded as "Macroeconomics" at least once, 6.5% of them are also coded as "Labor."

## 2.3 Pennsylvania: The keystone state between national and state legislative contexts

We considered two approaches to classifying Pennsylvania legislation to address a number of potential issues. Our first concern is that Table 4 shows codebooks are not entirely analogous. Second, these are different legislatures all together, the types of actions that are contained in resolutions, bills, or amendments can all be different. For example, state legislatures often set policy with direct democracy instruments, such as referenda. Also, state legislatures consider many more sincere opportunities to amend state constitutions. This is all to say, the model may have difficulty traveling between contexts. So initially the policy content of Pennsylvania bills was predicted using a model trained only on CBP data. This model tended to under-perform, and is omitted from this article. The rest of this section describes our methods based on a jointly-trained the model on both the CBP and PAPDP.

Table 4: CAP, CBP, and PAPDP Codebook Crosswalk

| Major Topic No. | CAP | CBP | PAPDP |
|---|---|---|---|
| (0100) | Macroeconomics | * | Fiscal and Economic Issues |
| (0200) | Civil Rights | * | Civil Rights and Liberties |
| (0300) | Health | * | Health |
| (0400) | Agriculture | * | Agriculture |
| (0500) | Labor | * | Labor, Employment, and Immigration |
| (0600) | Education | * | Education |
| (0700) | Environment | * | Environment |
| (0800) | Energy | * | Energy |
| (0900) | Immigration | * | Immigration |
| (1000) | Transportation | * | Transportation |
| (1200) | Law and Crime | * | Law, Crime, and Family |
| (1300) | Social Welfare | * | Social Welfare |
| (1400) | Housing | * | Community Development, Housing Issues |
| (1500) | Domestic Commerce | * | Banking, Finance, Domestic Commerce |
| (1600) | Defense | * | Defense |
| (1700) | Technology | * | Space, Science, Technology, Communications |
| (1800) | Foreign Trade | * | Foreign Trade |
| (1900) | International Affairs | * | International Affairs and Foreign Aid |
| (2000) | Government Operations | * | State Government Operations |
| (2100) | Public Lands | * | Public Lands and Water Management |
| (2300) | Culture | * | |
| (2400) | | | Local Government and Governance |
| (9999) | | Private Bills | |

Notes: * ⇒ as in the Comparative Agendas Project (CAP). This table does not depict a perfect crosswalk between major topics, as several subtopics within-major topic for the PAPDP map to different major topics in the CAP. A full description of these differences is available here: https://liberalarts.temple.edu/sites/liberalarts/files/CAP-PPPDP%2BCrosswalk%2BCodes.pdf. The "Immigration" topic in the PAPDP codebook is not used, but "retained for theoretical and conceptual reasons."

Given the semantic and conceptual distance between the Congressional and Pennsylvania corpora, it is an open question whether the model could be trained on both hand-coded datasets without suffering a loss in performance on either. To examine this, we trained the model on approximately 85% of the CBP and PAPDP hand-coded data and assess its performance on validation data from both, making no changes to the model architecture or methodological approach, not informing the model of the legislature from which a given stream of input data originated, and not allowing the model to "warm up" by e.g. training on the CBP or PAPDP

data in isolation first before training on both. We also preserved the "Local Government and Governance" topic to account for an area of the policy agenda which may be specific to the state legislative setting, and therefore useful to retain in generating new out-of-sample state legislative data. This can pose a threat to model performance in the sense that this topic is wholly "off-limits" for Congressional bills, as is "Immigration" and "Private Bills" for Pennsylvania bills. In other words, in order to perform well on the two corpora simultaneously, the model must implicitly learn the legislature which generated the input data to understand which topics are in play. Table 5 presents the jointly-trained model's performance on the validation data.

Table 5: Jointly-Trained (Un-Clustered) Model Performance on Validation Datasets

| *Cell values are Model $F_1$ Scores, with # of Bills in parentheses* | | | | |
|---|---|---|---|---|
| | US Bills | PA Bills | PA Resolutions | PA Amendments |
| Agriculture | 88 (1647) | 82 (93) | 85 (42) | 100 (1) |
| Civil Rights | 79 (983) | 67 (119) | 75 (77) | 0 (0) |
| Culture | 74 (103) | 77 (69) | 65 (87) | 100 (32) |
| Defense | 86 (3254) | 85 (121) | 87 (77) | 80 (3) |
| Domestic Commerce | 82 (2482) | 90 (655) | 82 (119) | 92 (13) |
| Education | 86 (1420) | 90 (412) | 92 (122) | 98 (87) |
| Energy | 86 (1369) | 91 (148) | 86 (43) | 93 (8) |
| Environment | 84 (1360) | 90 (368) | 85 (80) | 100 (4) |
| Foreign Trade | 91 (2187) | 94 (8) | 75 (12) | 0 (0) |
| Government Operations | 85 (5096) | 87 (620) | 79 (313) | 87 (16) |
| Health | 89 (2952) | 94 (524) | 93 (312) | 97 (45) |
| Housing | 80 (785) | 84 (251) | 76 (41) | 57 (5) |
| Immigration | 88 (543) | 0 (0) | 0 (0) | 0 (0) |
| International Affairs | 77 (869) | 90 (11) | 81 (37) | 0 (0) |
| Labor | 82 (1547) | 90 (210) | 82 (26) | 89 (5) |
| Law and Crime | 82 (2158) | 92 (1155) | 87 (201) | 99 (51) |
| Local Government and Governance | 0 (0) | 89 (436) | 74 (46) | 57 (3) |
| Macroeconomics | 76 (1909) | 86 (284) | 74 (62) | 88 (11) |
| Private Bills | 99 (7910) | 0 (0) | 0 (0) | 0 (0) |
| Public Lands | 88 (3902) | 83 (141) | 61 (22) | 86 (3) |
| Social Welfare | 84 (1693) | 88 (265) | 76 (73) | 94 (26) |
| Technology | 87 (764) | 80 (57) | 63 (17) | 0 (1) |
| Transportation | 85 (2106) | 88 (627) | 75 (61) | 100 (10) |
| Micro Average | 87 (47039) | 89 (6574) | 82 (1870) | 95 (324) |

Notes: "Micro Average" computed across all topics denotes that all true positives and false positives are counted globally, ignoring which topic they came from. This aggregation strategy means each topic contributes to the global average proportional to the number of bills in the support.

In terms of overall performance, the model appears to be almost entirely unaffected by the concatenation of the two legislative domains, retaining almost the exact same overall performance on both corpora as it had when trained on each one in isolation. Unlike the CBP, the PAPDP also code resolutions and amendments into policy areas. Model performance tends to be worse for resolutions than for bills, reflecting the fact that resolutions are perhaps one extra degree semantically out-of-sample relative to bills. The model also appears to excel at coding Pennsylvania amendments, though there are very few of them.

Again, a close inspection reveals that the model's misses are acceptable. We conducted a clustering procedure to map bills by semantic similarity, to be shown in Figure 6a, and it allows us to compare the "ground truth" subtopic hand-codes against the model's major topic predictions. Below are several examples of its most-prevalent errors:

- PA's "Labor - Migrant and Seasonal" for our model's "Immigration" (35% of the time)
- PA's "Macroeconomics - Unemployment Rate" for our model's "Labor" (35%)
- PA's "Defense - Alliances" for our model's "Foreign Trade" (23%), or "International Affairs" (19%)

- PA's "Domestic Commerce - General" for our model's "Macroeconomics" (21%)
- PA's "International Affairs - Western Europe" for our model's "Defense" (20%)
- PA's "Health - Drug and Alcohol Abuse" for our model's "Law and Crime" (18%)
- PA's "Environment - Land and Water" for our model's "Public Lands" (18%)

Some of these errors are to be expected (the PAPDP does not include a "Immigration" code), and others are often found within the Congressional data as well: "Domestic Commerce" vs. "Macroeconomics." Altogether, this proves to be a sturdy bridge between these disparate legislative contexts.

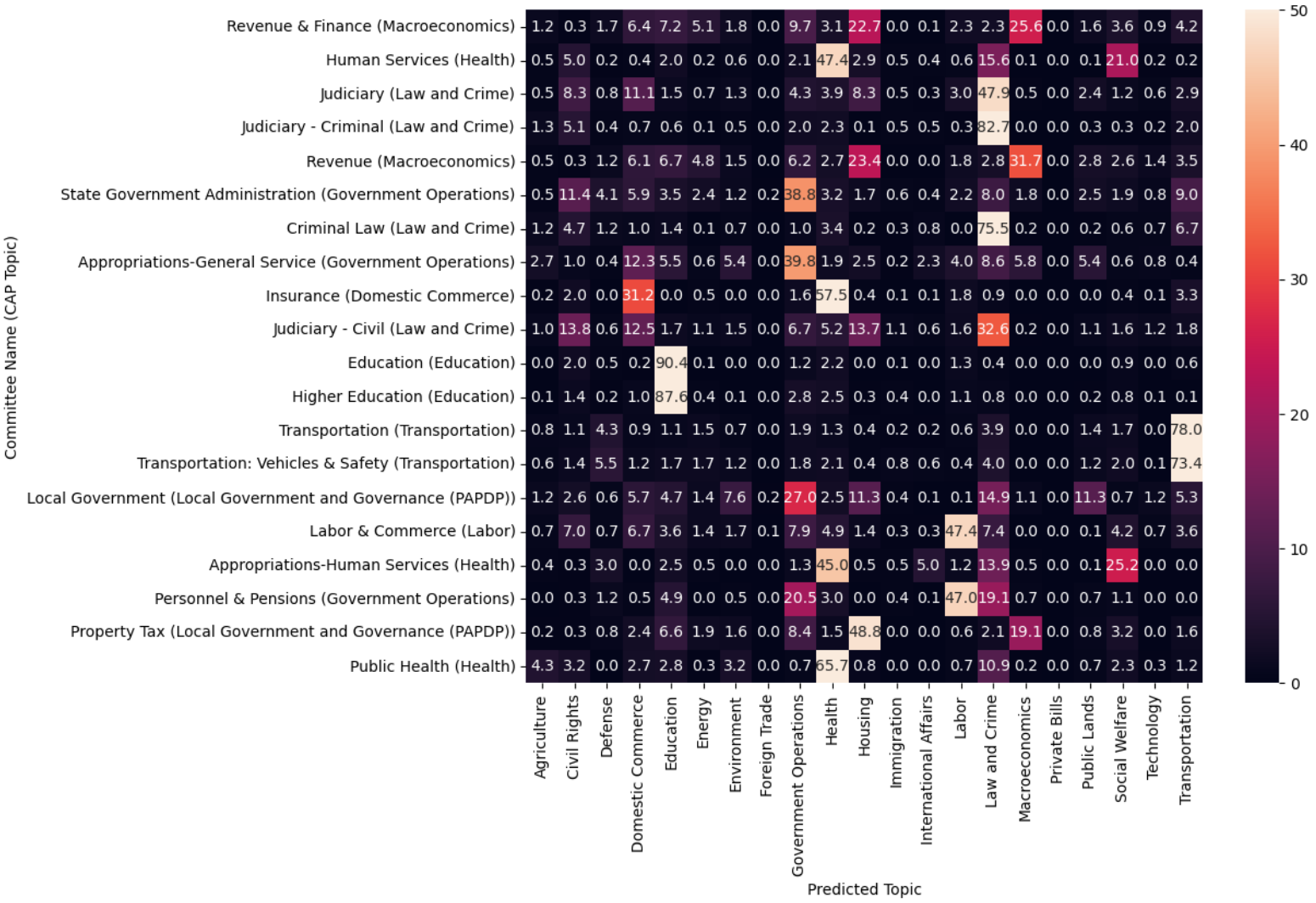
## 2.4 Illinois: Crossing state lines

Illinois presents an opportunity to verify using Pennsylvania as a keystone. It is similar to Pennsylvania in many ways. Like the Pennsylvania Constitution, the Illinois Constitution imposes that bills adhere to a "single subject" and have a "clear title." Also, it has a deep and rich legislative record. The Illinois General Assembly is a prolific producer of legislation, introducing the second-most bills and resolutions per session behind New York, and just ahead of Congress.

Illinois also provides an external measuring stick of the policy content of bills, which have otherwise not been sorted by human-coders. The General Assembly requires that bills be heard in a committee (typically, the "Second Reading") before their passage, meaning that all successful legislation will be observed as having at least one committee assignment. The Illinois House's "Rules" and the Senate's "Executive" Committees are the ruling committees over their respective chambers, but bills that will eventually be absorbed by either the Rules or Executive Committee are at least initially assigned to a policy area-dedicated committee.

After running the model that was jointly trained on Congress and Pennsylvania, we compare the predicted topics to bill that were referred to each committee throughout 2009-2023. Figure 3 presents the most commonly referred-to committees (save for "Assignments," "Executive," and "Rules") and the share of bills predicted as pertaining to each policy area. We had low expectations for a clean mapping between committees and CAP policy codes, as many bills are sent to committees for arcane matters of jurisdiction or path-dependent reasons local to one chamber. For example, in Congress, the "Ways and Means" committee that usually handles the tax code, had control of major portions of the 2010 Affordable Care Act, a landmark health care bill for both of the reasons mentioned above. With those caveats in mind there is a sensible focus for each committee with respect to the CAP policy codes. For example: "Revenue & Finance" (which we coded as "Macroeconomics") concerns mostly "Macroeconomics" and "Housing"; "Human Services" concerns "Health" and "Social Welfare"; "Judiciary" concerns "Law and Crime"; "Insurance" is "Domestic Commerce" and "Health"; the "Labor & Commerce" committee concerns "Labor"; and "Personnel & Pensions" concerns mostly "Labor" and "Government Operations."

Figure 3: Predicted Topic vs. Committee Assignment for Illinois Bills (2009-2023)




Notes: For the full depiction of my hand-coded topic assignments for Illinois committees, see Appendix **??**.

Yet again, the model's disagreements show that it is missing in sensible ways. For example, the model is split between "Agriculture" and "Transportation," for a bill amending the "Illinois Vehicle Code" declaring that it "shall not be unlawful for any person to drive or operate [vehicles for farming] to and from the home, farm, farm buildings, and any adjacent or nearby farm land." This is a genuinely difficult question for a machine learning model, or a human-coder to parse. Other examples draw similar confusion:

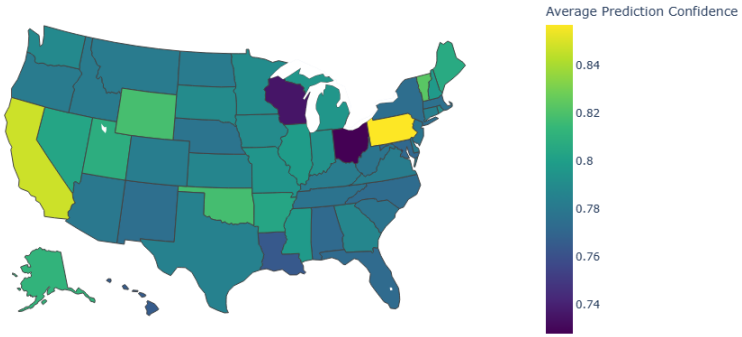
- "Defense" and "Education," for a bill amending the "School Code" concerning the "Reserve Officer's Training Corps scholarships."
- "Health" and "Social Welfare," for a bill amending the "Illinois Public Aid Code" concerning the "amount and nature of medical assistance."
- "Housing" and "Macroeconomics," for a (placeholder) appropriations bill appropriating "$2 from the General Revenue Fund to the Property Tax Appeal Board."

## 2.5 Expanding to all 50 states

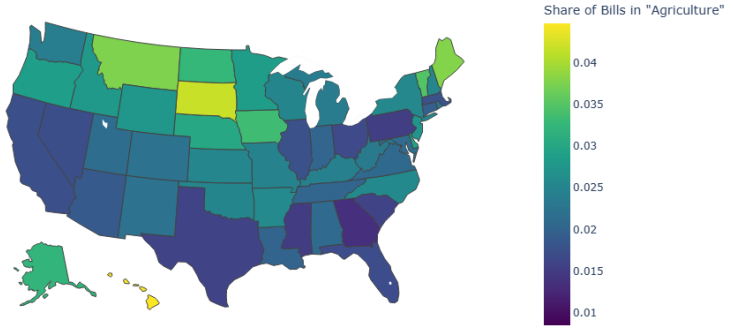
To expand to the remaining 48 states, we use the joint-trained CBP and PAPDP to bring the 50 states and Congress up-to-date through February of 2023. Figure 4 presents heatmaps depicting the model's average confidence (panel 4a) as well as the share of bills predicted as pertaining to "Agriculture" (panel 4b), "Energy" (4c), and "Environment" (4d) - three CAP topics over which we may hold well-defined priors regarding state-specific issue attention. Panel 4a indicates that the model is most confident for its predictions regarding Pennsylvania, possibly due to its Pennsylvania training data, and least-confident for Wisconsin and Ohio. This variation in model confidence across states may be indicative of variation in model performance, but it may also suggest variation in the degree to which bills are multi-faceted. Panel 4b indicates a higher share of "Agriculture" bills in e.g. South Dakota, Montana, Maine, and Nebraska, relative to e.g. Georgia, Texas, Florida, and Pennsylvania. The state with the largest share of attention to "Energy" is Alaska (in terms of number of bills introduced), While California and Maine have the largest share of "Environment" legislation.

Figure 4: Heatmaps of Model Predictions on all 50 States

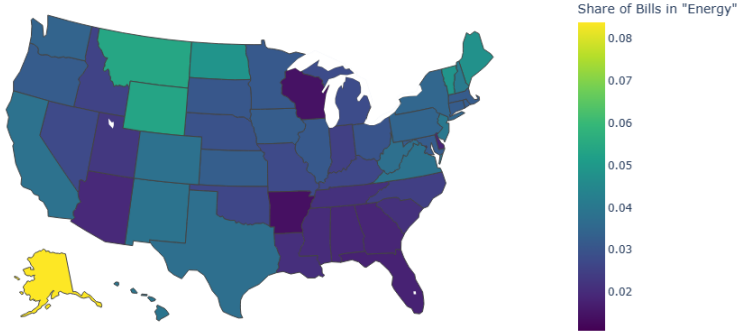
(a) Average Confidence in Predictions



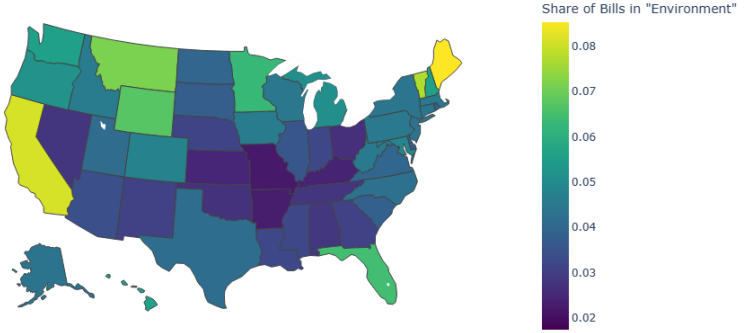

(b) Share of "Agriculture" Bills




(c) Share of "Energy" Bills




(d) Share of "Environment" Bills

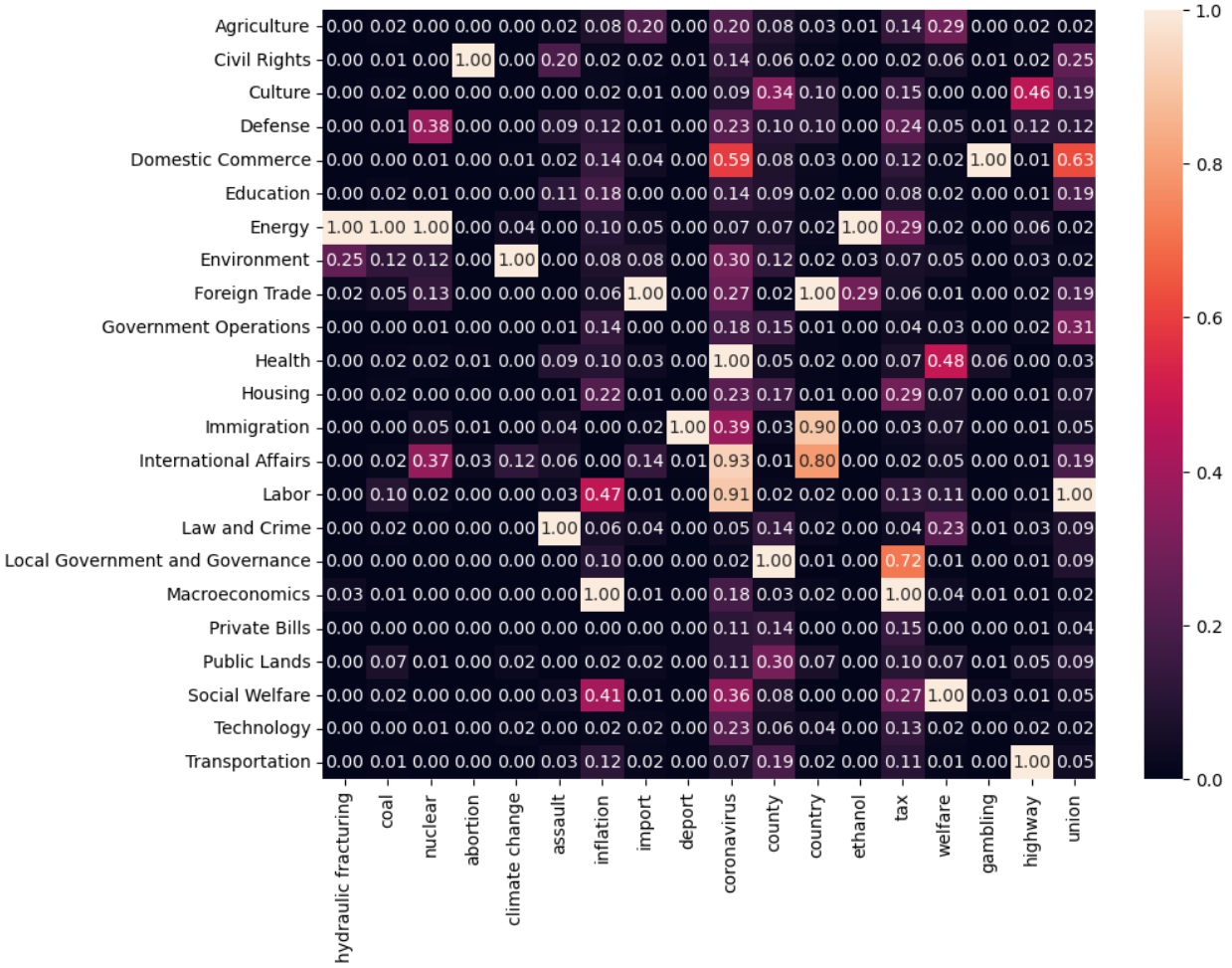



Notes: Subfigures concern bills only. Data are provided via *LegiScan*, 2009-2023.

To investigate the model's output more precisely, in Figure 5, we construct a set of search terms (case-insensitive) to return bills from the corpora of all 50 states and Congress. We make note of the predicted topic for each bill returned by a given search term, computing the share of bills per-topic which include the search term. We then normalize the computed shares per-topic over each search term to identify the topic to which the search term most frequently attends. To highlight a few examples, "Abortion" and "climate change" almost always denote "Civil Rights" and "Environment," respectively. Other notable relationships are:

- "**Coal**," "**nuclear**," and "**ethanol**": More frequently concern "Energy" than other topics, but "coal" and "nuclear" are also occasionally found in "Environment" bills. "Coal" is also occasionally found in "Labor" bills, while "nuclear" is frequently fond in "Defense" and "International Affairs."
- "**Coronavirus**": Found in many topics, but most frequently "Health," with "International Affairs," "Labor," and "Domestic Commerce" close behind.
- "**County**": Indicative of "Local Government and Governance," whereas "country" points to "Foreign trade," "Immigration," and "International Affairs."
- "**hydraulic fracturing**": is most commonly found in bills predicted to pertain to "Energy," and is four times less-frequently found in "Envirnment" bills, almost never found in "Macroeconomics," and never found elsewhere.

Figure 5: Search Term × Topic Frequencies (Normalized), All 50 States




Notes: Figure is column-normalized by dividing the number of bills containing the search term by-topic by the most-frequently returned topic, i.e. value of 1.0 indicates the predicted topic in which the search term most frequently occurs, and a value of 0.5 denotes that the topic= *row* includes the search term= *column* half as frequently as the most frequent one.

# 3 Technical Validation

This section contains a number of internal validations, which show how the model is consistent with its nearest training data, to show it is reliable. There are also external validations, so it can be compared with data that is far afield of its training data, in order to show the generalizability of the estimates. Overall, the model agrees very closely with the hand-coded efforts. We closely examine the rare instances where the model and the hand-coders disagree, and this exercise demonstrates that these deviations are often the result of CAP coders needing to select a "leading" topic, as these bills often straddle multiple potential policy areas.

## 3.1 In-domain: Comparing Congress over time

Over the years, CBP data has been were coded by several teams. The $80^{th}$-$92^{nd}$ Congresses were coded by a team at the University of Colorado led by Scott Adler, while the $93^{rd}$-$114^{th}$ were coded by a team at the University of Washington led by John Wilkerson. The $105^{th}$

Congress is the last purely hand-coded session (Hillard, Purpura, and Wilkerson 2008), with the $106^{th}$ and all Congresses thereafter employing hand-coders assisted by the "ensemble" model as developed in Collingwood and Wilkerson 2012 to code the vast majority of bills.[6] Therefore, we can partition Congresses intro three segments aftert the $92^{nd}$ and $105^{th}$ Congresses. This does require some tweaks to the evolving codebook, so we collapse "Culture" to be a subtopic of "Education," as only five bills from 1947-1972 were coded as "Culture."

The starkest increases and decreases in attention over time (in terms of number of bills introduced) by topic are to "Health" and "Private Bills," respectively. "Private Bills" are incredibly easy to identify, and were a commanding share of the dataset in the first section, but are now the second rarest topic, behind "Culture." This accords with reality. According to a manual published by the Congressional Research Service, "from 1817 through 1971, most Congresses enacted hundreds of private laws, but since then, the number has declined significantly as Congress has expanded administrative discretion to deal with many of the situations that tended to give rise to private bills."[7] Because "Private Bills" do not functionally operate or cover the same scope as "Public" bills, for clarity, I do not include their count in calculating the percentage share of the agenda covered by each topic in Table 6. The overall volume of legislation is also lower in recent memory than in more distant congresses, meaning the model has exposure to more examples that draw on older language in legislation. Another crucial difference between the three sections is the hand-coder consistency rates. The second section is leagues apart from the others in terms of hand-coder consistency rates, with an average inconsistency of 7.0% and a standard deviation in that figure across topics of 3.0%. The third section has, by far, the highest hand-coder inconsistency; even "Private Bills" are inconsistently coded 11% of the time.

Table 6: Topic Distribution for Congressional Bills by Section, 1947-2017 ($n = 466,975$)

| | First Section $80^{th}$-$92^{nd}$ | | Second Section $93^{rd}$-$105^{th}$ | | Third Section $106^{th}$-$114^{th}$ | |
|---|---|---|---|---|---|---|
| Topic | # Bills (% Share) | Inconsistency | # Bills (% Share) | Inconsistency | # Bills (% Share) | Inconsistency |
| Agriculture | 8167 (5.1%) | 7% | 5919 (4.2%) | 6% | 1990 (2.3%) | 21% |
| Civil Rights | 3436 (2.1%) | 19% | 3265 (2.3%) | 13% | 1852 (2.1%) | 31% |
| Culture | 5 (0.0%) | 50% | 328 (0.2%) | 15% | 61 (0.1%) | 48% |
| Defense | 18412 (11.5%) | 9% | 9062 (6.5%) | 6% | 6002 (6.9%) | 23% |
| Domestic Commerce | 7769 (4.9%) | 13% | 9785 (7.0%) | 7% | 6352 (7.3%) | 23% |
| Education | 6380 (4.0%) | 9% | 4618 (3.3%) | 6% | 4580 (5.3%) | 14% |
| Energy | 2236 (1.4%) | 10% | 6646 (4.7%) | 4% | 3945 (4.5%) | 17% |
| Environment | 4221 (2.6%) | 12% | 6023 (4.3%) | 5% | 3598 (4.1%) | 20% |
| Foreign Trade | 5169 (3.2%) | 9% | 6807 (4.9%) | 6% | 7971 (9.2%) | 12% |
| Government Operations | 26587 (16.6%) | 12% | 16323 (11.7%) | 7% | 8154 (9.4%) | 26% |
| Health | 5923 (3.7%) | 12% | 11461 (8.2%) | 3% | 10668 (12.3%) | 12% |
| Housing | 3907 (2.4%) | 16% | 3126 (2.2%) | 8% | 1363 (1.6%) | 31% |
| Immigration | 2478 (1.5%) | 10% | 1739 (1.2%) | 5% | 1743 (2.0%) | 18% |
| International Affairs | 2862 (1.8%) | 18% | 3504 (2.5%) | 10% | 2592 (3.0%) | 22% |
| Labor | 6980 (4.4%) | 17% | 5898 (4.2%) | 10% | 2631 (3.0%) | 30% |
| Law and Crime | 7762 (4.9%) | 12% | 8728 (6.2%) | 7% | 5139 (5.9%) | 24% |
| Macroeconomics | 7271 (4.5%) | 24% | 6398 (4.6%) | 8% | 3676 (4.2%) | 32% |
| Private Bills | 69389 | 1% | 9556 | 1% | 1047 | 11% |
| Public Lands | 19413 (12.1%) | 9% | 12376 (8.8%) | 6% | 7620 (8.8%) | 15% |
| Social Welfare | 7817 (4.9%) | 17% | 7858 (5.6%) | 8% | 2273 (2.6%) | 31% |
| Technology | 2342 (1.5%) | 7% | 2646 (1.9%) | 4% | 1758 (2.0%) | 26% |
| Transportation | 10791 (6.7%) | 12% | 7538 (5.4%) | 4% | 3039 (3.5%) | 19% |
| Total / Avg. (Std. Dev.) | 229,317 | 14.5% (9.2%) | 149,604 | 7.0% (3.0%) | 88,054 | 23.6% (8.5%) |

Notes: Congressional sessions were partitioned into contiguous "sections" by noting the various contexts in which the ground truth data were generated. The first section was coded by a team led by Scott Adler, while the second and third were coded by a team led by John Wilkerson. The third section incorporates the machine learning model developed in Collingwood and Wilkerson 2012 to generate labels for a large share of bills.

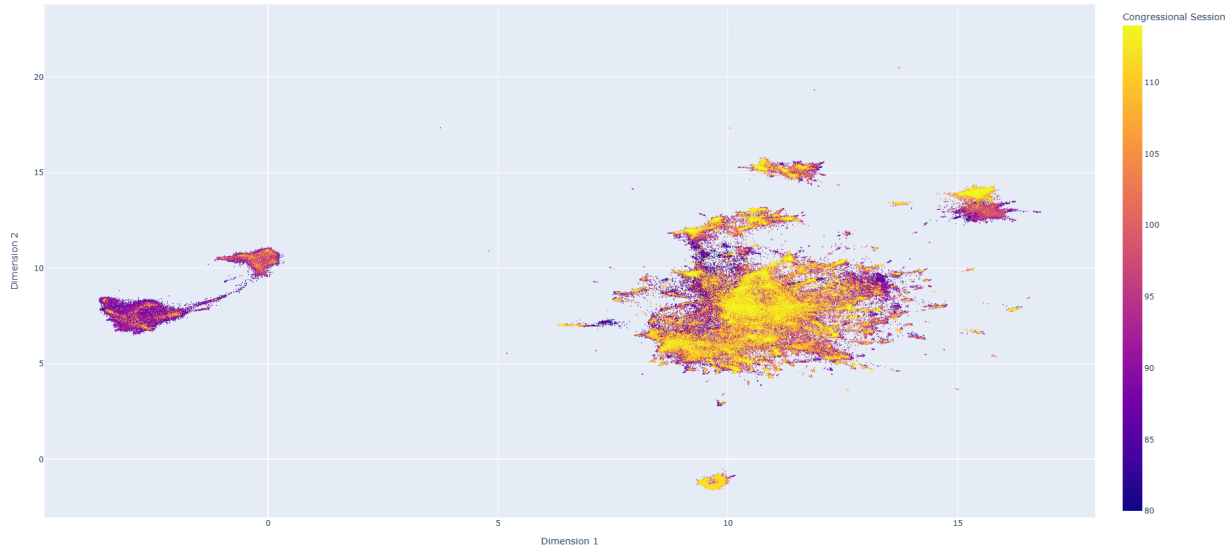
6. These temporal "sections" in the data are based on our understanding of the CBP's data creation process and correspondence with the CBP. Errors and misunderstandings are my own.

7. See: https://crsreports.congress.gov/product/pdf/R/R45287/3

To simulate the use-case of extending the trained model to unseen corpora, and to get a sense of the "semantic drift" present in legislative text, we project high-dimensional document embeddings derived from a pre-trained transformer model onto a 2-dimensional space using Universal Manifold Approximation and Projection (UMAP, see McInnes, Healy, and Melville 2018). In Figure 6a, for example, we recover the fact that the "Private Bills" policy area is largely an artifact of older Congresses, and identify that even within a tight semantic "cluster" such as bills pertaining to the "Internal Revenue Code" (Figure 6b), language shifts over time, emphasizing the need to construct meaningful tests of model generalizability. To achieve this, we combine UMAP embeddings with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN, Campello et al. 2013) to construct unsupervised semantic clusters of bills, restricting the model to training on a select subset of clusters, and validating its performance on the "out-of-sample" clusters.

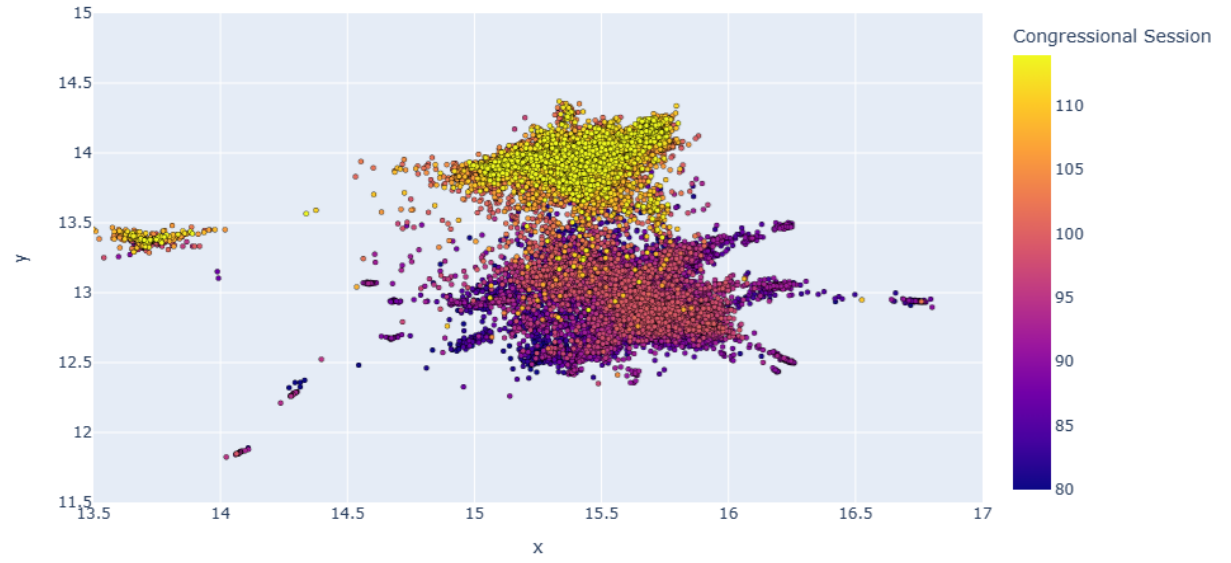
Figure 6: "Topic Drift" in Congressional Bills (80th-114th Congresses)

(a) All Congressional Bills




Notes: The "Private Bills" super-cluster to the "west" is mostly an artifact of earlier Congressional sessions, with private bills becoming far less common as time goes on. The "northeast" collection of bills mostly contains references to the "Internal Revenue Code," which comprise the scope of subfigure 6b.

(b) The "Internal Revenue Code" Super-Cluster




Notes: The "Internal Revenue Code" super-cluster refers to the "northeast" cluster depicted in Subfigure 6a. Semantically, it appears to be "drifting" north over time, indicating an evolution in the language used in bills mentioning the "Internal Revenue Code."

### 3.2 Comparison to Bag of Words models

Our transformer based estimates compare favorably to previous efforts. On a training set where Hillard, Purpura, and Wilkerson 2008 is 89% accurate, our model is 91.2% accurate.[8] If the sample is limited to the bills that the Hillard model is most confident on, their accuracy rises to 94%, and if we limit our model to predictions it is 85% confident in, our accuracy is 96.3%. We also calibrate our sample to compare to Collingwood and Wilkerson 2012 which had an F-1 score of 79.3, and our F-1 score is 84.1.[9]

### 3.3 Out of domain: Synthetic bills

This section tries to demonstrate how the model is "learning" to predict the policy content in legislation based on context clues. It tries to show how a deep neural network, trained to adjust its millions of parameters, focused on its own abstract optimization problem, can "explain" itself. One straightforward approach to exploring the model's strengths and weaknesses is to allow the researcher to generate "synthetic" examples - fake bill titles "drafted" by the researcher themselves. With the researcher having strong priors regarding the policy content of these synthetic examples, the model can be evaluated with respect to how accurately in uncovers it. Table 7 presents synthetic bill titles (of our own making) and reports the model's top three confidence scores.

8. Details in Appendix 5.1.1.
9. Details in Appendix 5.1.2.

Table 7: Synthetic Bill Examples

| | | | |
|---|---|---|---|
| *Panel A: In-Sample Phrases* | | | |
| discrimination 2093 | Civil Rights (94%) | Macroeconomics (2%) | Transportation (2%) |
| doctor 517 | Health (92%) | Education (3%) | Agriculture (2%) |
| doctoral 26 | Education (84%) | Health (9%) | Housing (2%) |
| food stamps 193 | Social Welfare (94%) | Domestic Commerce (2%) | Agriculture (1%) |
| internal revenue code 34038 | Macroeconomics (63%) | Government Operations (32%) | Transportation (1%) |
| loan forgiveness 130 | Education (43%) | Housing (29%) | Domestic Commerce (16%) |
| neighborhood 293 | Housing (90%) | Government Operations (7%) | Private Bills (2%) |
| smartphones 2 | Technology (86%) | Private Bills (3%) | Education (2%) |
| tariffs 119 | Foreign Trade (99%) | Private Bills (0%) | Energy (0%) |
| *Panel B: "Out-of-Sample" Phrases (Do Not Occur in CBP Data* | | | |
| coronavirus 0 | Health (90%) | Education (2%) | Agriculture (2%) |
| social distancing 0 | Health (65%) | Education (17%) | Environment (6%) |
| cryptocurrency 0 | Technology (36%) | Private Bills (15%) | Law and Crime (13%) |
| bitcoin 0 | Macroeconomics (48%) | Technology (12%) | Private Bills (12%) |
| *Panel C: Abbreviations* | | | |
| 401(k) 45 | Labor (50%) | Social Welfare (39%) | Macroeconomics (5%) |
| AARP 0 | Social Welfare (96%) | Health (1%) | Agriculture (1%) |
| HSA 17 | Social Welfare (83%) | Health (12%) | Labor (1%) |
| IRS 7022 | Macroeconomics (79%) | Government Operations (8%) | Domestic Commerce (3%) |
| UIUC 0 | Education (97%) | Housing (1%) | Private Bills (0%) |
| UVM 0 | Education (91%) | Housing (4%) | Transportation (1%) |
| UVA 2 | Education (86%) | Health (5%) | Defense (3%) |
| VA 110 | Defense (96%) | Foreign Trade (1%) | Public Lands (1%) |
| *Panel D: Proper Nouns* | | | |
| Congressional Bills Project 0 | Government Operations (97%) | Private Bills (2%) | Transportation (0%) |
| Albert Einstein 4 | Private Bills (48%) | Energy (19%) | Education (11%) |
| Ethan Dee 0 | Private Bills (90%) | Education (7%) | Environment (1%) |
| John Maynard Keynes 0 | Macroeconomics (57%) | Domestic Commerce (39%) | Energy (1%) |
| George Washington 145 | Government Operations (81%) | Private Bills (11%) | Defense (5%) |
| China 351 | Foreign Trade (49%) | International Affairs (48%) | Macroeconomics (1%) |
| India 89 | International Affairs (57%) | Foreign Trade (35%) | Macroeconomics (3%) |
| Indian 4019 | Public Lands (70%) | International Affairs (8%) | Government Operations (6%) |
| North Korea 70 | International Affairs (83%) | Defense (12%) | Foreign Trade (2%) |
| Russia 54 | International Affairs (92%) | Defense (3%) | Macroeconomics (2%) |
| *Panel E: "Tax" Homonyms* | | | |
| carbon tax 3 | Environment (77%) | Energy (11%) | Macroeconomics (7%) |
| estate tax 754 | Macroeconomics (54%) | Housing (40%) | Energy (2%) |
| gasoline tax 52 | Energy (93%) | Transportation (2%) | Environment (1%) |
| liquor tax 0 | Domestic Commerce (62%) | Macroeconomics (28%) | Health (4%) |
| sales tax 147 | Macroeconomics (95%) | Domestic Commerce (3%) | Energy (1%) |
| tax credit 2732 | Macroeconomics (93%) | Domestic Commerce (3%) | Social Welfare (2%) |
| *Panel E: "Environment" Homonyms* | | | |
| computer environment 0 | Technology (74%) | Education (17%) | Macroeconomics (2%) |
| learning environment 3 | Education (98%) | Housing (0%) | Technology (0%) |
| natural environment 15 | Environment (87%) | Public Lands (9%) | Labor (1%) |
| political environment 0 | Government Operations (95%) | Civil Rights (2%) | Macroeconomics (1%) |
| workplace environment 0 | Labor (81%) | Macroeconomics (7%) | Transportation (5%) |
| about the learning environment 0 | Education (94%) | Civil Rights (2%) | Housing (1%) |
| learning about the environment 0 | Environment (97%) | Health (1%) | International Affairs (0%) |
| about learning about the environment 0 | Environment (94%) | Education (3%) | Housing (1%) |
| about learning the environment 0 | Education (63%) | Environment (31%) | Housing (4%) |
| environmental factors 21 | Environment (93%) | International Affairs (1%) | Health (1%) |
| environmental factors affecting outcomes 0 | Health (57%) | Environment (27%) | Education (4%) |
| an environment conducive to growth 0 | Macroeconomics (88%) | Housing (4%) | Energy (2%) |



Notes: Apart from abbreviations, all search terms are case-insensitive.

Panel ($A$) of Table 7 presents several phrases that appear in Congressional bill titles, and thus have been seen in some way by hand-coders in various contexts. For example, mentions of "China" appear in 517 bill titles; when the model is made to form a prediction for the word "China," barring any other context, its leading prediction is the "International Affairs" topic, with a confidence score of 90%. The model's usage of word-piece, rather than word embeddings, aids in assigning a sensibly-confident "Education" prediction to the word "doctor*al*," whereas "doctor" confidently concerns "Health." The model's pre-training task on English Wikipedia and BooksCorpus (Zhu et al. 2015) generates a strong baseline familiarity with English; while "smartphones" only appears twice in Congressional bill titles, it is confidently mapped to the "Technology" topic. Panel ($B$) of Table 7 presents several examples that were not seen by the model during training and are not present at all in Congressional bill titles. The word "coronavirus" is constructed by RoBERTa by concatenating the word-pieces "cor" + "##on" + "##av" + "##irus"; even if never exposed to the virus' exact namesake, in general, the model is capable of inferring that [*word-pieces*] + "virus" most likely warrants a word embedding that falls within the boundaries of the CBP's "Health" topic.

Panel ($C$) of Table 7 presents several examples of abbreviations, several of which were not seen by the model during training and are not present at all in Congressional bill titles. Two abbreviations for the University of Illinois at Urbana-Champaign ("UIUC") and the University of Vermont ("UVM") are not present in the CBP data, and yet, the model predicts the "Education" topic for both of them. Again, one of the benefits of transfer learning is that the model starts with a strong baseline understanding of English, and, having "read" BooksCorpus (Zhu et al. 2015) and English Wikipedia, encountered both of these acronyms during pre-training.

Panel ($D$) of Table 7 explores the model's generalizability to proper nouns. The "Congressional Bills Project" itself perhaps unironically maps to the "Government Operations" topic. "Albert Einstein" made an appearance in four Congressional bill titles (and several more resolutions) from 1947-2017:

- To establish a national Albert Einstein Teacher Fellowship Program for outstanding secondary school science and mathematics teachers. *(102-HR-4346)*
- A bill to establish a national Albert Einstein Teacher Fellowship Program for outstanding secondary school science and mathematics teachers. *(102-S-2031)*
- To establish within the Department of Energy a national Albert Einstein Distinguished Educator Fellowship Program for outstanding elementary and secondary mathematics and science teachers. *(103-HR-4759)*
- A bill to establish within the National Laboratories of the Department of Energy a national Albert Einstein Distinguished Educator Fellowship Program. *(103-S-2104)*

with each bill hand-coded as "Education." However, if "Albert Einstein" is read in isolation, the model only partially treats it as "Education" (confidence = 11%), but more directly views it as "Energy" (19%) or "Private Bills" (48%), likely owing to the fact that it is a proper noun and "Private Bills" usually entail the phrase "for the relief of" + [*proper noun*]. The unknown person "Ethan Dee" exemplifies this phenomenon, with the model now far more confident in the "Private Bills" prediction. Interestingly, "John Maynard Keynes" receives a "Macroeconomics" (57%) and "Domestic Commerce" (39%) label while never appearing in the CBP dataset (but most likely appearing in RoBERTa's pre-training data).

Panels ($E$) and ($F$) of Table 7 present homonyms of "tax" and "environment," respectively, to demonstrate how these words, in isolation, might map strongly to particular topics, but in the presence of surrounding context, entail entirely different policy areas. For example, in Panel ($E$), a "carbon" tax maps to "Environment," a "gasoline" tax to "Energy," and a "liquor" tax to "Domestic Commerce."[10] In panel ($F$), the "computer," "learning," "natural,"

10. The exact phrase "liquor tax" never appears in a Congressional bill title, but of course, many bills discuss the "sale of liquor" and subject it to some form of "tax."

"political," and "workplace" environments map to "Technology," "Education," "Environment," "Government Operations," and "Labor," respectively. Moreover, if a bill is "about the learning environment," it maps to "Education," but if a bill entails "learning about the environment," it maps to "Environment." A bill discussing "environmental factors affecting outcomes" would map to "Health" (with a strong background note of "Environment"), whereas a bill discussing "an environment conducive to growth" is predicted as attending to "Macroeconomics."

Another by-hand approach to testing the model's generalizability is to create "perturbations" of a given synthetic bill and observe how those perturbations affect its prediction. For example, a synthetic bill titled "This is a bill about the learning environment" should return the same prediction whether or not the model is allowed to read the words "This," "is," "a," "bill," "about," or "the." In Table 8, we mask each of these words one-by-one, and report the model's predicted confidence scores. The model's confidence in its prediction is mostly unaffected by masking a single word, save for masking the word "learning," as it appears to be the model's main semantic connection to the "Education" topic.

Table 8: Synthetic Bill Example, with "Masked" Language

| Bill Title | $1^{st}$-Best Prediction | $2^{nd}$-Best Prediction | $3^{rd}$-Best Prediction |
|---|---|---|---|
| Public Law No. 346. This is a bill concerning the learning environment. | Education (95%) | Civil Rights (1%) | Public Lands (1%) |
| [MASK] Law No. 346. This is a bill concerning the learning environment. | Education (95%) | Government Operations (1%) | Civil Rights (1%) |
| Public[MASK] No. 346. This is a bill concerning the learning environment. | Education (95%) | Government Operations (1%) | Civil Rights (1%) |
| Public Law[MASK]. 346. This is a bill concerning the learning environment. | Education (94%) | Civil Rights (2%) | Government Operations (1%) |
| Public Law No[MASK] 346. This is a bill concerning the learning environment. | Education (95%) | Civil Rights (1%) | Public Lands (1%) |
| Public Law No.[MASK]. This is a bill concerning the learning environment. | Education (96%) | Civil Rights (1%) | Public Lands (1%) |
| Public Law No. 346[MASK]. This is a bill concerning the learning environment. | Education (95%) | Government Operations (1%) | Civil Rights (1%) |
| Public Law No. 346.[MASK] is a bill concerning the learning environment. | Education (94%) | Government Operations (2%) | Civil Rights (1%) |
| Public Law No. 346. This[MASK] a bill concerning the learning environment. | Education (95%) | Civil Rights (1%) | Government Operations (1%) |
| Public Law No. 346. This is[MASK] bill concerning the learning environment. | Education (94%) | Civil Rights (2%) | Government Operations (1%) |
| Public Law No. 346. This is a[MASK] concerning the learning environment. | Education (94%) | Civil Rights (2%) | Housing (1%) |
| Public Law No. 346. This is a bill[MASK] the learning environment. | Education (94%) | Civil Rights (1%) | Technology (1%) |
| Public Law No. 346. This is a bill concerning[MASK] learning environment. | Education (95%) | Civil Rights (1%) | Technology (1%) |
| Public Law No. 346. This is a bill concerning the[MASK] environment. | Environment (91%) | Agriculture (3%) | Transportation (1%) |
| Public Law No. 346. This is a bill concerning the learning[MASK]. | Education (94%) | Social Welfare (2%) | Public Lands (1%) |
| Public Law No. 346. This is a bill concerning the learning environment[MASK] | Education (95%) | Civil Rights (1%) | Public Lands (1%) |

Notes: The "[MASK]" token denotes that the word is "masked" from the model, i.e. it is unable to observe the word, and instead knows only that some word occupies the "[MASK]" position. Results were similar when simply deleting the word in the "[MASK]" position.

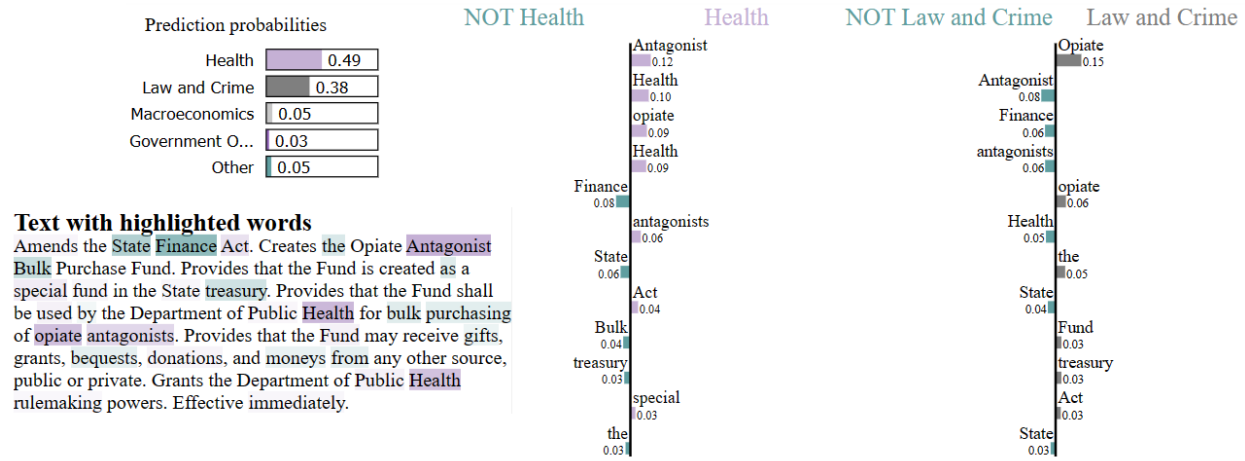
It perhaps goes without saying that a self-designed test of the model's robustness might paint

a generous picture. To formalize this approach of testing the model's robustness to perturbations in the input data, algorithms such as LIME (Ribeiro, Singh, and Guestrin 2016) can be used to generate "locally interpretable" explanations for the model's output sampling from a distribution of word-masking options. LIME first randomly chooses a number of words to mask, and then randomly chooses which specific words to mask. This process is iterated `num_samples`[11] times, generating a strong approximation for the most "important" words in the document with respect to the classification decision. In Figure 7, we highlight the model's most important words, approximated using LIME, for two bills originating in the Illinois General Assembly - a legislative domain that is temporally and conceptually outside the scope of the model's original training data. In panel 7a, the LIME algorithm returns that the word "Finance" detracts from both of the model's leading predictions, "Health" and "Law and Crime," but the word is quickly overwhelmed by the following sentences which discuss "opiate" and "health." "Opiate" contributes to both "Health" and "Law and Crime" predictions, but repeated mentions of "health" appear to drive the model toward the "Health" topic. In panel 7b, the model unambiguously predicts the "Law and Crime" topic owing to mentions of e.g. "felony," "firearm," "ammunition," and "knowingly" (an adverb common to the "Law and Crime" topic). It is also worth noting that in both of these examples, no single word dominates the model's prediction rationale, echoing the desire for a model that generalizes well and does not place too much stock into specific features.

11. A LIME hyperparameter.

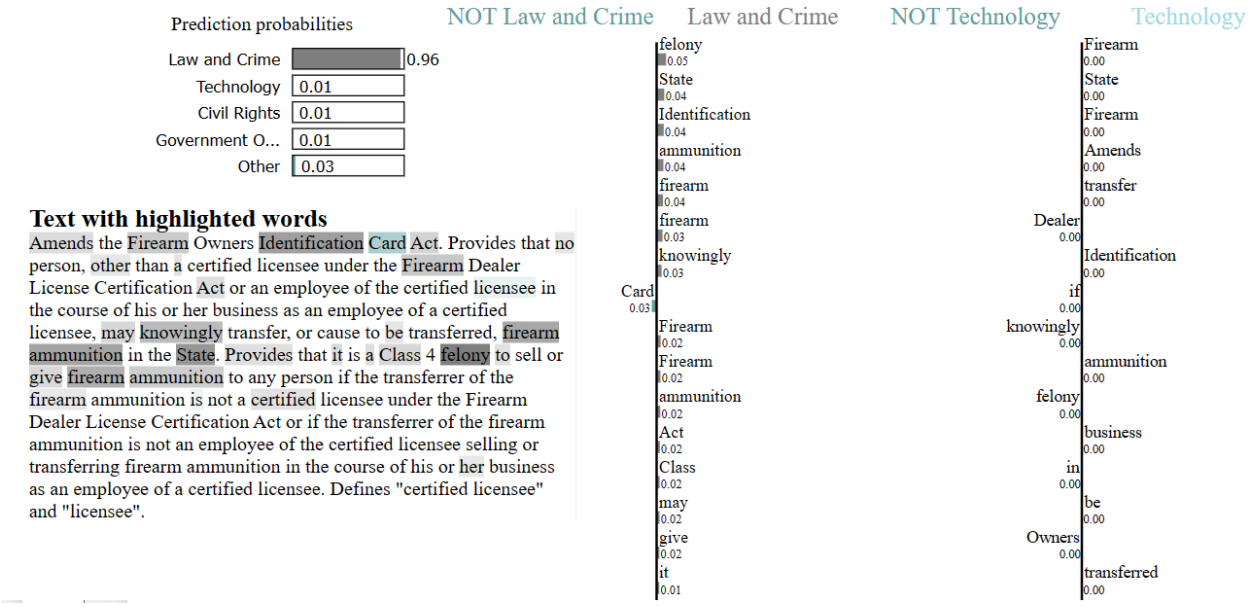
Figure 7: Explaining Model Predictions for Example Bills from Illinois

(a) Example Bill using LIME (*IL-HB-2526*, 2023)




Notes: The `num_samples` hyperparameter was set to `5000`, i.e. $5,000$ random selections over the number of words and which to be masked from the model.

(b) Example Bill using LIME (*IL-HB-1057*, 2023)




Notes: The `num_samples` hyperparameter was set to `5000`, i.e. $5,000$ random selections over the number of words and which to be masked from the model.

### 3.4 Know Thyself: Predicting the CAP Descriptions

As a final form of validation, the CAP master codebook[12] - the set of instructions for the classification system itself - can be used as another form of validation data. We take the descriptions of each subtopic to be the input data, and the major topics to which they belong to be the ground truth labels associated with those descriptions. Table 9 presents the model's performance when predicting the major topics associated with the CAP subtopic descriptions. The model correctly places 207 of the 212 subtopic descriptions into their proper major topics.

Table 9: Model Performance on the CAP Master Codebook

| Topic | Precision | Recall | $F_1$ Score | Support |
|---|---|---|---|---|
| Agriculture | 100.0 | 100.0 | 100.0 | 9 |
| Civil Rights | 100.0 | 100.0 | 100.0 | 10 |
| Defense | 94.7 | 100.0 | 97.3 | 18 |
| Domestic Commerce | 100.0 | 93.3 | 96.6 | 15 |
| Education | 100.0 | 100.0 | 100.0 | 9 |
| Energy | 100.0 | 100.0 | 100.0 | 9 |
| Environment | 100.0 | 100.0 | 100.0 | 11 |
| Foreign Trade | 100.0 | 100.0 | 100.0 | 8 |
| Government Operations | 94.1 | 88.9 | 91.4 | 18 |
| Health | 100.0 | 100.0 | 100.0 | 17 |
| Housing | 100.0 | 100.0 | 100.0 | 11 |
| Immigration | 100.0 | 100.0 | 100.0 | 1 |
| International Affairs | 92.3 | 100.0 | 96.0 | 12 |
| Labor | 100.0 | 100.0 | 100.0 | 9 |
| Law and Crime | 100.0 | 92.3 | 96.0 | 13 |
| Macroeconomics | 90.0 | 100.0 | 94.7 | 9 |
| Public Lands | 100.0 | 85.7 | 92.3 | 7 |
| Social Welfare | 100.0 | 100.0 | 100.0 | 7 |
| Technology | 100.0 | 100.0 | 100.0 | 10 |
| Transportation | 90.0 | 100.0 | 94.7 | 9 |
| Micro Average | 97.6 | 97.6 | 97.6 | 212 |
| Macro Average | 98.1 | 98.0 | 98.0 | 212 |

Notes: "Micro Average" computed across all topics denotes that all true positives and false positives are counted globally, ignoring which topic they came from. This aggregation strategy means each topic contributes to the global average proportional to the number of bills in the support. A "macro averaging" strategy first computes the per-topic value, and then takes a simple average of the each topic's value. Each topic thereby contributes equally to the global average.

Per Table 10, three of the five that the model fails to classify belong to "Government Operations," and one apiece belong to "Law and Crime" and "Public Lands." In four of these five cases, the model's confidence in its prediction is below 60%, indicating that it is not overly "convinced" of its (incorrect) prediction.

12. See: https://www.comparativeagendas.net/pages/master-codebook

Table 10: Model Errors on the CAP Master Codebook

| Subtopic Description | $1^{st}$-Best Prediction | $2^{nd}$-Best Prediction | $3^{rd}$-Best Prediction |
|---|---|---|---|
| Includes issues related to police and other general domestic security responses to terrorism, such as special police (1227, Law and Crime) | Defense (42%) | Law and Crime (25%) | Domestic Commerce (9%) |
| Includes issues related to domestic commerce research and development (1598, Domestic Commerce) | Transportation (54%) | Macroeconomics (29%) | Domestic Commerce (7%) |
| Includes issues related to tax administration, enforcement, and auditing for both individuals and corporations (2009, Government Operations) | Macroeconomics (93%) | Domestic Commerce (3%) | Government Operations (2%) |
| Includes issues related to claims against the government, compensation for the victims of terrorist attacks, compensation policies without other substantive provisions (2015, Government Operations) | International Affairs (60%) | Defense (19%) | Government Operations (14%) |
| Includes issues related to territorial and dependency issues and devolution (2105, Public Lands) | Government Operations (47%) | Law and Crime (30%) | Defense (14%) |

Notes: In parentheses in the "Subtopic Description" column are the CAP subtopic codes and the major topics to which they belong.

In the fifth case, though, for CAP topic 2009 ("Tax Administration"), the model places extremely high confidence in the "Macroeoconomics" topic, perhaps conflating it with topic 107 ("Tax Code") which reads, "Includes issues related to tax policy, the impact of taxes, and tax enforcement." Examining the hand-coder inconsistencies associated with subtopics lends credibility to this explanation. In Table 11, I report the top 25 hand-coder inconsistencies by subtopic in the CBP data from the $93^{rd}$ through the $114^{th}$ Congress. In constructing this table, we noted several codes for documents which we believed to be in error, on the grounds that they do not exist in either the CBP or CAP codebooks. We corrected these anomalies by-hand (a total of 27 bills), and additionally recoded bills with the CBP code 609 ("Arts and Humanities") as code 2300 ("Culture") to match the CAP codebook. Table 11 shows that the conflation of topic 107 with 2009 is among the most common hand-coder inconsistencies, with nearly all that rank above it being inconsistencies generating subtopic mismatches which are within, rather than across major topics. One of the two hand-coder inconsistencies conflating two major topics that rank as more inconsistent than topic 107 and 2009 has an inconsistency rate of 28% for "(Agriculture) Animal and Crop Disease" and "(Defense) R&D," and all 24 bills which cause this inconsistency are copies of the same title which reads: "A bill to prohibit the military departments from using dogs in connection with any research or other activities relating to biological or chemical warfare agents." The second one is the CBP-specific "Private Bills" topic being conflated with the reasonable second choice: "(Government Operations) Claims against the government." Topics 107 and 2009 share significant conceptual overlap, generating (justifiable) hand-coder disagreements, and as a consequence, the two topics are a natural point of confusion for the model.

Table 11: Top 25 CBP Subtopic Hand-Coder Inconsistencies ($93^{rd}$-$114^{th}$ Congresses, With Manual Adjustments)

| CBP Subtopic ... | ... Also Coded As | Inconsistency Rate | # Bills Concerned |
|---|---|---|---|
| Social Welfare - Elderly Assistance (4509) | Social Welfare - General (748) | 49% | 5257 |
| Social Welfare - Elderly Assistance (4509) | Social Welfare - Other (52) | 32% | 4561 |
| Agriculture - Animal and Crop Disease (586) | Defense - R&D (76) | 28% | 662 |
| Law and Crime - Agencies (3620) | Law and Crime - Police, Fire, and Weapons Control (391) | 28% | 4011 |
| Social Welfare - Nutrition Assistance (2090) | Social Welfare - Low-Income Assistance (98) | 27% | 2188 |
| Defense - Personnel Issues (7664) | Defense - Veteran Affairs and Other Issues (541) | 25% | 8205 |
| Civil Rights - General (887) | Civil Rights - Other (46) | 19% | 933 |
| Law and Crime - Agencies (3620) | Law and Crime - Other (91) | 19% | 3711 |
| Private Bills - Private Bills (10603) | Govt. Ops. - Claims against the government (431) | 18% | 11034 |
| Education - Elementary & Secondary (2118) | Education - Excellence (897) | 17% | 3015 |
| Macroeconomics - Tax Code (6264) | Govt. Ops. - Tax Administration (667) | 16% | 6931 |
| Education - Elementary & Secondary (2118) | Education - Underprivileged (494) | 16% | 2612 |
| Social Welfare - Elderly Assistance (4509) | Labor - Other (46) | 16% | 4555 |
| Social Welfare - Nutrition Assistance (2090) | Energy - Other (38) | 15% | 2128 |
| Govt. Ops. - Political Campaigns (2827) | Civil Rights - Voting Rights (377) | 15% | 3204 |
| Education - Elementary & Secondary (2118) | Education - Other (230) | 15% | 2348 |
| Public Lands - Public Lands (5618) | Public Lands - General (630) | 15% | 6248 |
| Technology - Space (359) | Technology - Commercial Use of Space (119) | 14% | 478 |
| Technology - Telecommunications (1102) | Technology - General (249) | 13% | 1351 |
| Defense - Personnel Issues (7664) | Defense - Foreign Operations (320) | 13% | 7984 |
| Education - Underprivileged (494) | Education - R&D (51) | 13% | 545 |
| Social Welfare - General (748) | Civil Rights - Handicap Discrimination (235) | 13% | 983 |
| Environment - General (711) | Environment - Other (103) | 13% | 814 |
| Immigration - Immigration (3482) | Labor - Migrant and Seasonal (202) | 12% | 3684 |
| Agriculture - Animal and Crop Disease (586) | Agriculture - Other (181) | 12% | 767 |

Notes: Underlined are hand-coder inconsistencies at the subtopic level which induce major topic inconsistencies. In other words, the non-underlined rows denote inconsistencies that are irrelevant for a model trained to predict major topic codes. The topic codes for 27 bills were manually corrected to align with the current CBP and CAP codebooks. For details regarding the code adjustments I made, see Appendix **??**.

### 3.5 State-level validations

Table 12 presents the topic distributions for Pennsylvania bills and resolutions, and reprints the distributions for its Congressional counterpart for comparability.

Table 12: Predicted Topic Distribution for Congress (2018-2023) and Pennsylvania (2010-2023)

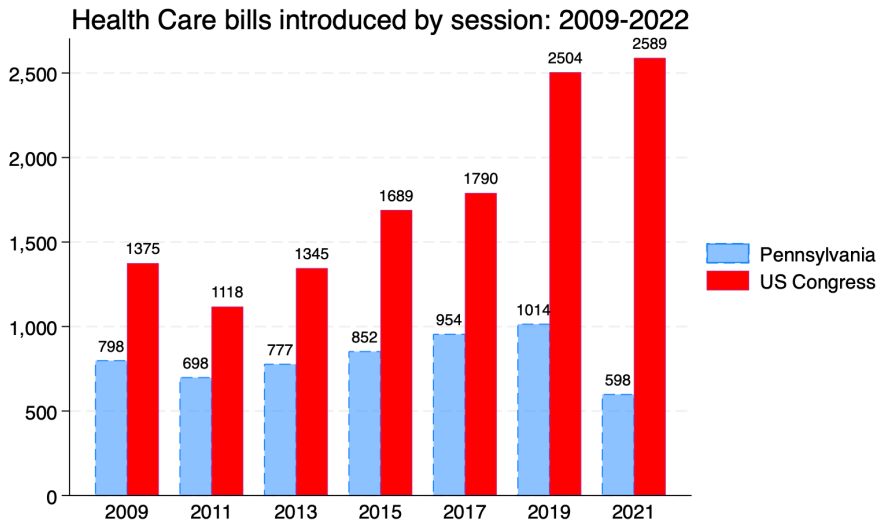
| | Congress (2018-2023) | | Pennsylvania (2010-2023) | |
|---|---|---|---|---|
| Topic | Bills | Resolutions | Bills | Resolutions |
| Agriculture | 537 (2%) | 63 (1%) | 509 (2%) | 252 (2%) |
| Civil Rights | 1612 (5%) | 501 (11%) | 1024 (3%) | 847 (8%) |
| Culture | 85 (0%) | 300 (6%) | 664 (2%) | 917 (8%) |
| Defense | 1921 (6%) | 302 (6%) | 797 (2%) | 703 (6%) |
| Domestic Commerce | 2190 (7%) | 124 (3%) | 3026 (9%) | 408 (4%) |
| Education | 1678 (5%) | 275 (6%) | 2758 (8%) | 784 (7%) |
| Energy | 1349 (4%) | 66 (1%) | 1148 (3%) | 153 (1%) |
| Environment | 1445 (5%) | 110 (2%) | 1504 (5%) | 304 (3%) |
| Foreign Trade | 474 (2%) | 28 (1%) | 24 (0%) | 37 (0%) |
| Government Operations | 2657 (9%) | 821 (17%) | 3371 (10%) | 1057 (9%) |
| Health | 4679 (15%) | 564 (12%) | 3537 (11%) | 3028 (27%) |
| Housing | 605 (2%) | 16 (0%) | 1243 (4%) | 99 (1%) |
| Immigration | 953 (3%) | 37 (1%) | 67 (0%) | 24 (0%) |
| International Affairs | 1594 (5%) | 805 (17%) | 60 (0%) | 261 (2%) |
| Labor | 1039 (3%) | 43 (1%) | 973 (3%) | 110 (1%) |
| Law and Crime | 2018 (7%) | 251 (5%) | 5759 (17%) | 935 (8%) |
| Local Government and Governance | 0 (0%) | 5 (0%) | 1455 (4%) | 184 (2%) |
| Macroeconomics | 886 (3%) | 62 (1%) | 1709 (5%) | 121 (1%) |
| Private Bills | 112 (0%) | 1 (0%) | 0 (0%) | 0 (0%) |
| Public Lands | 1970 (6%) | 101 (2%) | 512 (2%) | 100 (1%) |
| Social Welfare | 933 (3%) | 125 (3%) | 1177 (4%) | 564 (5%) |
| Technology | 998 (3%) | 56 (1%) | 189 (1%) | 60 (1%) |
| Transportation | 1076 (3%) | 54 (1%) | 1824 (5%) | 250 (2%) |
| Total | 30811 | 4710 | 33330 | 11198 |

Notes: Topic predictions were generated using the jointly-trained (CBP and PAPDP) model with clustering, as depicted in Table **??**. Bill data for Congress and Pennsylvania are provided via *LegiScan*.

Instances where the difference in the model's confidence between its first- and second-best predictions is less than 5 percentage points offer compact examples of the model's comprehension of topics; while a hand-coder might disagree with which particular topic won by virtue of having a slightly higher confidence score, hard-to-classify examples showcase the model's ability to identify the several themes present in these bills. The model's first- and second-best predictions appear to be reasonable, for example the model is split between:

- "Civil Rights" (50.9%) and "Health" (46.2%) for "An Act providing for transgender health benefits."
- "Culture" (43.0%) and "Transportation" (40.4%) for "An Act designating a bridge ... as the Brigadier General Anna Mae. V. McCabe Hays Memorial Bridge."
- "Defense" (49.1%) and "Education" (47.0%) for "An Act amending Title 51 (Military Affairs) ... further providing for educational leave of absence."

Figure 8: Number of pieces of "Health Care" legislation (bills and resolutions) introduced in Pennsylvania and the US Congress by session, 2009-2022.




- "Education" (50.0%) and "Law and Crime" (48.8%) for "An Act amending . . .the Public School Code of 1949, in safe schools, further providing for definitions and for policy relating to bullying."
- "Energy" (49.7%) and "Environment" (45.4%) for "An Act establishing a well impact fee; providing for distribution of fees; establishing the Local Government Shale Impact Fund, the Environmental Shale Impact Mitigation Fund and the Road and Bridge Shale Impact Mitigation Account; and providing for the powers and duties of the Department of Revenue."
- "Health" (47.3%) and "Labor" (46.6%) for "An Act providing for health care assistance for certain steelworkers . . ."
- "Local Government and Governance" (48.0%) and "Housing" (45.2%) for "An Act amending Title 72 (Taxation and Fiscal Affairs) of the Pennsylvania Consolidated Statutes, providing for property tax payments."

Researchers can take heart that the aggregate relationship between these contexts is similar, and the examples where the model disagreed, tend to be close misses.

The data also respond to real world situations. Figure 8 shows the number of "health care" bills introduced in Congress and the Pennsylvania legislature by session. Generally Congress considers more health care legislation. However, both legislatures observed record highs in attention to health care in 2019-2020, the session when the covid pandemic set in. This was also an era characterized by state leadership within the American federal system (Murray and Murray 2023). Interestingly, in 2021-2022, Democrats took hold of Congress and paid high amounts of attention to health care, and attention to the topic subsided in Pennsylvania, perhaps because of the federal leadership on the issue.

Figure 8 demonstrates another advantage of these data, as they allow for inferences in the amount of attention legislators are paying to a topic. Previous dictionary-based methods (e.g. Garlick 2023) used to code state legislatures were consistent over time, but were not reliable at assessing levels of attention. In other words, dictionary methods were good at measuring the changes in the number of bills that mentioned the word "health" over time, and why there's no reason to suspect that term would be biased for health care, the same assumption does not apply to field where lanugage has evolved, for example the word: "uber" becoming synonymous

with transportation, but not existing in that context before 2011 or so.

# 4 Data Records and Usage Notes

Our model output is available in two forms at Open Science Forum: https://osf.io/e2unp/?view_only=e0302a513e7e4cc9999feab2045d47b3. One file has individual confidence estimate for each policy area on each legislative document (40.4 million observations), and the master dataset (1.7 million observations), which identifies a single leading policy area for each legislative document, as well as the next two highest policy areas (Top $K$ Agreement). Both of these are drawn from legiscan, so they follow legiscan's naming conventions for bills and provide a link back to where the information was scraped on legiscan. Further work will link these data to other sources, such as the OpenStates project.

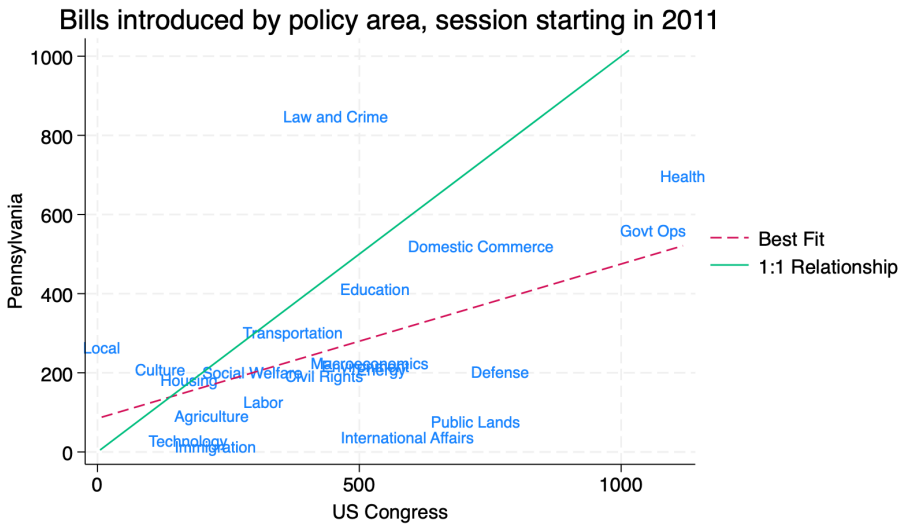
Reporting Top $K$ Agreement and confidence scores grants us additional insights into the nature of the Congressional bills data. The difference in the model's confidence between its first- and second-most confident predictions indicates how strongly it considers the secondary topic to be present. If the model is 50% confident in "Macroeconomics" and 45% confident in "Domestic Commerce," loosely speaking, it appears as if both topics are roughly equivalently present, effectively generating a multi-*label* prediction. Because we lack an intuitive confidence threshold over which to declare a bill should be multi-label, we instead defer to the downstream researcher to inspect the model's per-topic confidence scores, and assess for themselves what threshold is appropriate for their use-case. For example, a researcher interested in recalling any bill that is at least somewhat concerned with "Civil Rights" as a policy area might want to pull bills for which the model's confidence in "Civil Rights" is at least (say) 10%. A researcher faced with a gargantuan pile of "Health" bills might wish to narrow their scope to only those bills that firmly attend to "Health" by looking for confidence scores above 90%.

We recommend aggregating state legislative sessions into biennia. The vast majority of states have two-year sessions following an election in November of even years, like the US Congress. Four states start legislative sessions in even years, following elections in odd-numbered years (LA, MS, NJ, VA). We group bills by the first-year of its typical two-year session (46 states), or the second year of those sessions (LA, MS, NJ, VA). Using these two-year sessions allows us to make sensible temporal comparisons across the federal system, like Figure 9, which shows the number of bills introduced in Congress and Pennsylvania in the 2011-12 legislative session. The fit line being to the right of the 1:1 line shows that Congress typically considers more bills per policy area. The exceptions to this trend are policies with an intensely local focus, like "Law and Crime," owing to state legislative oversight over their criminal justice system and police departments, or "Local Government" itself.

Future channels of research could include using the transformer architecture to code other state-level political documents. For example, nearly every state has institutions producing text similar to what is in the national CAP, such as front pages of papers of record, Gubernatorial "State of the State" addresses, Supreme Court decisions, and rulemaking procedures. If the CAP can be scaled to the state context, and travel over state lines, there is little reason to believe the model could not be used on these types of data. It is particularly useful to have them estimated in a common space as well, as there are many important questions about American federalism (e.g. McCann, Shipan, and Volden 2015, Garlick 2023, Murray and Murray 2023) that are exposed to measurement error from the different contexts.

This paper offers lessons that other researchers should be kept in mind. First, it is useful to have a quality dataset on both sides of the bridge between levels of the federal system, because state and national data are not inherently analogous. Our model trained on only national data underperformed the one trained on national and state data. Second, it is important to apply these models with care, and tinker when necessary. These paper benefitted from untold numbers of iterations to arrive in its robust final form. But our results indicate that the start-up costs with adopting these new tools are a worthy investment.

Figure 9: Number of pieces of legislation introduced in Congress and Pennsylvania by policy, 2011-12




Pennsylvania includes both regular bills and resolutions. Figure excludes "Private Bills" and "Foreign Trade."

## References


Adler, E Scott, and John Wilkerson. 2015. "Congressional Bills Project: 1998-2014." *NSF 00880066 and 00880061* 1 (textbackslashurlhttp://www.congressionalbills.org/). http://www.congressionalbills.org/.

Anzia, Sarah F. 2019. "Looking for influence in all the wrong places: How studying subnational policy can revive research on interest groups." *The Journal of Politics* 81 (1): 343–351.

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. 2019. "Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data." *American Political Science Review* 113 (4): 883–901.

Baumgartner, Frank R., and Bryan D. Jones. 2002. *Policy dynamics.* Chicago: University of Chicago Press. Book.

Campello, Ricardo JGB, Davoud Moulavi, Jörg Sander, and Arthur Zimek. 2013. "Density-based clustering based on hierarchical density estimates." In *Data Mining (ICDM), 2013 IEEE 13th International Conference on,* 160–168. IEEE.

Collingwood, Loren, and John Wilkerson. 2012. "Tradeoffs in accuracy and efficiency in supervised learning methods." *Journal of Information Technology & Politics* 9 (3): 298–318.

Firth, J.R. 1957. "A synopsis of linguistic theory, 1930-1955." *Studies in linguistic analysis,* https://ci.nii.ac.jp/naid/10020680394/.

Garlick, Alex. 2023. "Laboratories of Politics: There is Bottom-up Diffusion of Policy Attention in the American Federal System." *Political Research Quarterly* 76 (1): 29–43.

Publisher: Cambridge Univ Press

Gilens, Martin, and Benjamin I Page. 2014. "Testing theories of American politics: Elites, interest groups, and average citizens." *Perspectives on politics* 12 (03): 564–581.

Hillard, Dustin, Stephen Purpura, and John Wilkerson. 2008. "Computer-assisted topic classification for mixed-methods social science research." *Journal of Information Technology & Politics* 4 (4): 31–46.

Jones, Bryan D. 2016. "The comparative policy agendas projects as measurement systems: response to dowding, hindmoor and martin." *Journal of Public Policy* 36 (1): 31–46.

Lapinski, John S. 2013. *The Substance of Representation: Congress, American Political Development, and Lawmaking.* Princeton University Press.

Lee, Frances E. 2009. *Beyond ideology: Politics, principles, and partisanship in the US Senate.* University of Chicago Press.

Publisher: CSF Associates

McCann, Pamela J Clouser, Charles R Shipan, and Craig Volden. 2015. "Top-Down Federalism: State Policy Responses to National Government Discussions." *Publius: The Journal of Federalism,* 1–31.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* ACM.

McLaughlin, Joseph P, Paul Wolfgang, J Wesley Leckrone, Justin Gollob, Jason Bossie, Jay Jennings, and Michelle J Atherton. 2010. "The Pennsylvania policy database project: A model for comparative analysis." *State Politics & Policy Quarterly* 10 (3): 320–336.

Murray, Gregg R, and Susan M Murray. 2023. "Following the science? Examining the issuance of stay-at-home orders related to COVID-19 by US governors." *American Politics Research* 51 (2): 147–160.

Palmer, Alexis, Noah A Smith, and Arthur Spirling. 2024. "Using proprietary language models in academic research requires explicit justification." *Nature Computational Science* 4 (1): 2–3.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. ""Why should i trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,* 1135–1144.

Širinić, Daniela, and Dario Nikić Čakar. 2019. "Croatian Political Agendas." *Comparative Policy Agendas: Theory, Tools, Data,* 74.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *CoRR* abs/1706.03762. arXiv: 1706.03762. http://arxiv.org/abs/1706.03762.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144.*

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. *XLNet: Generalized Autoregressive Pretraining for Language Understanding.* arXiv: 1906.08237 [cs.CL].

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books." In *Proceedings of the IEEE international conference on computer vision,* 19–27.

# 5 Appendix

## 5.1 Replications of Bag of Words Models to code Congress

### 5.1.1 Compared with Hillard et al. (2008)

Hillard, Purpura, and Wilkerson 2008 demonstrated that a machine learning model could classify the majority of Congressional bills in a way that agrees with hand-coders with a high degree of accuracy. Their procedure involved interpreting documents as bags-of-words, including the usual dimensionality reduction steps of removing stop-words, removing numbers, stemming words, and removing case sensitivity. They combined a Naïve Bayes classifier, a support vector machine, a *MaxEnt* model, and a *Boostexter* model in an "ensemble" approach where the model's predictions from each classifier are combined in a vote system. Where the model's classifiers agree with each other (85% of all bills), their approach achieves a 94% accuracy; where they disagree, the accuracy is 61%.

For benchmarking model performance against Hillard, Purpura, and Wilkerson 2008, we report results for predicting major topics. They train their model on the $80^{th}$ through $105^{th}$ Congress — approximately 374,000 bills — and split the data 50-50 into training and validation sets, and we do the same. Further, to match their procedure, we do not sort the training and validation datasets in any particular way with respect to duplicate bill titles. At the time of their writing, "Culture" was a subtopic of "Education," and "Immigration" as a subtopic of "Labor." I collapse bills from the "Culture" and "Immigration" topics back to "Education" and "Labor," respectively, to match their procedure.

Table 13 presents the model benchmarked against Hillard, Purpura, and Wilkerson 2008. The overall accuracy of their model was 89.0%, and our model attains 91.2%. They also looked specifically at their best 85% of predictions, where their "ensemble" model unanimously votes[13] for the same topic to assign to a bill; for this subset of bills, they achieve 94% accuracy. For our model's best 85% of predictions (its 85% most confident predictions), it achieves an accuracy of 96.3%. If the desired accuracy is 94%, their model, again, can reach this for 85% of bills, whereas our model can for 93.7%.

Table 13: (Accuracy) Model Benchmark - Hillard, Purpura, and Wilkerson 2008

| | Accuracy (Overall) | Accuracy (Best 85% of Predictions) | % of Bills Achieving 94% Accuracy |
|---|---|---|---|
| Hillard, Purpura, and Wilkerson 2008 | 89.0% | 94% | 85% |
| Dee (2022) | 91.2% | 96.3% | 93.7% |

Notes: To match Hillard, Purpura, and Wilkerson 2008, the model was trained on the $80^{th} - 105^{th}$ Congresses, and duplicate bill titles were not dropped. The best 85% predictions reflect the case where Hillard, Purpura, and Wilkerson 2008's ensemble model unanimously agrees on the bill's topic, and likewise examining my model's 85% most confident predictions. When Hillard, Purpura, and Wilkerson 2008 focuses on the bills for which their ensemble model agrees, they achieve 94% accuracy. In the rightmost column, I examine what percentage of bills can be predicted by my model with a criterion of a minimum 94% accuracy.

### 5.1.2 Compared with Collingwood and Wilkerson (2012)

Collingwood and Wilkerson 2012 demonstrated that a machine learning model could be trained to emulate the hand-coders of the CBP and do so for an appreciable level of accuracy given the small dataset with which they experimented. As a result, the CBP "now relies on [the machine learning model developed in their work] to classify a large proportion of bills at similarly high levels of reliability [compared to hand-coders]" (Collingwood and Wilkerson 2012). The model

13. For their main results, they build a support vector machine, a *MaxEnt* model, and a *Boostexter* model. Where the three models agree on the predicted topic of the bill, the "ensemble" of all three reflects a unanimous vote for that topic.

developed in that paper is an ensemble of four off-the-shelf machine learning algorithms that utilize a bag of words representation of language. In their main results, to explore the potential for a machine learning model to be a viable replacement for hand-coders when new data are introduced, they artificially restrict their training data to be a stratified sample of anywhere from 100 to 1,000 bills per topic, depending on the specification, or a purely random draw from the entire corpus. In the interest of space, I report results for these two sample size extremes of 100 and 1,000 bills per topic, stratify the training data to ensure a uniform class distribution, and present the model's $F_1$ score as a means of balancing the relative importance of *Precision* and *Recall*.

To further mimic the analysis from Collingwood and Wilkerson 2012, we restrict the temporal scope of the Congressional bills data to the $90^{th}$-$106^{th}$ Congresses. At the time of their writing, the "Culture" and "Immigration" issue areas did not exist as their own major topics, but instead as subtopics of "Education" and "Labor," respectively, and thus we return bills that are now coded in the former areas to the latter. The approach to handling duplicate bill titles in the data in Collingwood and Wilkerson 2012 was to drop those duplicates, and we do the same here. Regarding the inconsistently hand-coded bill titles, we lack information as to which copy of each bill title survived this process, and thus cannot identify which topic each bill should be coded as. For $n$ =100 and $n$ =1,000 bills per topic, I train the model 30 times, redrawing the training and validation samples each time to effectively "bootstrap" the model's performance. Table 14 presents the model benchmarked against Collingwood and Wilkerson 2012 for $n$ =100 and $n$ =1,000 bills per topic. The left panel reports model performance for the case where there are $n$ =100 bills per topic in the training and validation sets. Across all topics, the model out-performs Collingwood and Wilkerson 2012 by an average of 11.8 percentage points. Redrawing the training and validation sets for a total of 30 iterations also reveals that the model is far more stable, with a standard deviation in per-topic performance of 3.1 percentage points across training runs. The right panel reports model performance for the case where there are $n$ =1,000 bills per topic in the training and validation sets. For most topics, the model using $n$ =100 performs at least as well as Collingwood and Wilkerson 2012 using $n$ =1,000. When the model is trained on $n$ =1,000 bills per topic, it out-performs Collingwood and Wilkerson 2012 by an average of 4.8 percentage points. Moreover, the model is far more consistent across training runs, with a standard deviation in per-topic $F_1$ scores of approximately 0.8 percentage points across training iterations.

Table 14: ($F_1$ Score) Model Benchmark - Collingwood and Wilkerson 2012

| *All values are the model's $F_1$ Score in percentage point terms* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $n = 100$ per topic ($N = 2,000$) | | | | $n = 1,000$ per topic ($N = 20,000$) | | | |
| | CW (2012) | | Dee (2022) | | CW (2012) | | Dee (2022) | |
| Topic | Mean | Std. Dev. | Mean | Std. Dev. | Mean | Std. Dev. | Mean | Std. Dev. |
| Macroeconomics | 54.3 | 7.0 | 75.8 (+21.5) | 4.1 (-2.9) | 68.2 | 2.1 | 76.8 (+8.6) | 1.2 (-0.9) |
| Civil Rights | 60.4 | 8.7 | 78.0 (+17.6) | 2.0 (-6.7) | 75.4 | 2.2 | 82.1 (+6.7) | 0.9 (-1.3) |
| Health | 71.9 | 7.5 | 83.8 (+11.9) | 3.5 (-4.0) | 80.8 | 1.9 | 86.7 (+5.9) | 0.8 (-1.1) |
| Agriculture | 73.2 | 7.6 | 80.3 (+7.1) | 1.2 (-6.4) | 83.1 | 1.8 | 85.5 (+2.4) | 0.9 (-0.9) |
| Labor | 65.9 | 8.4 | 75.5 (+9.6) | 4.6 (-3.8) | 76.8 | 2.0 | 81.1 (+4.3) | 1.4 (-0.6) |
| Education | 74.0 | 7.6 | 82.7 (+8.7) | 2.9 (-4.7) | 83.6 | 1.8 | 86.8 (+3.2) | 0.5 (-1.3) |
| Environment | 67.4 | 8.0 | 81.3 (+13.9) | 1.9 (-6.1) | 79.6 | 2.0 | 84.7 (+5.1) | 1.0 (-1.0) |
| Energy | 77.1 | 7.3 | 85.6 (+8.5) | 3.3 (-4.0) | 86.7 | 1.7 | 89.3 (+2.6) | 1.0 (-0.7) |
| Transportation | 68.5 | 8.0 | 82.1 (+13.6) | 3.5 (-4.5) | 81.3 | 2.0 | 85.8 (+4.5) | 0.7 (-1.3) |
| Law and Crime | 57.9 | 8.5 | 80.1 (+22.2) | 2.0 (-6.5) | 74.0 | 2.2 | 82.1 (+8.1) | 0.8 (-1.4) |
| Social Welfare | 73.6 | 7.2 | 82.0 (+8.4) | 3.3 (-3.9) | 80.3 | 2.0 | 83.2 (+2.9) | 1.3 (-0.7) |
| Housing | 73.9 | 7.5 | 82.8 (+8.9) | 3.1 (-4.4) | 82.4 | 1.8 | 85.8 (+3.4) | 1.1 (-0.7) |
| Domestic Commerce | 48.9 | 9.1 | 69.9 (+21.0) | 2.9 (-6.2) | 67.1 | 2.6 | 74.9 (+7.8) | 0.9 (-1.7) |
| Defense | 65.6 | 8.6 | 76.1 (+10.5) | 2.0 (-6.6) | 78.0 | 2.0 | 82.5 (+4.5) | 1.1 (-0.9) |
| Technology | 74.2 | 7.4 | 85.6 (+11.4) | 3.9 (-3.5) | 84.5 | 1.7 | 88.8 (+4.3) | 0.7 (-1.0) |
| Foreign Trade | 80.3 | 6.6 | 81.2 (+0.9) | 3.8 (-2.8) | 86.0 | 1.7 | 87.1 (+1.1) | 0.6 (-1.1) |
| International Affairs | 61.1 | 8.1 | 76.0 (+14.9) | 3.5 (-4.6) | 75.8 | 2.0 | 83.3 (+7.5) | 0.7 (-1.3) |
| Government Operations | 51.4 | 8.6 | 66.9 (+15.5) | 4.4 (-4.2) | 64.8 | 2.6 | 73.5 (+8.7) | 1.3 (-1.3) |
| Public Lands | 70.6 | 7.6 | 79.7 (+9.1) | 4.3 (-3.3) | 80.1 | 2.1 | 84.3 (+4.2) | 1.0 (-1.1) |
| Private Bills | 95.1 | 3.1 | 97.2 (+2.1) | 1.1 (-2.0) | 96.6 | 1.1 | 98.2 (+1.6) | 0.3 (-0.8) |
| Avg. Across Topics | 68.3 | 7.6 | 80.1 (+11.8) | 3.1 (-4.5) | 79.3 | 2.0 | 84.1 (+4.8) | 0.8 (-1.2) |

Notes: For comparability, the "Immigration" topic is combined with the "Labor" topic, and the "Culture" topic with the "Education" topic, to match the Comparative Agends Project codebook at the time of Collingwood and Wilkerson 2012. All values reported refer to training and validating on the CBP data from the $90^{th} - 106^{th}$ Congresses. "Duplicate" bills, in terms of bill titles, were dropped from the sample at-random. Values reported for Collingwood and Wilkerson 2012 come from their Table 3.