AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Keyword-Assisted Topic Models

Shusei Eshima and **Kosuke Imai** Harvard University
Tomoya Sasaki Massachusetts Institute of Technology

Abstract: *In recent years, fully automated content analysis based on probabilistic topic models has become popular among social scientists because of their scalability. However, researchers find that these models often fail to measure specific concepts of substantive interest by inadvertently creating multiple topics with similar content and combining distinct themes into a single topic. In this article, we empirically demonstrate that providing a small number of keywords can substantially enhance the measurement performance of topic models. An important advantage of the proposed keyword-assisted topic model (keyATM) is that the specification of keywords requires researchers to label topics prior to fitting a model to the data. This contrasts with a widespread practice of post hoc topic interpretation and adjustments that compromises the objectivity of empirical findings. In our application, we find that keyATM provides more interpretable results, has better document classification performance, and is less sensitive to the number of topics.*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/RKNNVL.

Textual data represent the most fundamental way of recording and preserving human communication and activities. Social scientists have analyzed texts to measure a variety of substantive concepts such as political ideology and policy positions (Laver, Benoit, and Garry 2003; Otjes and Green-Pedersen 2021). A typical process would require researchers to read and manually classify relevant documents into different categories based on a codebook prepared specifically for measuring substantive concepts of interest (e.g., Bauer 2000). Given the lack of scalability of this traditional approach, social scientists are increasingly relying on fully automated content analysis based on machine learning models (Grimmer and Stewart 2013). In particular, probabilistic topic models have been widely used to uncover the content of documents and to explore the relations between discovered topics and meta-information such as author characteristics (e.g., Blei, Ng, and Jordan 2003; Grimmer 2010; Roberts, Stewart, and Airoldi 2016).

Unfortunately, although topic models can *explore* themes of a corpus (e.g., Roberts et al. 2014), they do not necessarily *measure* specific concepts of substantive interest. Although researchers have also relied upon topic models for measurement purposes (e.g., Bagozzi and Berliner 2018; Blaydes, Grimmer, and McQueen 2018; Barberá et al. 2019; Dietrich, Hayes, and O'brien 2019; Grimmer 2013; Martin and McCrain 2019), they acknowledge that these fully automated models often inadvertently create multiple topics with similar content and combine different themes into a single topic (Chang et al.

Shusei Eshima, Graduate Student, Department of Government, Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138 (shuseieshima@g.harvard.edu). Kosuke Imai, Professor, Department of Government and Department of Statistics, Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138 (imai@harvard.edu). Tomoya Sasaki, Graduate Student, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue E53, Cambridge, MA 02142 (tomoyas@mit.edu).

The proposed methodology is implemented via an open-source software package keyATM, which is available at https://cran.r-project.org/package=keyATM. We thank Doug Rice and Yutaka Shinada for sharing their data, Luwei Ying, Jacob Montgomery, and Brandon Stewart for sharing with us their experience of setting up validation exercises, and Soichiro Yamauchi for advice on methodological and computational issues. We also thank Soubhik Barari, Matthew Blackwell, Max Goplerud, Andy Halterman, Masataka Harada, Hiroto Katsumata, Gary King, Dean Knox, Shiro Kuriwaki, Will Lowe, Luke Miratrix, Hirofumi Miwa, Daichi Mochihashi, Santiago Olivella, Yon Soo Park, Reed Rasband, Hunter Rendleman, Sarah Mohamed, Yuki Shiraito, Tyler Simko, and Diana Stanescu, as well as seminar participants at the Institute for Quantitative Social Science Applied Statistics Workshop, the Japanese Society for Quantitative Political Science 2020 Winter Meeting, International Methods Colloquium, Annual Conference of the Society for Political Methodology (2020), and Annual Conference of the American Political Science Association (2020) for helpful discussions and comments on this project. Lastly, we thank the editors and our three anonymous reviewers for providing us with additional comments.

American Journal of Political Science, Vol. 68, No. 2, April 2024, Pp. 730–750

 DOI: 10.1111/ajps.12779

2009; Morstatter and Liu 2016; Newman, Bonilla, and Buntine 2011). These undesirable features may impede obtaining a clear interpretation of topics and adequate measurements of substantive concepts. This mismatch is not surprising because these models do not directly incorporate the information about topics of interest. Researchers would not know whether a model yields topics that properly measure substantive concepts until they fit the model. For this reason, scholars have emphasized the importance of human validation (Grimmer and Stewart 2013; Ying, Montgomery, and Stewart 2022).

Another unsatisfactory characteristic of the current approaches is that researchers must interpret and label estimated topics after model fitting. This task is crucial especially when topic models are used for measurement. Researchers must make a post hoc decision about drawing a connection between estimated topics and substantive concepts of interest. Together with commonly used post hoc adjustments of topics such as topic merging or word reweighting (Bischof and Airoldi 2012), this widespread practice can compromise the scientific objectivity of empirical findings. Finally, the empirical results obtained under probabilistic topic models are known to be sensitive to the number of topics (Boyd-Graber, Mimno, and Newman 2014; Roberts, Stewart, and Tingley 2016).

In this article, we propose the keyword-assisted topic models (keyATM) that allow researchers to label topics via the specification of keywords *prior to* model fitting. This semisupervised approach avoids post hoc interpretation and adjustments of topics because researchers can use keywords to specify the substantive concepts to be measured. Unlike the popular models such as the latent Dirichlet allocation (LDA, Blei, Ng, and Jordan 2003) and the structural topic model (STM, Roberts, Stewart, and Airoldi 2016), the keyATM methodology is not an unsupervised topic model that only uses unlabeled data. Rather, it is a semisupervised topic model that combines a small amount of information with a large amount of unlabeled data. As opposed to a supervised approach, keyATM only requires a small number of keywords for each concept of interest and does not necessitate manual labeling of many documents. We emphasize that keyATM is most useful when a researcher is interested in measuring specific topics. If the goal is to explore topics, the existing topic models such as LDA might be more appropriate.

We empirically demonstrate that providing topic models with a small number of keywords substantially improves their performance and better serves the purpose of measurement. We assess the performance of topics models both qualitatively, by examining the most frequent words for each estimated topic and employing crowd-sourced validation (Ying, Montgomery, and Stewart 2022), and quantitatively, by comparing the document classification with human coding and conducting a crowdsource validation study.

The proposed keyATM methodology builds upon the model originally introduced by Jagarlamudi, Daumé III, and Udupa (2012) by making the following improvements. First, keyATM can have topics with no keyword. Researchers may not be able to prepare keywords for all topics that potentially exist in a corpus (King, Lam, and Roberts 2017). Thus, we leave room for *exploring* a corpus with these additional no-keyword topics. Second, researchers can characterize document–topic distribution with meta-information to study how estimated topics vary as a function of document-level covariates and over time. Third, keyATM is a fully Bayesian approach and estimates hyperparameters to improve the model performance (Wallach, Mimno, and McCallum 2009).

Related Methods

Since the pioneering work of Blei, Ng, and Jordan (2003), numerous researchers have worked on improving topic interpretability (Blei 2012). We neither claim that keyATM is better than other existing approaches nor attempt to provide a comprehensive review of this large literature here. Instead, we briefly compare keyATM with the existing closely related models. Most importantly, the base keyATM adds the aforementioned improvements to the model of Jagarlamudi, Daumé III, and Udupa (2012). Li et al. (2019) propose a model similar to the base keyATM under the assumption that each document has a single keyword topic and some topics with no keyword. In contrast, keyATM allows each document to belong to multiple keyword topics.

Some researchers have also used human inputs to restrict the parameters of topic models. For example, Chemudugunta et al. (2008) incorporate keywords into a topic model. However, unlike keyATM, their model assumes that prespecified keywords cannot belong to other topics and each topic with such keywords cannot have other keywords. Hu, Boyd-Graber, and Satinoff (2011) propose a topic model where researchers refine the discovered topics by iteratively adding constraints such that certain sets of words are forced to appear together in the same topic. keyATM does not place such a strict constraint and also avoids an iterative refinement process.

Other researchers have incorporated substantive knowledge by placing an informative prior distribution over topic–word distributions. Lu et al. (2011)

modify the values of prior parameters for keywords, essentially inflating their frequency. Fan, Doshi-Velez, and Miratrix (2019) employ a similar strategy when constructing informative priors using a combination of term frequency-inverse document frequency (TF-IDF) and domain knowledge. Although these models require researchers to specify the particular values of hyper prior parameters that directly control the importance of keywords, our approach imposes a restriction on the structure of prior distribution so that the models are allowed to learn from the data the importance of keywords (Supporting Information [SI], p. 3, explains this connection in detail). Hansen, Ringger, and Seppi (2013) and Wood et al. (2017) use external corpus such as Wikipedia to construct either topic–word distributions or hyperparameters.

Placing a certain structure on the prior information is a common strategy. For example, Newman, Bonilla, and Buntine (2011) model the structural relations among different words using an external data. Xie, Yang, and Xing (2015) incorporate the information about the similarity of words through a Markov random field, whereas Andrzejewski, Zhu, and Craven (2009) specify a set of words that have a similar probability within a certain topic through a Dirichlet Forest prior distribution.

In contrast to these existing approaches, keyATM directly incorporates a small number of keywords into a topic–word distribution. We believe that this simplicity of keyATM is particularly appealing for social scientists.

The Base keyATM

We begin by describing the base keyATM, which we will extend in various ways. The base keyATM improves the model of Jagarlamudi, Daumé III, and Udupa (2012) by allowing some topics to have no keyword and estimating hyperparameters for better empirical performance. Our application demonstrates that this base keyATM yields results superior to those of LDA qualitatively and quantitatively.

Model

Suppose that we have a total of D documents and each document d has N_d words. These documents contain a total of V unique words. Let w_{di} represent the ith word in document d where $\mathcal{W}_d = \{w_{d1}, w_{d2}, \ldots, w_{dN_d}\}$ represents the set of all words used in document d. We are interested in identifying the topics that underlie each document. We consider two types of topics: topics with keywords, which are of primary interest to researchers and are referred to as *keyword topics*, and topics without keywords, which we call *no-keyword topics*. Suppose that we have a total of K topics and the first $\widetilde{K}$ of them are keyword topics, that is, $\widetilde{K} \leq K$. For each keyword topic k, researchers provide a set of L_k keywords, which is denoted by $\mathcal{V}_k = \{v_{k1}, v_{k2}, \ldots, v_{kL_k}\}$. Note that the same keywords may be used for different keyword topics and keywords are a part of total V unique words.

Our model is based on the following data generation process. For each word i in document d, we first draw the latent topic variable $z_{di} \in \{1, 2, \ldots, K\}$ from the topic distribution of the document,

$$z_{di} \overset{\text{indep.}}{\sim} \mathsf{Categorical}(\boldsymbol{\theta}_d),$$

where $\boldsymbol{\theta}_d$ is a K-dimensional vector of topic probabilities for document d with $\sum_{k=1}^{K} \theta_{dk} = 1$. This document–topic distribution $\boldsymbol{\theta}_d$ characterizes the relative proportion of each topic for document d.

If the sampled topic is one of the no-keyword topics, then we draw the word w_{di} from the corresponding word distribution of the topic,

$$w_{di} \mid z_{di} = k \overset{\text{indep.}}{\sim} \mathsf{Categorical}(\boldsymbol{\phi}_k) \quad \text{for } k \in \{\widetilde{K}+1, \widetilde{K}+2, \ldots, K\},$$

where $\boldsymbol{\phi}_k$ is a V-dimensional vector of word probabilities for topic k with $\sum_{v=1}^{V} \phi_{kv} = 1$. This probability vector represents the relative frequency of each word within topic k.

On the other hand, if the sampled topic has keywords, we draw a Bernoulli random variable s_{di} with success probability π_k for word i in document d. If this variable is equal to 1, then word w_{di} is drawn from the set of keywords for the topic based on probability vector $\widetilde{\boldsymbol{\phi}}_k$. In contrast, if s_{di} is equal to 0, then we sample the word from the standard topic–word distribution of the topic $\boldsymbol{\phi}_k$. Therefore, we have,

$$s_{di} \mid z_{di} = k \overset{\text{indep.}}{\sim} \mathsf{Bernoulli}(\pi_k) \quad \text{for } k \in \{1, 2, \ldots, \widetilde{K}\},$$

$$w_{di} \mid s_{di}, z_{di} = k \overset{\text{indep.}}{\sim} \begin{cases} \mathsf{Categorical}(\boldsymbol{\phi}_k) & \text{if } s_{di} = 0 \\ \mathsf{Categorical}(\widetilde{\boldsymbol{\phi}}_k) & \text{if } s_{di} = 1 \end{cases} \quad \text{for } k \in \{1, 2, \ldots, \widetilde{K}\},$$

where π_k represents the probability of sampling from the set of keywords, and $\widetilde{\boldsymbol{\phi}}_k$ is a V dimensional vector of word probabilities for the set of keywords of topic k, that is, $\mathcal{V}_k$. Thus, L_k of V elements in $\widetilde{\boldsymbol{\phi}}_k$ have positive values and the others are 0. We use the following prior distributions:

$$\pi_k \overset{\text{i.i.d.}}{\sim} \mathsf{Beta}(\gamma_1, \gamma_2) \quad \text{for } k = 1, 2, \ldots, \widetilde{K}$$

$$\boldsymbol{\phi}_k \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\boldsymbol{\beta}) \quad \text{for } k = 1, 2, \ldots, K$$
$$\widetilde{\boldsymbol{\phi}}_k \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\widetilde{\boldsymbol{\beta}}) \quad \text{for } k = 1, 2, \ldots, \widetilde{K} \tag{1}$$

$$\boldsymbol{\theta}_d \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\boldsymbol{\alpha}) \quad \text{for } d = 1, 2, \ldots, D \tag{2}$$

$$\alpha_k \overset{\text{indep.}}{\sim} \begin{cases} \text{Gamma}(\tilde{\eta}_1, \tilde{\eta}_2) & \text{for } k = 1, 2, \ldots, \widetilde{K} \\ \text{Gamma}(\eta_1, \eta_2) & \text{for } k = \widetilde{K}+1, \widetilde{K}+2, \ldots, K, \end{cases} \tag{3}$$

where with slight abuse of notation we specify the prior distribution for $\widetilde{\boldsymbol{\phi}}_k$, with the constant hyperparameter value placed only on keyword elements of the topic.

In typical applications, the choice of hyperparameters matters little so long as the amount of data are sufficiently large.[1] The only exception is the prior for π_k, which controls the influence of keywords. We use the uniform prior distribution for π_k, that is, $\gamma_1 = \gamma_2 = 1$ as a noninformative prior, which we find works well across a variety of applications.

As shown above, keyATM is based on a mixture of two distributions, one with positive probabilities only for keywords and the other with positive probabilities for all words. This mixture structure makes the prior means for the frequency of user-selected keywords given a topic greater than those of nonkeywords in the same topic. In addition, the prior variance is also larger for the frequency of keywords given a topic than for nonkeywords. This encourages keyATM to place greater importance on keywords a priori while allowing the model to learn from the data about the precise degree to which keywords matter for a given topic. Because keyword topics are distinct, keyATM is less prone to the label switching problem.

Sampling Algorithm

We next describe the sampling algorithm. To improve the empirical performance of the model, we use term weights that help prevent highly frequent words from dominating the resulting topics (Wilson and Chew 2010, hereafter wLDA). We use a collapsed Gibbs sampling algorithm to sample from the posterior distribution by integrating out the variables $(\boldsymbol{\theta}, \boldsymbol{\phi}, \widetilde{\boldsymbol{\phi}}, \boldsymbol{\pi})$ (Griffiths and Steyvers 2004). This yields a Markov chain of $(\boldsymbol{z}, \mathbf{s}, \boldsymbol{\alpha})$ and helps address the identifiability problem regarding $\boldsymbol{\phi}_{kv}$, $\widetilde{\boldsymbol{\phi}}_{kv}$, and π_k.

From the expression of the collapsed posterior distribution, it is straightforward to derive the conditional posterior distribution of each parameter. First, the sampling distribution of topic assignment for each word i in document d is given by,

$$\Pr(z_{di} = k \mid \mathbf{z}^{-di}, \mathbf{w}, \mathbf{s}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma})$$
$$\propto \begin{cases} \dfrac{\beta_v + n_{kv}^{-di}}{\sum_v \beta_v + n_k^{-di}} \cdot \dfrac{n_k^{-di} + \gamma_2}{\tilde{n}_k^{-di} + \gamma_1 + n_k^{-di} + \gamma_2} \cdot \left(n_{dk}^{-di} + \alpha_k\right) \\ \quad \text{if } s_{di} = 0, \\ \dfrac{\tilde{\beta}_v + \tilde{n}_{kv}^{-di}}{\sum_{v \in \mathcal{V}_k} \tilde{\beta}_v + \tilde{n}_k^{-di}} \cdot \dfrac{\tilde{n}_k^{-di} + \gamma_1}{\tilde{n}_k^{-di} + \gamma_1 + n_k^{-di} + \gamma_2} \cdot \left(n_{dk}^{-di} + \alpha_k\right) \\ \quad \text{if } s_{di} = 1, \end{cases}$$

where n_k^{-di} ($\tilde{n}_k^{-di}$) represents the number of words (keywords) in the documents assigned to topic (keyword topic) k excluding the ith word of document d. Similarly, n_{kv}^{-di} ($\tilde{n}_{kv}^{-di}$) denotes the number of times word (keyword) v is assigned to topic (keyword topic) k again excluding the ith word of document d, and n_{dk}^{-di} represents the number of times word v is assigned to topic k in document d excluding the ith word of document d.

Next, we sample s_{di} from the following conditional posterior distribution:

$$\Pr(s_{di} = s \mid \mathbf{s}^{-di}, \boldsymbol{z}, \mathbf{w}, \boldsymbol{\beta}, \widetilde{\boldsymbol{\beta}}, \boldsymbol{\gamma})$$
$$\propto \begin{cases} \dfrac{\beta_v + n_{z_{di}v}^{-di}}{\sum_v \beta_v + n_{z_{di}}^{-di}} \cdot \left(n_{z_{di}}^{-di} + \gamma_2\right) & \text{if} \quad s = 0, \\ \dfrac{\tilde{\beta}_v + \tilde{n}_{z_{di}v}^{-di}}{\sum_{v \in \mathcal{V}_{z_{di}}} \tilde{\beta}_v + \tilde{n}_{z_{di}}^{-di}} \cdot \left(\tilde{n}_{z_{di}}^{-di} + \gamma_1\right) & \text{if} \quad s = 1. \end{cases}$$

Finally, the conditional posterior distribution of α_k is given by,

$$p(\alpha_k \mid \boldsymbol{\alpha}_{-[k]}, \mathbf{s}, \boldsymbol{z}, \mathbf{w}, \tilde{\boldsymbol{\eta}}) \propto \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right) \prod_{d=1}^D \Gamma(n_{dk} + \alpha_k)}{\Gamma(\alpha_k) \prod_{d=1}^D \Gamma\left(\sum_{k=1}^K n_{dk} + \alpha_k\right)} \cdot \alpha_k^{\tilde{\eta}_1 - 1} \exp(-\tilde{\eta}_2 \alpha_k), \tag{5}$$

for $k = 1, 2, \ldots, \widetilde{K}$. For $k = \widetilde{K}+1, \ldots, K$, the conditional distribution is identical except that $\tilde{\eta}_1$ and $\tilde{\eta}_2$ are replaced with η_1 and η_2. We use an unbounded slice sampler to efficiently sample from a large parameter space (Mochihashi 2020).

As mentioned earlier, we apply the weighting method of wLDA when computing these word counts in the collapsed Gibbs sampler so that frequently occurring words do not overwhelm other meaningful words. Based on the information theory, Wilson and Chew (2010) propose a weighting scheme based on $-\log_2 p(v)$ where $p(v)$ is estimated using an observed frequency of term v. The weight for a term v is defined as

$$m(v) \;=\; -\log_2 \frac{\sum_{d=1}^D \sum_{i=1}^{N_d} \mathbb{1}(w_{di} = v)}{\sum_{d=1}^D N_d}. \tag{6}$$

[1]The default values are: $\gamma_1 = \gamma_2 = 1$, $\beta = 0.01$, $\tilde{\beta} = 0.1$, $\eta_1 = 2$, $\eta_2 = 1$, and $\tilde{\eta}_1 = \tilde{\eta}_2 = 1$. But, these values can be adjusted reflecting one's prior knowledge.

Then, the weighted word counts used in the collapsed Gibbs sampler are given by

$$n_{kv} = m(v)\sum_{d=1}^{D}\sum_{i=1}^{N_d}\mathbb{1}(w_{di}=v)\mathbb{1}(s_{di}=0)\mathbb{1}(z_{di}=k),$$

$$\tilde{n}_{kv} = m(v)\sum_{d=1}^{D}\sum_{i=1}^{N_d}\mathbb{1}(w_{di}=v)\mathbb{1}(s_{di}=1)\mathbb{1}(z_{di}=k),$$

$$n_{dk} = \sum_{i=1}^{N_d} m(w_{di})\mathbb{1}(z_{di}=k),$$

where $m(v)=1$ for all v corresponds to the unweighted sampler.

Model Interpretation

To interpret the fitted keyATM, we focus on two quantities of interest. The topic–word distribution represents the relative frequency of words for each topic, characterizing the topic content. The document–topic distribution represents the proportions of topics for each document, reflecting the main themes of the document.

We obtain a single topic–word distribution $\boldsymbol{\phi}_k^*$ for keyword topics by combining both $\boldsymbol{\phi}_k$ and $\widetilde{\boldsymbol{\phi}}_k$ according to the following mixture structure assumed under the model for each word v of topic k,

$$\phi_{kv}^* = (1-\pi_k)\phi_{kv} + \pi_k\widetilde{\phi}_{kv}. \tag{7}$$

Because both $\boldsymbol{\phi}_k$ and $\widetilde{\boldsymbol{\phi}}_k$ are marginalized out, we compute the marginal posterior mean as our estimate of topic–word distribution,

$$\begin{aligned}
&\mathbb{E}[\phi_{kv}^* \mid \mathbf{w}] \\
&\quad = \mathbb{E}\Big\{\mathbb{E}[\phi_{kv}^* \mid \beta_v, \widetilde{\beta}_v, \boldsymbol{\gamma}, \mathbf{s}, \boldsymbol{z}, \mathbf{w}] \;\Big|\; \mathbf{w}\Big\} \\
&\quad = \begin{cases}
\mathbb{E}\left[\dfrac{n_k+\gamma_2}{\tilde{n}_k+\gamma_1+n_k+\gamma_2}\cdot\dfrac{\beta_v+n_{kv}}{\sum_{v'}\beta_{v'}+n_k}\right. \\
\qquad \left. +\dfrac{\tilde{n}_k+\gamma_1}{\tilde{n}_k+\gamma_1+n_k+\gamma_2}\cdot\dfrac{\tilde{\beta}_v+\tilde{n}_{kv}}{\sum_{v'\in\mathcal{V}_k}\tilde{\beta}_{v'}+\tilde{n}_{kv}} \;\Big|\; \mathbf{w}\right] \\
\qquad \text{if } v\in\mathcal{V}_k \\
\mathbb{E}\left[\dfrac{n_k+\gamma_2}{\tilde{n}_k+\gamma_1+n_k+\gamma_2}\cdot\dfrac{\beta_v+n_{kv}}{\sum_{v'}\beta_{v'}+n_k} \;\Big|\; \mathbf{w}\right] \\
\qquad \text{if } v\notin\mathcal{V}_k,
\end{cases}
\end{aligned} \tag{8}$$

where $n_k=\sum_{v=1}^{V} n_{kv}$, $\tilde{n}_k=\sum_{v=1}^{V}\tilde{n}_{kv}$, and the second equality follows from the conditional independence relations assumed under the model. Similarly, although the document–topic distribution θ_{dk} is also marginalized out, we compute its marginal posterior,

$$\begin{aligned}
\mathbb{E}[\theta_{dk} \mid \mathbf{w}] &= \mathbb{E}\Big\{\mathbb{E}[\theta_{dk} \mid \alpha_k, \boldsymbol{z}, \mathbf{w}] \;\Big|\; \mathbf{w}\Big\} \\
&= \mathbb{E}\left[\frac{\alpha_k+n_{dk}}{\sum_{k'=1}^{K}\alpha_{k'}+n_{dk'}} \;\Big|\; \mathbf{w}\right],
\end{aligned} \tag{9}$$

for each document d and topic k.

Empirical Evaluation

We assess the empirical performance of the base keyATM by analyzing the texts of Congressional bills, using labels and keywords compiled by the Comparative Agenda Project (CAP).[2] We show that keyATM yields more interpretable topic– word distributions than wLDA. Recall that the only difference between keyATM and wLDA is the existence of keyword topics. In addition, topic–word distributions obtained from keyATM are more consistent with the human-coded labels given by the CAP. Finally, we validate the topic classification of these bills against the corresponding human coding obtained from the Congressional Bills Project (CBP).[3] We find that keyATM outperforms wLDA, illustrating the improved quality of estimated document–topic distributions. A greater correspondence between the human coding and keyATM outputs suggests the advantage of keyATM when used for measuring specific topics of interest.

Data and Setup. We analyze the Congressional bills that were subject to floor votes during the 101st to 114th Sessions.[4] These bills are identified via Voteview[5] and their texts are obtained from `congress.gov`.[6] There are a total of 4,421 such bills with an average of 316 bills per session. We preprocess the raw texts by first removing stop words via the R package `quanteda` (Benoit et al. 2018), then pruning words that appear less than 11 times in the corpus, and lemmatizing the remaining words via the Python library `NLTK` (Bird, Klein, and Loper 2009).[7] After preprocessing, we have on average 5,537 words per bill and 7,776 unique words in the entire corpus.

[2] https://www.comparativeagendas.net , last accessed on December 10, 2019.

[3] http://www.congressionalbills.org/codebooks.html , last accessed on December 10, 2019.

[4] These sessions are chosen for the availability of data.

[5] https://voteview.com , last accessed on December 10, 2019.

[6] https://www.congress.gov/ , last accessed on December 10, 2019.

[7] See SI, p. 4, for details.

TABLE 1 Frequency of Each Topic and Its Most Frequent Keywords

Topic Label	Count	Percentage	Most Frequent Keywords
Government operations	864	19.54	administrative capital collection
Public lands	464	10.50	land resource water
Defense	433	9.79	security military operation
Domestic commerce	392	8.87	cost security management
Law & crime	274	6.20	code family court
Health	272	6.15	cost health payment
International affairs	207	4.68	committee foreign develop
Transportation	191	4.32	construction transportation air
Macroeconomics	177	4.00	cost interest budget
Environment	163	3.69	resource water protection
Education	138	3.12	education area loan
Energy	132	2.99	energy vehicle conservation
Technology	131	2.96	transfer research technology
Labor	111	2.51	employee benefit standard
Foreign trade	110	2.49	agreement foreign international
Civil rights	102	2.31	information contract right
Social welfare	73	1.65	assistance child care
Agriculture	68	1.54	product food market
Housing	65	1.47	housing community family
Immigration	52	1.18	immigration refugee citizenship
Culture	2	0.05	cultural culture

Notes: The table presents the label of each topic from the Comparative Agendas Project codebook, the number and proportion of the bills classified by the human coders of the Congressional Bills Project for each topic, and three most frequent keywords associated with each topic. Note that the *Culture* topic only has two keywords and the same keywords may appear for different topics.

The maximum document length is 152,624 and the minimum is 26.

These bills are ideal for our empirical evaluation because the CBP uses human coders to assign a primary policy topic that follows CAP to each bill, enabling us to validate the automated classification of topic models against the manual coding.[8] We derive keywords of each topic from the brief description provided by the CAP. We make this process as automatic as possible to reduce the subjectivity of our empirical validation (see SI, p. 4). SI (pp. 11–14) demonstrates that our empirical results shown below are robust to selection of different keywords. Table 1 presents the 21 CAP topics, the number and proportion of the bills assigned by the CBP human coders to each topic, and their most frequent keywords (see Table S1 in SI, p. 4, for the full list of keywords).

We fit keyATM and wLDA to this corpus. For both models, we use a total of $K = \widetilde{K} = 21$ topics and do not include any additional topics because the CAP topics are designed to encompass all possible issues in this corpus. This setting means that for keyATM, all topics have some keywords. We use the default prior specification of the keyATM package (see footnote). For wLDA, we use the exactly same implementation and specification as keyATM with the exception of using no keyword, that is, $\pi_k = 0$ for all k. We run five independent Markov chains with different random starting values obtained from the prior distribution of each model. We run the Markov chain Monte Carlo (MCMC) algorithms for 3,000 iterations and obtain the posterior means of ϕ^*_{kv} and θ_{dk} using Equations (8) and (9), respectively (SI, p. 14, analyzes the convergence).

Topic Interpretability. We begin by examining the interpretability of the resulting topics. We focus on the topic–word distributions and show that words with high probabilities given a topic are consistent with the topic's label. For keyATM, there is no need to label topics with prespecified keywords after model fitting. In contrast, wLDA requires the post hoc labeling of the resulting topics. Here, we determine the topic labels such that the document classification performance of wLDA is

[8]Master Codebook. The Policy Agendas Project at the University of Texas at Austin, 2019. Available at https://www.comparativeagendas.net/pages/master-codebook , last accessed on December 10, 2019.

TABLE 2 Comparison of Top 10 Words for Six Selected Topics between keyATM and wLDA

Labor		**Transportation**		**Foreign Trade**	
keyATM	wLDA	keyATM	wLDA	keyATM	wLDA
employee	apply	**transportation**	transportation	product*	air
benefit	tax	**highway**	highway	**trade**	vessel
individual	amendment	safety	safety	change	airport
rate	end	carrier	vehicle	**agreement**	transportation
compensation	taxable	**air**	carrier	good	aviation
period	respect	code*	motor	tobacco*	administrator
code*	period	system	system	head	aircraft
payment*	individual	vehicle*	strike	article	carrier
determine	case	**airport**	rail	free	administration
agreement*	relate	motor	code	chapter	coast
Immigration		**Law & Crime**		**Government Operations**	
keyATM	wLDA	keyATM	wLDA	keyATM	wLDA
security*	alien	intelligence*	security	expense	congress
alien	attorney	attorney	information	appropriation	house
immigration	child	**crime**	intelligence	remain	senate
homeland*	crime	**court**	homeland	authorize	office
border*	immigration	**enforcement**	committee	necessary	committee
status	grant	**criminal**	director	transfer*	commission
nationality	enforcement	**code**	system	expend	representative
describe	person	offense	foreign	exceed	congressional
individual	court	person	government	office	strike
employer*	offense	**justice**	office	activity	bill

Notes: The table shows the ten words with the highest estimated probability for each topic under each model. For keyATM, the pre-specified keywords for each topic appear in bold letters, whereas the asterisks indicate the keywords specified for another topic. Both models use term weights described in Sampling Algorithm.

maximized (see SI, p. 27). This leads to a less favorable empirical evaluation for keyATM. Below, we show that even in this setting, keyATM significantly outperforms wLDA.

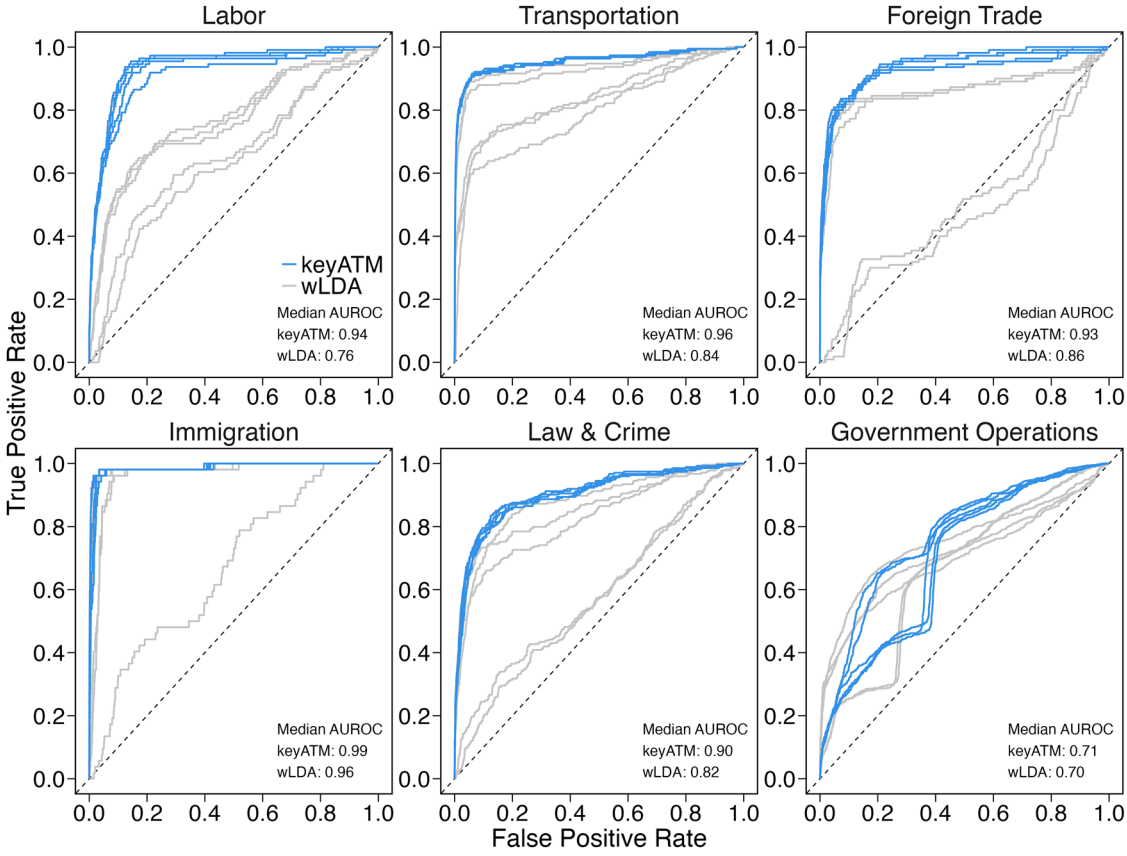
Table 2 presents 10 words with the highest estimated probabilities for six selected topics under each model (see Table S2 in SI, p. 5, for the remaining 15 topics). For keyATM, the keywords of each topic appear in bold letters whereas the asterisks indicate the keywords from another topic. Each model's result is based on the MCMC draws from one of the five chains that has the median performance in terms of the overall area under the receiver operating characteristics (AUROC).[9]

The results demonstrate several advantages of keyATM. First, the *Labor* topic of wLDA includes many unrelated terms and does not contain any terms related to this topic whereas keyATM lists many keywords among the most frequent words for the topic, such as "benefit," "employee," and "compensation." Second, wLDA does not find meaningful terms for the *Foreign trade* topic and instead creates two topics (labeled as *Transportation* and *Foreign trade*) whose most frequent terms are related to the *Transportation* topic. In contrast, the top words selected by keyATM represent the content of the *Foreign trade* topic, whereas those for the *Transportation* capture the substantive meaning of the topic well. As shown in the full top words table in SI (p. 5), wLDA fails to create topics whose top words contain terms related to *Labor* or *Foreign trade*.

Similarly, wLDA has difficulty in selecting the words that represent the *Law & crime* topic and cannot distinguish it from the *Immigration* topic. Indeed, the *Immigration* topic for wLDA includes the keywords of the *Law & crime* topic such as "crime," "court," and "enforcement." In contrast, keyATM selects many keywords

[9]For wLDA, it is difficult to combine multiple chains due to the label switching problem. There is no such problem for keyATM because the topics are labeled before model fitting.

FIGURE 1 Comparison of the ROC Curves between keyATM and wLDA for Six Selected Topics



Notes: Each line represents the ROC curve from one of the five Markov chains with different starting values for keyATM (blue lines) and wLDA (gray lines). The median AUROC indicates the median value of AUROC among five chains for each model. The plots show that keyATM has a better topic classification performance than wLDA with the exception of the "Government operations" topic. The results of keyATM are also less sensitive to the starting values.

among the top 10 words for each of these two topics without conflating them. This result is impressive because the bills whose primary topic is the *Immigration* topic account only for 1.18% of all bills. Thus, keyATM can measure *Law & crime* and *Immigration* as two distinct topics whereas wLDA fails to separate them. Finally, both keyATM and wLDA are unable to identify the meaningful words for the *Government operations* topic, which is the most frequent topic in our corpus. SI (p. 7) explains why both models fail to uncover this particular topic.

Topic Classification. Next, to evaluate the quality of topic–document distributions, we compare the automated classification of keyATM and wLDA with human coding. The proximity between estimated topic–document distributions and human coding implies better measurement. Specifically, we compare the estimated topic–document distribution, $\hat{\theta}_{dk}$ given in Equation (9), with the primary policy topic assigned by the CBP human coders. Although the topic models allow each document to belong to multiple topics, the CBP selects only one primary topic for each bill. Despite this difference, we independently evaluate the classification performance of keyATM against that of wLDA via the ROC curves based on $\hat{\theta}_{dk}$ for each topic k. As noted earlier, our evaluation setting favors wLDA because wLDA topics are matched with the CBP topics by maximizing its classification performance.

Figure 1 presents the ROC curves for the same six selected topics as those shown in Table 2 (see SI, p. 7, for the other topics). Each line represents the ROC curve based on one of the five Markov chains with different starting values for keyATM (blue lines) and wLDA (gray lines). We find that keyATM outperforms wLDA except for the *Government operations* topic. The results are consistent with the qualitative evaluation based on Table 2. For example, the poor performance of both models for the *Government operations* is not surprising given that their

top words are not informative about the topic content (SI, p. 7, explains this underperformance). When compared to wLDA, keyATM has a much better classification performance for the *Labor*, *Transportation*, *Foreign trade*, and *Law & crime* topics, where its topic interpretability is superior. Finally, the ROC curves for keyATM are less sensitive to different starting values than those for wLDA with the exception of the *Government operations* topic.

The Covariate keyATM

Next, we extend the base keyATM by incorporating covariates for the document–topic distribution. The inclusion of covariates is useful as social scientists often have meta-information about documents (e.g., authorship). We adopt the Dirichlet-Multinomial regression framework of Mimno and McCallum (2008) rather than the logistic normal regression approach of the STM in Roberts, Stewart, and Airoldi (2016) so that the collapsed Gibbs sampler strategy for the base keyATM can be used.

Model

Suppose that we have an M-dimensional covariate $\mathbf{x}_d$ (including an intercept) for each document d. We model the document–topic distribution using these covariates in the following fashion (in place of Equations (2) and (3)),

$$\boldsymbol{\theta}_d \overset{\text{indep.}}{\sim} \mathsf{Dirichlet}(\exp(\boldsymbol{\lambda}^\top \mathbf{x}_d)),$$
$$\lambda_{mk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2),$$

for each $d = 1, 2, \ldots, D$, where $\boldsymbol{\lambda}$ is an $M \times K$ matrix of coefficients and λ_{mk} is the (m, k) element of $\boldsymbol{\lambda}$. The sampling algorithm and the model interpretation are straightforward extensions of the base keyATM (SI, p. 1).[10]

Empirical Evaluation

We evaluate the empirical performance of the covariate keyATM against that of STM using the Japanese election manifesto data (Catalinac 2016). Analyzing manifestos of Japan's Liberal Democratic Party (LDP) candidates, the author finds that the 1994 electoral reform is associated with a relative increase in the topics about programmatic policies and a decline in the topics about pork barrel. Because the manifestos come from eight elections and the author focuses on LDP candidates, we include the election-year and LDP dummies as the covariates. We find that the covariate keyATM yields more interpretable topics and its results are less sensitive to the total number of topics chosen by researchers than STM.

Data and Setup. We analyze a total of 7,497 manifestos (Shinada 2006). Catalinac (2016) preprocessed the data by tokenizing Japanese sentences, removed punctuations and stop words, and cleaned up the documents based on an author-defined dictionary. We use the document–term matrix from the original study so that the preprocessing steps remain identical. In Japanese elections, every registered political candidate is given a fixed amount of space in a government publication, in which their manifesto can be printed. This document, containing the manifestos of all candidates, is then distributed to all registered voters. After preprocessing, the average number of words is about 177 (the maximum is 543 and the minimum is 4), whereas the number of unique terms is 2,832.

These manifestos cover 3,303 unique candidates who ran in the eight consecutive elections held between 1986 and 2009. Because Japanese electoral campaigns are heavily restricted, the manifestos represent one of the few ways in which candidates communicate their policy goals to voters.

Keyword Construction. Unlike the validation study presented in the "Empirical Evaluation" section of the base keyATM, we do not have human-coded topics for this data set. In the original article, the author applies LDA with 69 topics and labels all the topics after fitting the model by carefully examining the 15 most frequent words for each topic. The author, then, discusses how the estimated topic proportions change after the 1994 electoral reform. To apply the covariate keyATM, we must develop a set of keywords for topics. Unfortunately, we cannot merely use the most frequent words identified by Catalinac (2016) with LDA as the keywords because that would imply analyzing the same data as the one used to derive keywords.

To address this problem, we independently construct keywords using the questionnaires of the UTokyo-Asahi Surveys (UTAS), which is a collaborative project between the University of Tokyo and the Asahi Shimbun, a major national newspaper. Table 3 presents the resulting 16 topics and their keywords (two pork barrel and 14

[10]We do not directly model the correlation across topics. Although we have explored alternative modeling strategies including the Logistic-Normal approach of Roberts, Stewart, and Airoldi (2016), we find that the proposed models generally perform well without directly modeling the correlation structure.

TABLE 3 Keywords for Each Topic

Type	Topic Label	Keywords
Pork barrel	Public works	employment, public, works
	Road construction	road, budget
Programmatic	Regional devolution	rural area, devolve, merger
	Tax	consumption, tax, tax increase
	Economic recovery	economic climate, measure, fiscal policy, deficit
	Global economy	trade, investment, industry
	Alternation of government	government, alternation
	Constitution	constitution
	Party	party, political party
	Postal privatization	postal, privatize
	Inclusive society	women, participate, civilian
	Social welfare	society, welfare
	Pension	pension
	Education	education
	Environment	environment, protection
	Security	defense, foreign policy, self defense

Notes: The left and middle columns show the types of policies and topic labels assigned by Catalinac (2016). The corresponding keywords in the right column are obtained from the UTokyo-Asahi Surveys (UTAS). This results in the removal of five policy areas (sightseeing, regional revitalization, policy vision, political position, and investing more on human capital) that do not appear in the UTAS.

programmatic policy topics). Because most UTAS questions consist of a single sentence, we typically choose nouns that represent each topic's substantive meanings (SI, p. 19, explains details of keyword construction).

Finally, we fit the covariate keyATM and STM, using seven election-year indicator variables and another indicator for the LDP candidates. We examine the degree to which the results are sensitive to model specification by varying the total number of topics. Specifically, in addition to the 16 keyword topics, we include different numbers of extra topics with no keyword. We try 0, 5, 10, and 15 no-keyword topics. We fit keyATM for 3,000 iterations with a thinning of 10 and the default hyperparameter values. We use the default settings of the STM package.

Topic Interpretability. Table 4 lists the 10 most frequent words for each of the six selected topics. These topics are chosen because they are easier to understand without the knowledge of Japanese politics (see SI, p. 21, for the results of the other topics). We match each topic of STM with that of keyATM by applying the Hungarian algorithm to the estimated topic–word distributions so that the overall similarity between the results of the two models is maximized.

We find that the covariate keyATM produces more interpretable topics, judged by these 10 most frequent words, than STM. For example, keyATM identifies, for the *Road construction* topic, the terms such as "development," "construction," and "track" as well as two assigned keywords, "road" and "budget." This makes sense given that developing infrastructures such as road and railway tracks is considered as one of the most popular pork barrel policies in Japan. For the *Education* topic, STM does not include "education," whereas keyATM includes two selected keywords, "children" and "education." Finally, for the *Security* topic, keyATM lists terms such as "peace," "safe," and "international," whereas the terms selected by STM broadly cover international economy and politics.

Topic Discrimination. Good topic models should yield topics distinct from one another, which means that we would like different topics to have different words representing them. The bar plots in Figure 2 present the number of times that each of the 38 keywords appears in the top 10 (left panel) or 15 (right panel) words of keyword topics. As expected, the covariate keyATM (blue shaded bars) assigns the same keywords to fewer topics than STM (gray bars). In particular, more keywords appear as the most frequent terms only for a single topic under keyATM than under STM.

Covariate Effects. One key hypothesis of Catalinac (2016) is that after the 1994 electoral reform, LDP candidates adopted electoral strategies to pursue more programmatic policies. The author tests this hypothesis

TABLE 4 Comparison of Top 10 Words for Six Selected Topics between the Covariate keyATM and STM

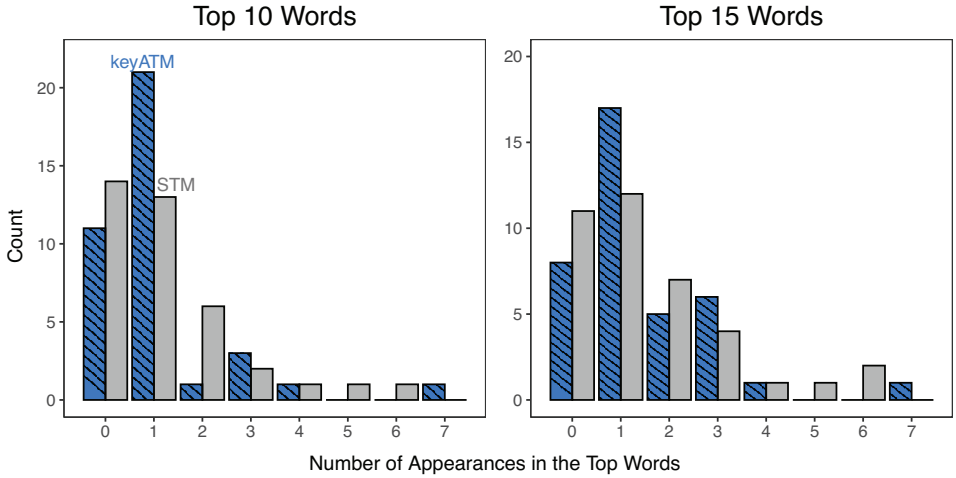
Road Construction		Tax		Economic Recovery	
keyATM	STM	keyATM	STM	keyATM	STM
development	tax	Japan	Japan	reform	reform
road	reduce tax	**tax**	citizen	**measure**	postal
city	yen	citizen	JCP	society∗	privatize
construction	housing	JCP	politic	Japan	Japan
tracks	realize	**consumption**	tax	**economic climate**	rural area
budget	daily life	politic	consumption	reassure	country
realize	move forward	**tax increase**	tax increase	economy	citizen
promote	city	oppose	oppose	institution	safe
move forward	education	business	business	safe	government
early	measure	protect	protect	support	pension
Inclusive Society		**Education**		**Security**	
keyATM	STM	keyATM	STM	keyATM	STM
politic	politic	politic	Japan	Japan	society
civilian	reform	Japan	person	**foreign policy**	Japan
society∗	new	person	country	peace	world
participate	realize	**children**	politic	world	economy
peace	citizen	**education**	necessary	economy	environment
welfare∗	government	country	problem	country	international
aim	daily life	make	children	citizen	education
human rights	rural area	force	force	**defense**	country
realize	corruption	have	have	safe	peace
consumption∗	change	problem	future	international	aim

Notes: The table shows the ten words with the highest estimated probabilities for each topic under each model. For keyATM, the prespecified keywords for each topic appear in bold letters whereas the asterisks indicate the keywords specified for another topic.

by plotting the estimated topic proportions for each election year. Here, we take advantage of the fact that the covariate keyATM and STM can directly incorporate covariates. The quantities of interest are the election-year proportions of the pork barrel and programmatic topics for LDP politicians. Specifically, we first compute, for each topic, the posterior mean of document–topic probability for LDP manifestos within each election year by using Equation (S2) is in SI (p. 2) with the appropriate values of covariates. We then compute the sum of these posterior mean proportions for each policy type as an estimate of the total proportion. In addition, we examine the sensitivity of these results to the choice of the total number of no-keyword topics for both keyATM and STM.

Figure 3 plots the sum of estimated topic proportions corresponding to pork barrel (blue) and programmatic policies (red), for the LDP candidates. All topics in STM are referred to as no-keyword topics because they do not have preassigned keywords. The total number of topics is the same between two models. The plot omits credible intervals because they are too narrow to be visible. Consistent with the original analysis, we generally find that in the first election after the 1994 electoral reform, the proportion of programmatic policy topics substantially increased whereas the proportion of pork barrel topics remain virtually unchanged. For the covariate keyATM, this finding is consistent across all model specifications except the model without no-keyword topics (dotted lines with solid diamonds). This model without any no-keyword topic is not credible in this application because these keywords taken from UTAS do not cover the entire contents of manifestos. In contrast, the performance of STM is much more sensitive to the total number of topics. The change after the electoral reform is also less stark when compared to the covariate keyATM. In sum, the covariate keyATM yields more reasonable and robust results than STM.

FIGURE 2 Exclusivity of Keywords across Topics



Notes: Left (right) bar plot shows the number of times that each of the 38 keywords appears as the 10 most frequent (15 most frequent) words of keyword topics. These words are less likely to be shared across topics under the covariate keyATM (blue shaded bars) than under the STM (gray bars).

The Dynamic keyATM

The second extension we consider is the dynamic modeling of document–topic distributions. Researchers are often interested in investigating how the prevalence of topics changes over time (Clarke and Kocak 2020). Building on the collapsed Gibbs sampler used for the base keyATM, we apply the Hidden Markov Model (HMM). HMM has been used to introduce time dynamic components in various applications (e.g., Quinn et al. 2010; Park 2012; Knox and Lucas 2021; Olivella, Pratt, and Imai 2022, see SI, p. 22, for details). An alternative modeling strategy is to use time fixed effects either in the covariate keyATM or in a regression model fitted to the output from the base keyATM. Unlike these approaches, the dynamic keyATM incorporates time ordering and smoothly models time trend while properly accounting for uncertainty.

Model

Suppose that we have a total of T time periods and each document d belongs to one of these time periods, $t[d] \in \{1, 2, \ldots, T\}$. The HMM is based on the idea that each time period belongs to one of the latent discrete states. Assume that we have a total of R such states and use $h_t \in \{1, 2, \ldots, R\}$ to denote the latent state for time t. Following Chib (1998), we only allow for one-step forward transition, implying that the probability of transition from state r to state r', that is, $p_{rr'} = \Pr(h_{t+1} = r' \mid h_t = r)$, is equal to zero unless $r' = r + 1$. This assumption considerably simplifies the estimation without sacrificing model fit so long as we have a sufficiently large number of states. The resulting Markov transition probability matrix is given by,

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & 0 & \cdots & 0 & 0 \\ 0 & p_{22} & p_{23} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p_{R-1,R-1} & p_{R-1,R} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

The prior distribution for the probability of no transition is uniform, that is, $p_{rr} \overset{\text{i.i.d.}}{\sim} \mathsf{Uniform}(0, 1)$ for $r = 1, 2, \ldots, R$. Finally, the dynamic keyATM allows the topic proportion $\boldsymbol{\theta}_d$ to evolve over time by letting $\boldsymbol{\alpha}$ to vary across the latent states. Modeling $\boldsymbol{\alpha}$ instead of $\boldsymbol{\theta}_d$ makes keyATM less sensitive to the short-term temporal variation. Thus, instead of Equation (3) we have,

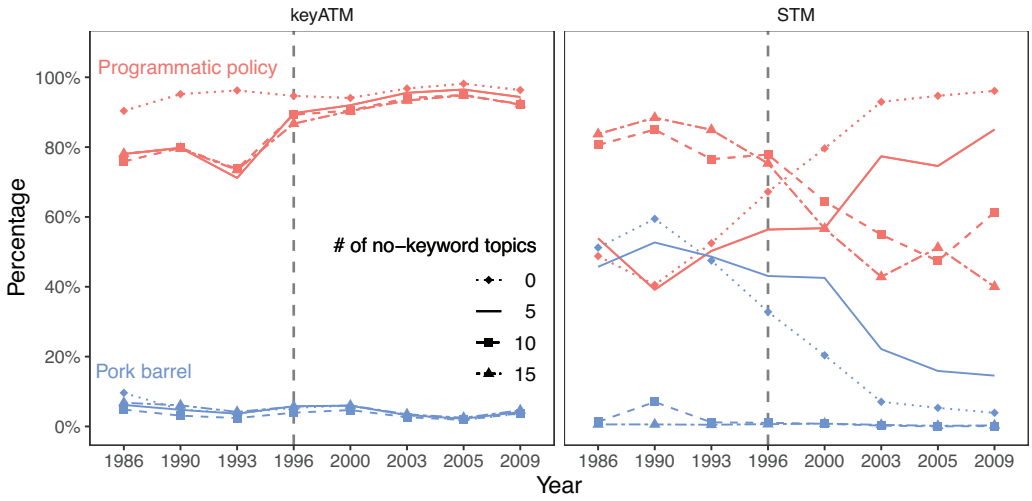
$$\alpha_{rk} \overset{\text{i.i.d.}}{\sim} \mathsf{Gamma}(\eta_1, \eta_2) \quad \text{for } r = 1, 2, \ldots, R \text{ and } k = 1, 2, \ldots, K. \tag{10}$$

The sampling algorithm and the model interpretation are straightforward extension of the base keyATM (SI, p. 2).

Empirical Evaluation

In this section, we empirically evaluate the performance of the dynamic keyATM by analyzing the corpus of the United States Supreme Court opinions from the

FIGURE 3 Programmatic Policy Topics Increase Right after the 1994 Electoral Reform



Notes: The results based on the covariate keyATM (left panel) shows that the estimated proportion of programmatic policy topics increased in the 1996 election right after the election reform. The results are not sensitive to the number of topics except when there is no additional no-keyword topic. Note that all topics in STM are referred to as no-keyword topics because they do not have preassigned keywords. The total number of topics is the same between two models. The results based on STM vary substantially across different numbers of topics.

Supreme Court Database (SCD) project.[11] Like the Congressional bill data set analyzed in the "Empirical Evaluation" section of the base keyATM, the key advantage of this data set is that human coders have identified the primary topic of each opinion and each topic comes with keywords. The only difference between the dynamic keyATM and wLDA is the existence of keyword topics. We show that the dynamic keyATM yields more interpretable topics and better classification performance than the dynamic wLDA. Moreover, the time trend of topic prevalence estimated with keyATM is closer to the human coding than that of wLDA without keywords.

Data and Setup. We analyze a total of 17,245 Supreme Court opinions written between 1946 and 2012, with an average of 265 opinions per year (Rice 2017).[12]

We preprocess these texts using the same procedure used in the "Data and Setup" section of the base keyATM, yielding a corpus with an average of 1,298 words per document and a total of 9,608 unique words. The maximum number of words for a document is 30,767, whereas the minimum is 1.

The SCD project used human coders to identify the primary issue area for each opinion.[13] According to the project website, there are a total of 278 issues across 14 broader categories. We use the aggregate 14 categories as our keyword topics, that is, $\widetilde{K} = 14$. We obtain the keywords from the issue descriptions provided by the SCD project. We apply the same preprocessing procedure used in the "Data and Setup" section of the base keyATM. SI (p. 23) provides further details about the keyword construction. Table 5 presents these 14 topics from the SCD project, the number and proportion of the opinions classified to each topic by the SCD human coders, and their five most frequent keywords (see Table D.3 in SI, p. 23, for the full list of keywords).

We fit the dynamic keyATM and wLDA to this corpus. Because the SCD topic categories are supposed to be comprehensive, we do not include any additional topics that do not have keywords, that is, $K = \widetilde{K} = 14$. Therefore, all topics have some keywords for keyATM. For the HMM specification, we use a total of five states, that is, $R = 5$, because five states performed the best in terms of the commonly used perplexity measure. For the hyperparameters, we use the default values provided by the keyATM package. Finally, the implementation for the dynamic wLDA is identical to that of the dynamic

[11] http://scdb.wustl.edu/, accessed December 10, 2019.

[12] The authors thank Douglas Rice for generously sharing the text data of the Supreme Court opinions.

[13] Scholars who study judicial politics have used this issue code (Rice 2017).

TABLE 5 Frequency of Each Topic and Its Most Common Keywords

Topic Label	Count	Percentage	Most Frequent Keywords
Criminal procedure	4268	24.75	right rule trial evidence justice
Economic activity	3062	17.76	federal right claim evidence power
Civil rights	2855	16.56	right public provision party constitutional
Judicial power	1964	11.39	federal right district rule claim
First amendment	1795	10.41	amendment first public party employee
Due process	738	4.28	right defendant constitutional employee process
Federalism	720	4.18	federal tax regulation property support
Unions	664	3.85	right employee standard union member
Federal taxation	529	3.07	federal claim provision tax business
Privacy	290	1.68	right regulation information freedom privacy
Attorneys	188	1.09	employee attorney official bar speech
Interstate relations	119	0.69	property interstate dispute foreign conflict
Miscellaneous	50	0.29	congress authority legislative executive veto
Private action	3	0.02	evidence property procedure contract civil

Notes: The table presents the label of each topic from the Supreme Court Database (SCD) codebook, the number and proportion of the opinions assigned to each topic by the SCD human coders, and five most frequent keywords associated with each topic. Note that the same keywords may appear for different topics.

keyATM with the exception of setting $\pi = 0$ (i.e., no keyword).

As in the "Empirical Evaluation" section of the base keyATM, we run five independent Markov chains for 3,000 iterations for each model with different starting values independently sampled from the prior distribution. We compute the posterior means of ϕ^*_{kv} and θ_{dk} using Equations (8) and S4. After fitting the models, we match the resulting topics from the dynamic wLDA with the SCD topics by maximizing its classification performance (see the "Topic Interpretability" section of the base keyATM). There is no need to apply this procedure to the dynamic keyATM because the topic labels are determined when specifying keywords before fitting the model. Thus, our empirical evaluation provides the least (most) favorable setting for the dynamic keyATM (wLDA).

Topic Interpretability. We first compare the interpretability of the topics obtained from the dynamic keyATM and wLDA. Table 6 presents the 10 words with the highest estimated probabilities defined in Equation (7) for selected six topics (see Table S11 in the SI, p. 23, for the remaining eight topics). For the dynamic keyATM, the prespecified keywords appear in bold letters whereas the asterisks indicate the keywords specified for another topic. The results for each model are based on the MCMC draws from one of the five chains that has the median performance in terms of the overall AUROC.

We find the resulting topics of the dynamic keyATM are at least as interpretable as those discovered by the dynamic wLDA. For example, the top 10 words selected by keyATM for the *First amendment* topic contains the relevant keywords such as "first," "amendment," "speech," and "religious." In addition, keyATM can collect substantively meaningful terms even when only a small number of keywords appear in top frequent words. For example, for keyATM, only one of the 19 keywords, "tax," appears in the list of the top 10 words for the *Federal taxation* topic. And yet, the other words on the list, such as "income" and "pay," are highly representative of the substantive meaning of the topic. Finally, both keyATM and wLDA fail to identify the meaningful terms for the *Privacy* topic. SI (p. 26) shows that this is because the keywords for the *Privacy* topic do not frequently appear in the opinions assigned to this topic by the SCD project.

Topic Classification. Next, we compare the classification performance of the dynamic keyATM and wLDA with the human coding from the SCD project. We apply the same procedure as the one used in the "Topic Classification" section of the base keyATM and compute the ROC curve and AUROC based on the estimated topic–document distribution, $\hat{\boldsymbol{\theta}}_d$, given in Equation (S4). As mentioned earlier, the results are more favorable to wLDA because we match its topics with the SCD topics by maximizing the AUROC of the wLDA.

Figure 4 presents the ROC curves for the same six selected topics as those shown in Table 6 (see SI, p. 24, for the other topics). Each line represents the ROC curve based on one of the five Markov chains for the dynamic keyATM (blue lines) and wLDA (gray lines)

TABLE 6 Comparison of 10 Top Words for Selected Six Topics between the Dynamic keyATM and Dynamic wLDA

Criminal Procedure		First Amendment		Unions	
keyATM	wLDA	keyATM	wLDA	keyATM	wLDA
trial	trial	**public**	public	**employee**	employee
jury	jury	**amendment**	first	**union**	union
defendant*	petitioner	**first**	speech	board	labor
evidence	evidence	government	amendment	**labor**	employer
criminal	defendant	may	interest	**employer**	board
sentence	counsel	interest	party	agreement	agreement
petitioner	right	**speech**	may	employment*	contract
judge	rule	right*	right	contract*	employment
conviction	make	can	political	**work**	bargaining
counsel	judge	**religious**	government	**bargaining**	work
Federal Taxation		**Civil Rights**		**Privacy**	
keyATM	wLDA	keyATM	wLDA	keyATM	wLDA
tax	tax	district*	school	search*	child
property*	property	**school**	district	officer	benefit
pay	income	**discrimination**	religious	police	interest
income	pay	election*	discrimination	amendment*	medical
payment	bank	**equal**	county	arrest	plan
interest	interest	county	election	warrant	provide
benefit*	corporation	vote	vote	fourth	parent
amount	payment	**plan**	equal	evidence*	woman
plan*	amount	one	education	person	may
fund*	business	race	student	use	statute

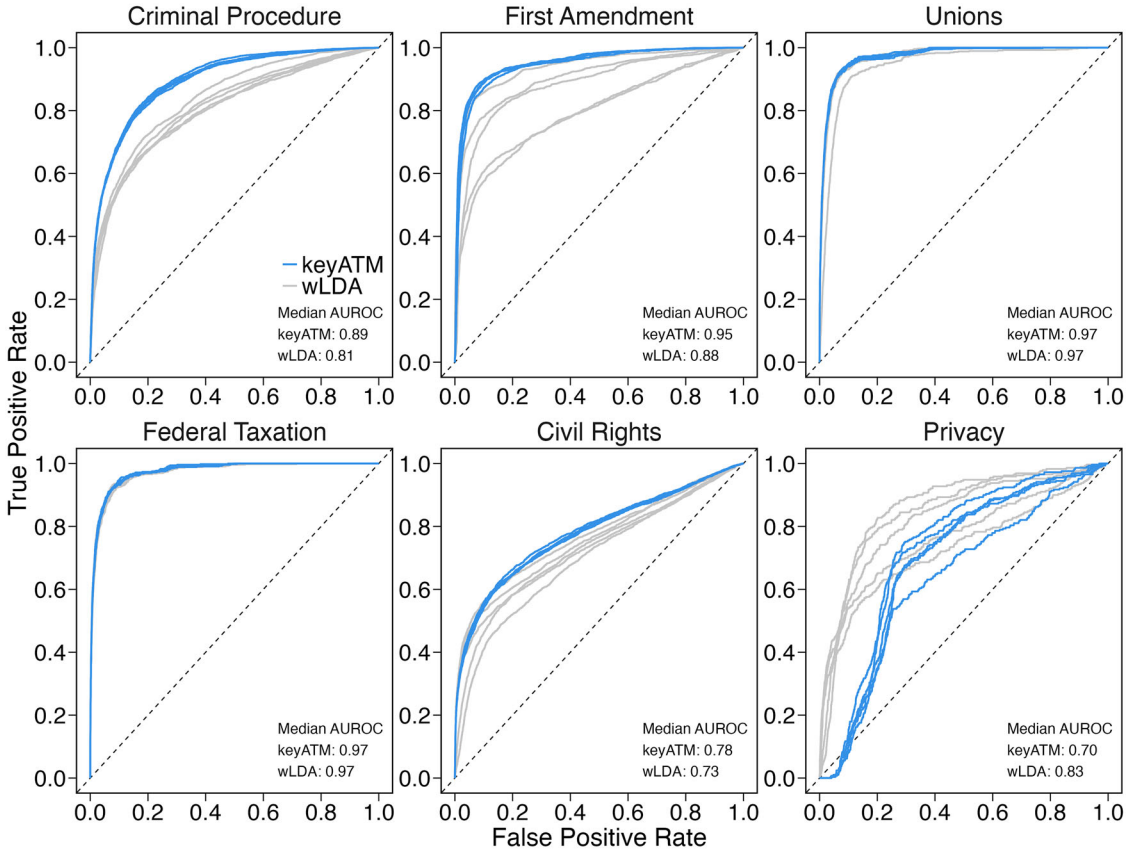
Notes: The table shows the ten words with the highest estimated probabilities for each topic under each model. For the dynamic keyATM, the pre-specified keywords for each topic appear in bold letters, whereas the asterisks indicate the keywords specified for another topic.

whereas the AUROC value is based on the chain with the median performance. We find that keyATM outperforms wLDA except for the *Privacy* topic. Recall that for this topic, the top words identified by both models have little substantive relevance. Lastly, the ROC curves for keyATM are in general less sensitive to different starting values than those for wLDA, again, except for the *Privacy* topic.

Time Trends of Topic Prevalence. Finally, we compare the time trend of estimated topic prevalence between each of the two topic models and the SCD human coding. We first calculate the mean proportion of each topic by using all documents that belong to each time point (see SI, p. 3, for details). For the SCD human coding, we compute the proportions of documents that are assigned to the topic of interest in each year. Note that the topic models assign multiple topics to each document, whereas the SCD coding classifies each document only to one of the 14 topics. As a result, these two proportions are not directly comparable. Therefore, we use the standardized measure to focus on relative time trends for comparison (i.e., subtract its mean from each data point and then divide it by its standard deviation).

Table 7 presents the correlation of the estimated topic prevalence between each of the two topic models and the SCD human coding (see SI, p. 25, for the other topics). We find that keyATM exhibits a higher correlation with the human coding for most topics than wLDA. For the *Privacy* topic, both keyATM and wLDA only weakly correlate with the human coding. This result is not surprising given the poor performance of keyATM for this topic in terms of both topic interpretability and classification (see SI, p. 26).

FIGURE 4 Comparison of the ROC Curves between the Dynamic keyATM and Dynamic wLDA for Six Selected Topics



Notes: Each line represents the ROC curve from one of the five Markov chains with different starting values for the dynamic keyATM (blue lines) and wLDA (gray lines). The plots show that keyATM has a better topic classification performance than wLDA with the exception of the *Privacy* topic. The median AUROC indicates the median value of AUROC among five chains for each model. The results of keyATM are also less sensitive to the starting values.

Crowdsource Validation

We conduct additional validation through crowdsourcing regarding the results about the superior topic interpretability of keyATM over wLDA. We find that keyATM generally improves the interpretability of estimated topics and their correspondence with labels without sacrificing topic coherency.

The Validation Methodology

We follow the validation framework proposed by Chang et al. (2009) and Ying, Montgomery, and Stewart (2022) to evaluate the resulting topics. First, we measure the semantic coherency of each topic via Random 4 Word Set Intrusion (R4WSI, coherency task; Ying, Montgomery, and Stewart 2022). A worker sees four different word sets (each word set consists of four words). Three word sets are randomly selected from the top words of one topic whereas the other set is randomly selected from those of a different topic.[14] The worker is then asked to identify one word set that is the most unrelated to other three.

Second, we measure the consistency between human labels and topics through the modified R4WSI coherency-and-label task (Ying, Montgomery, and Stewart 2022) where we also show each worker a label from the topic used to generate the three word sets. We then ask them to choose an unrelated word set from the four word sets. Using all three empirical applications, we apply both methods to keyATM and its baseline counterpart and compare their performance (see SI, p. 27, for details). We do not include Word Intrusion (WI) and

[14]Each word is generated from the top 20 words of a topic based on their posterior probability.

TABLE 7 Comparison of the Time Trends of Topic Prevalence between the Dynamic keyATM / wLDA and the SCD Human Coding for Six Selected Topics

Topic	Dynamic keyATM	Dynamic wLDA
Criminal procedure	0.83	0.07
First amendment	0.82	0.64
Unions	0.78	0.74
Federal taxation	0.80	0.79
Civil rights	0.76	0.70
Privacy	0.07	0.18

Notes: The table shows the correlation between the estimated topic prevalence in Equation S5 and the SCD human coding. To compare the estimated topic prevalence and the SCD human coding, we use the standardized measure that subtracts its mean from each data point and then divide it by its standard deviation. The results show that keyATM exhibits a higher correlation with the human coding for most topics than the dynamic wLDA.

Top 8 Word Set Intrusion (T8WSI) in our validation exercise because, as Ying, Montgomery, and Stewart (2022) note, both are difficult tasks, and the latter is particularly known to be sensitive to the choice of displayed words.[15]

We recruit English-speaking workers from Amazon Mechanical Turk for the base and dynamic applications and Japanese-speaking workers from CrowedWorks for the covariate application, between December 2020 and January 2021. Recruited workers are directed to a Qualtrics survey that we designed for this validation. Workers who agree to participate in this task are asked to complete 11 tasks of the same kind from a single empirical application: five tasks for keyATM, five tasks for the baseline model, and one gold-standard task, which is similar to the other tasks but is much less ambiguous.[16] We use the gold-standard task to assess worker's attention and remove the responses of any worker who fails to provide a correct answer.

We use Bayesian hierarchical logistic regressions to analyze the validation data for each application.[17] Because the number of observations is small for any given topic, partial pooling helps improve the precision of estimates.[18] The outcome is an indicator variable Y_i, which equals one if a worker correctly answers the task. Our predictor is the indicator variable X_i, which equals one if the task is based on the outputs of keyATM. We also include worker-specific and topic-specific random intercepts as well as random coefficients for each topic to account for topic heterogeneity. Formally, for each task conducted by worker j, each response i with topic k is modeled as,

$$\Pr(Y_i = 1) = \text{logit}^{-1}\big((\beta_0 + \delta_{0k[i]} + \gamma_{j[i]}) + (\beta_1 + \delta_{1k[i]})X_i\big),$$

where $(\delta_{0k}, \delta_{1k})^\top \sim \mathcal{N}(0, \mathbf{\Sigma})$ and $\gamma_j \sim \mathcal{N}(0, \sigma)$. Our quantity of interest is the difference in the predicted probabilities of correct responses between keyATM and its baseline counterpart.

Findings

Figure 5 presents the main validation results. We plot the differences in the predicted probabilities based on Bayesian hierarchical logistic regressions for pooled effects and for six topics shown in the previous sections (see SI, p. 30, for the other topics). First, the pooled estimates show that the base keyATM performs better in both tasks than wLDA. The differences in predicted probabilities are both positive and statistically significant at 95% level. Although the topic-specific estimates are less precise, the base keyATM performs well in the same topics (e.g., *Law & crime*) as those identified in section "Topic Interpretability" section of the base keyATM. On the other hand, the model performs poorly for the *Government operations*, which is also consistent with the result shown earlier.

Second, the covariate keyATM performs better in the coherency-and-label task than STM, whereas their performance is similar in terms of topic coherency. The estimated topics from keyATM match well with the preassigned labels. Finally, there is no statistically meaningful difference between the performance of the dynamic keyATM and wLDA. This may be because the court opinions involve technical vocabularies and complex concepts and hence they pose significant evaluation challenges to crowdsource workers (Ying, Montgomery, and Stewart 2022). For example, connecting the label *Civil rights* to such words as "district," "school," and "discrimination" requires some background knowledge.

Keyword Selection

The performance of keyATM critically depends on the quality of keywords. This section briefly discusses how one might select keywords for keyATM. In section

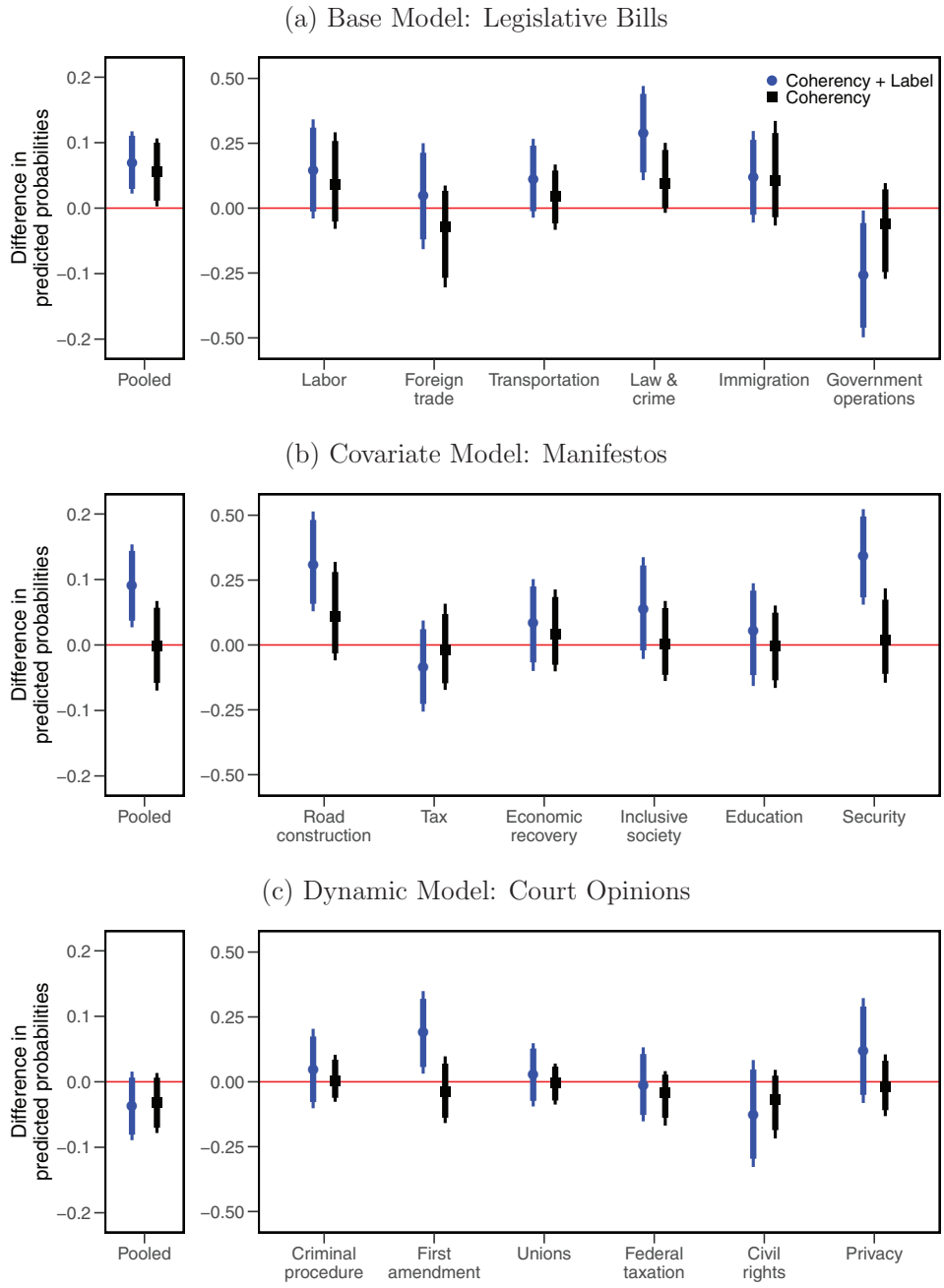
[15] SI (p. 28) shows results from an alternative design.

[16] Some workers participate in two sets of 11 tasks with two different methods.

[17] SI (p. 30) provides descriptive statistics.

[18] In the case of base models, for example, the average number of observations for each topic is 25 whereas the total number of observations is 545 for each task.

FIGURE 5 Comparison of the Performance of Validation Results between keyATM and Its Baseline Counterpart



Notes: Each point represents the difference in the predicted probabilities. The thick and thin vertical lines indicate the 95% and 90% credible intervals, respectively. Black and blue lines show the results from the coherency task (R4WSI) and the coherency-and-label task (modified R4WSI), respectively. The lines in the left panels represent the pooled results whereas the lines in the right panels represent the results from six topics shown in the previous sections.

"Keyword Construction" section of the covariate keyATM, we have shown how to construct keywords in the context of specific empirical application. In general, keyATM results in poor topic interpretability and classification when selected keywords do not frequently occur in one's corpus or do not discriminate their topics from others (see Figure S3 in the SI, p.10). We also conduct an experiment to examine the performance of keyATM by randomly removing some keywords. Our analysis shows that, although keyATM outperforms wLDA even when

some keywords are removed, the choice of keyword matters when the number of keywords is extremely small. For example, the *Immigration* topic in the first application, which only contains three keywords, performs poorly when the word "immigration" is removed from the keyword set. Thus, researchers should examine the relative frequency of candidate keywords and choose the ones that substantively match with the topics of interest. See SI (p. 14) for the full set of results.

Given the importance of keyword selection, future research should study how to choose good keywords. For example, King, Lam, and Roberts (2017) show the promising performance of automated keyword selection algorithms, whereas Watanabe and Zhou (2022) propose a dictionary-making procedure based on the average frequency entropy.

Concluding Remarks

Social scientists have utilized fully automated content analysis based on probabilistic topic models to measure a variety of concepts from textual data. To improve the quality of measurement, we propose the keyATM, which require researchers to label topics of interest before fitting a model. We have empirically shown that providing standard models with a small number of keywords can substantially improve the interpretability and classification performance of the resulting topics. There are several potential extensions of keyATM. For example, researchers can incorporate a small amount of human-coded labels assigned to each document to further improve the performance of keyATM. If such labels are based on single topics (as is the case for the CBP and SCD), this approach yields a mixed-membership version of document labels so that each document can belong to multiple topics.

References

Andrzejewski, David, Xiaojin Zhu, and Mark Craven. 2009. "Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors." In Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09). Association for Computing Machinery, pp. 25–32.

Bagozzi, Benjamin E., and Daniel Berliner. 2018. "The Politics of Scrutiny in Human Rights Monitoring: Evidence from Structural Topic Models of US State Department Human Rights Reports." *Political Science Research and Methods* 6(4): 661–77.

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.

Bauer, Martin W. 2000. "Classical Content Analysis: A Review." In Martin W. Bauer and George Gaskell (Eds.), *Qualitative Researching with Text, Image and Sound: A Practical Handbook for Social Research*. London: SAGE Publications. 132–51.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "quanteda: An R package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3(30): 774.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media.

Bischof, Jonathan M., and Edoardo M. Airoldi. 2012. "Summarizing Topical Content with Word Frequency and Exclusivity." In Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML'12). 9–16.

Blaydes, Lisa, Justin Grimmer, and Alison McQueen. 2018. "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds." *Journal of Politics* 80(4): 1150–67.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4): 77–84.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation. " *Journal of Machine Learning Research* 3: 993–1022.

Boyd-Graber, Jordan, David Mimno, and David Newman. 2014. "Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements." In E. M. Airoldi, D. Blei, E. A. Erosheva, and S. E. Fienberg (Eds.), *Handbook of Mixed Membership Models and Their Applications*. Boca Raton, FL: CRC Press, 3–41.

Catalinac, Amy. 2016. "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections." *Journal of Politics* 78(1): 1–18.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS '09). 288–96.

Chemudugunta, Chaitanya, America Holloway, Padhraic Smyth, and Mark Steyvers. 2008. "Modeling Documents by Combining Semantic Concepts with Unsupervised Statistical Learning." In Proceedings of the 7th International Conference on The Semantic Web (ISWC '08). Springer-Verlag, Berlin, Heidelberg, 229–44.

Chib, Siddhartha. 1998. "Estimation and Comparison of Multiple Change-Point Models." *Journal of Econometrics* 86(2): 221–41.

Clarke, Killian, and Korhan Kocak. 2020. "Launching Revolution: Social Media and the Egyptian Uprising's First Movers." *British Journal of Political Science* 50(3): 1025–45.

Dietrich, Bryce J., Matthew Hayes, and Diana Z. O'brien. 2019. "Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech." *American Political Science Review* 113(4): 941–62.

Fan, Angela, Finale Doshi-Velez, and Luke Miratrix. 2019. "Assessing Topic Model Relevance: Evaluation and Informative Priors." *Statistical Analysis and Data Mining* 12(3): 210–22.

Griffiths, Thomas L., and Mark Steyvers. 2004. "Finding Scientific Topics. " *Proceedings of the National Academy of Sciences of the United States of America* 101: 5528–5235.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1): 1–35.

Grimmer, Justin. 2013. "Appropriators Not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation." *American Journal of Political Science* 57(3): 624–42.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3): 267–97.

Hansen, Joshua A., Eric K. Ringger, and Kevin D. Seppi. 2013. "Probabilistic Explicit Topic Modeling Using Wikipedia." In: Gurevych, I., Biemann, C., Zesch, T. (eds) Language Processing and Knowledge in the Web. Springer, Berlin, Heidelberg. 69–82.

Hu, Yuening, Jordan Boyd-Graber, and Brianna Satinoff. 2011. "Interactive Topic Modeling." In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA. Association for Computational Linguistics. 248–57.

Jagarlamudi, Jagadeesh, Hal Daumé III, and Raghavendra Udupa. 2012. "Incorporating Lexical Priors into Topic Models." In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France. Association for Computational Linguistics. 204–13.

King, Gary, Patrick Lam, and Margaret E. Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61(4): 971–88.

Knox, Dean, and Christopher Lucas. 2021. "A Dynamic Model of Speech for the Social Sciences." *American Political Science Review* 15(2): 649–66.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(2): 311–31.

Li, Chenliang, Shiqian Chen, Jian Xing, Aixin Sun, and Zongyang Ma. 2019. "Seed-Guided Topic Model for Document Filtering and Classification." *ACM Transactions on Information Systems* 37(1): 1–37.

Lu, Bin, Myle Ott, Claire Cardie, and Benjamin. Tsou. 2011. Multi-Aspect Sentiment Analysis with Topic Models, 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, Canada, 81–88.

Martin, Gregory J., and Joshua McCrain. 2019. "Local News and National Politics." *American Political Science Review* 113(2): 372–84.

Mimno, David, and Andrew McCallum. 2008. Topic Models Conditioned on Arbitrary Features with Dirichlet-Multinomial Regression. In Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI '08). AUAI Press, Arlington, Virginia, USA, 411–18.

Mochihashi, Daichi. 2020. "Unbounded Slice Sampling." ISM Research Memorandum No. 1209 .

Morstatter, Fred, and Huan Liu. 2016. "A Novel Measure for Coherence in Statistical Topic Models." In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Berlin, Germany. Association for Computational Linguistics. 543–48.

Newman, David, Edwin V. Bonilla, and Wray Buntine. 2011. "Improving Topic Coherence with Regularized Topic Models." In Proceedings of the 24nd International Conference on Neural Information Processing Systems (NIPS'11), 496–504.

Olivella, Santiago, Tyler Pratt, and Kosuke Imai. 2022. "Dynamic Stochastic Blockmodel Regression for Network Data: Application to International Militarized Conflicts." *Journal of the American Statistical Association* 117(539): 1068–1081.

Otjes, Simon, and Christoffer Green-Pedersen. 2021. "When Do Political Parties Prioritize Labour? Issue Attention between Party Competition and Interest Group Power." *Party Politics* 27(4): 619–30.

Park, Jong Hee. 2012. "A Unified Method for Dynamic and Cross-Sectional Heterogeneity: Introducing Hidden Markov Panel Models." *American Journal of Political Science* 56(4): 1040–54.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1): 209–28.

Rice, Douglas R. 2017. "Issue Divisions and US Supreme Court Decision Making." *Journal of Politics* 79(1): 210–22.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2016. Navigating the Local Modes of Big Data: The Case of Topic Models. In Computational Social Science: Discovery and Prediction, Analytical Methods for Social Research, ed. R. Michael Alvarez. Cambridge: Cambridge University Press. Chapter 2, 51–97.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4): 1064–82.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111(515): 988–1003.

Shinada, Yutaka. 2006. "Senkyo Koyaku Seisaku Deta Nitsuite [Election manifesto policy data]. " *Nihon Seiji Kenkyu [Japanese Political Studies]* 3: 63–91.

Wallach, Hanna M., David Mimno, and Andrew McCallum. 2009. "Rethinking LDA: Why Priors Matter." In Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS '09). 1973–81.

Watanabe, Kohei, and Yuan Zhou. 2022. "Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches." *Social Science Computer Review* 40(2): 346–66.

Wilson, Andrew T., and Peter A. Chew. 2010. Term Weighting Schemes for Latent Dirichlet Allocation. In Human

Language Technologies. The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10). Association for Computational Linguistics, USA, 465–73.

Wood, J., P. Tan, W. Wang, and C. Arnold. 2017. "Source-LDA: Enhancing Probabilistic Topic Models Using Prior Knowledge Sources." 2017 IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA, USA, 411–22.

Xie, Pengtao, Diyi Yang, and Eric Xing. 2015. "Incorporating Word Correlation Knowledge into Topic Modeling." In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado. Association for Computational Linguistics, 725–34.

Ying, Luwei, Jacob M. Montgomery, and Brandon M. Stewart. 2022. "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures." *Political Analysis* 30(4): 570–89, https://doi.org/10.1017/pan.2021.33

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Sampling Algorithm and Model Interpretation

Appendix B: Additional Information for the Base keyATM

Appendix C: Additional Information for the Covariate keyATM

Appendix D: Additional Information for the Dynamic keyATM

Appendix E: Additional Information for the Topic Matching of wLDA

Appendix F: Additional Information for the Validation