Katharyn Loweth and Wendy Shi
Text as Data
Professor Asimovic
April 7, 2025

Replication Exercise #2 Report: Keyword Assisted Topic Models by Eshima, Imai, & Sasaki (2023)

Topic models have become a widely used tool to explore and understand text data. However, there are two key limitations with typical topic models like Latent Dirichlet Allocation (LDA). First, the model-created topics may not help researchers measure the prevalence of a defined concept because the different aspects of the concept may be split across topics (Eshima et al., 2023). Second, the researchers who built the model have to interpret and label the topics after model fitting (Ethima et al., 2023). In this paper by Eshima, Imai, and Sasaki (2023), the authors argue that keyword assisted topic models (keyATM), which is a semi-supervised approach that allows researchers to label topics via the specification of keywords *prior* to model fitting, address these limitations of standard topic modeling approaches. Through multiple case studies that compare the results of keyATM to LDA models, the authors demonstrate that keyATM provides more consistent and interpretable results for text data.

Differences and Similarities

For our replication exercise, we focused on the first case study in the paper that compared the results of the base keyword assisted topic modeling versus the weighted LDA model. While KeyATMs have the capacity to include covariates in model fitting, no covariates were considered in this exercise.

The dataset for the replication exercise consisted of Congressional bills that were subject to floor votes during the 101st to 114th Sessions, a total 4,421 bills with an average of 316 bills per session. The authors obtained the text from congress.gov. The authors also used 21 labels and associated keywords for these bills that were previously compiled by the Comparative Agenda Project (CAP) as part of their analysis.

We used the corpus prepared by the authors and associated keywords linked in the authors' repository for our analysis. As part of their preprocessing steps, the authors pruned words that appeared less than 11 times in the corpus and lemmatized the remaining words via the Python library NLTK. In the prepared corpus, there were 5,537 words per bill and 7,776 unique words in the entire corpus.

Similarities:

KeyATM and wLDA Topic Topwords:

In the paper, the author used the default setting for the keyATM package to generate model results. Due to computational and time constraints, we reduced the number of iterations included in the model fitting from 1500 (the default) to 100 iterations for both the keyATM and wLDA models. However, even with the lower iterations, the model results are stable (see Figure 1 in Appendix). Using the same package, we also successfully generated wLDA results for 100

iterations. The top keywords for six topics – Labor, Transportation, Foreign Trade, Immigration, Law & Crime, and Government Operations- are included in the tables below. Table 1 includes our replicated results, and Table 2 contains the original results.

Table 1: Our KeyATM and wLDA topword results

Replication Results

Labor		Transportation		Foreign Trade	
KeyATM	wLDA	KeyATM	wLDA	KeyATM	wLDA
employee	apply	**transportation**	transportation	**agreement**	trade
standard	tax	**air**	highway	**trade**	change
training	taxable	cost [1]	safety	code [11]	agreement
benefit	amendment	**construction**	carrier	payment [3]	good
employment	relate	**maintenance**	vehicle	**export**	head
individual	end	safety	system	change	chapter
compensation	respect	**highway**	grant	good	article
work	credit	system	motor	commerce [14]	subchapter
employer	income	carrier	code	chapter	free
labor	rule	commercial [14]	rail	head	new
Immigration		Law & Crimr		Government Operation	
KeyATM	wLDA	KeyATM	wLDA	KeyATM	wLDA
alien	alien	**code**	security	expense	congress
immigration	attorney	**court**	committee	appropriation	house
homeland [15]	child	**enforcement**	submit	remain	code
status	code	attorney	stat	authorize	representative
nationality	court	security [14]	intelligence	necessary	senate
describe	crime	person	page	expend	congressional
application	enforcement	intelligence [15]	information	office	committee
visa	immigration	**justice**	director	exceed	veteran
attorney	offense	**crime**	technology	budget [1]	budget
period	person	grant	system	assistance [12]	office

Table 2: Authors' Original Results Included in Paper (Eshima et al., 2023)

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

The bolded words in both tables represent keywords under each predefined topic that occurs in the corresponding topic results. As we can see from the results, the keywords follow high consistency, with Labor, Transportation, Foreign Trade, and Immigration containing many overlapped words from the manually defined keywords for each topic. Government operation is considered a vague concept; neither the author's results nor our results were able to incorporate original keywords into the final topic word distribution. Despite deviating slightly from the original keywords, our models still find about 80% of overlapped words in the corpus that fit into government operations.

We believe the slight discrepancy in keyATM model performance is caused by the difference in the number of iterations between the author's model and our replication. However,

from a qualitative standpoint, the words captured for each topic all show high relevance to the pre-defined topic.

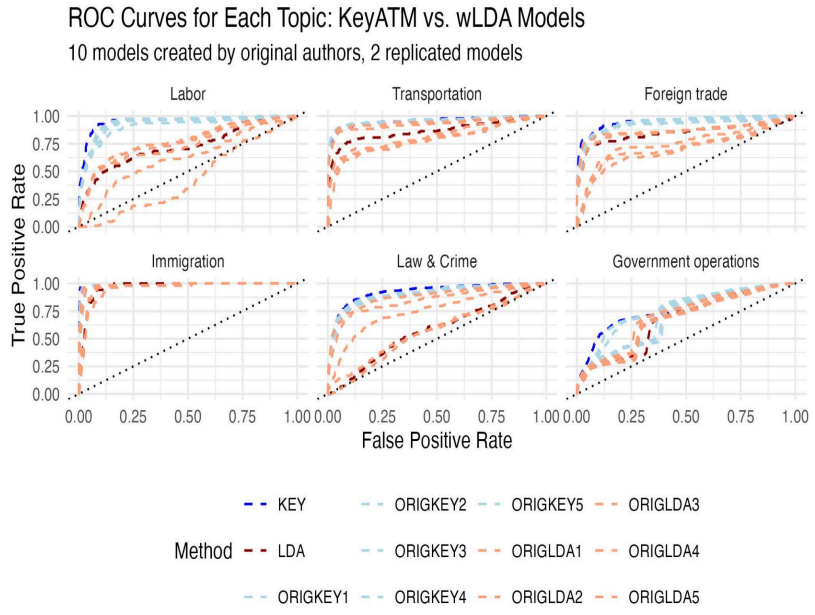
With the same package, we also draw visualizations for the keyATM estimated topic distributions across the corpus with bar plots compared to real distribution and keyword distribution with line plots that help understand the corpus. (See Figure 2, 3, 4 in Appendix).

KeyATM Classification Performance:

While both KeyATM and wLDA allow documents to relate to multiple documents (as seen through the models' theta output), the authors were also interested in how the performance of the two types of models could be interpreted as part of a classification exercise. Using the labels from the CAP, the authors applied these labels as the "true" label for the bill topic and performed a binary classification of whether the document was predicted correctly or not based on the topic with the documents' largest theta (predicted dominant topic). The authors compared the results of the two models using ROC curves.

Based on our modeling results, we confirmed their results that the KeyATM consistently outperformed the wLDA model in distinguishing topics across several topics, as seen below in Figure 1. The light blue lines in the graph below are the authors' original KeyATMs, and the dark blue line is our replicated results. Across topics, the light and dark blue lines are very close together or indistinguishable and are nearly always above the red lines representing wLDA results. The one exception is government operations, but as previously noted, the topic was not as well-defined by keywords as the other topics.

Figure 1



Differences:

wLDA Classification Performance:

While we were able to successfully implement our own KeyATM model and the authors' original KeyATM models from their online repository, our wLDA results varied. Comparing the authors' published results seen in Figure 6 in the Appendix, our results both for our own wLDA model and based on the author's original wLDA models from their online repository did not always align with the authors, especially for topics like foreign trade, based on the ROC curves.

We suspect that the discrepancies in our and the authors' results may be due to how topics were formed in the LDA model. As we previously stated, one of the limitations of LDA models is that different aspects of a defined concept can be split across multiple topics. Looking through the results of the authors' model, for example, words related to transportation have been divided across two topics, one more focused on land transportation and the other on air transportation. Thus, identifying a single LDA topic to align with the predefined label/topic as a comparison was challenging in our replication. The authors also did not specify in their write-up or code how they identified the corresponding wLDA topics from each model.

Autopsy

To reproduce the results of keyATM and wLDA performance comparison and ROC curve visualization, we utilized the base_main.R file and the saved keyATM and wLDA models from the authors' public repository. Additionally, we used the keyATM package that the authors built to finish the replication process.

When generating the keyATM results in the base_main.R file, we discovered that the authors used functions to help generate word-topic distributions that were stored in a separate common utility R file. We were not able to successfully load this file due to local memory issues and deploy these functions in our R workplace. Additionally, while some of the functions were part of the current keyATM package, others were not. We suspect that this could be caused by package updates when many functions written in the utility functions were no longer available in the updated keyATM package. Therefore, we opted for the keyATM R package created by the same author to generate the model results with preprocessed data and keywords.

The base_main.R file shows how the authors checked the keyATM model performance against the wLDA model. Using the function from the keyATM package, we generated the wLDA model with the author's preprocessed data. We were also able to use the supplied code for the multiROC package to replicate the authors' results. However, the author described that they implemented a Markov Chain Model to ensure the model results for wLDA converge after 3000 iterations, but this code was not included in the code file, and the package did not have a way to implement Markov Chains.

To interpret the wLDA model results and align them with the true labels and keyATM results, we had to manually access the top word results and assign topic names to the 21 models. The order and formation of the topics for the wLDA models varied across models. The generic Topic name for the wLDA model results creates complexity when plotting our own model results against the author's original results. Thus, we could not ensure that our interpretation and selection of the wLDA topics for the comparison matched the authors'. In addition, we need to

alter column names when merging results for different iterations in a format that multiROC recognizes.

The author provides very detailed replicated results, which contain 100+ files of model results, R code, and a Read-me file that specifies the functionality of each document. The code is very clean and does not contain unnecessary chunks that do not serve the objective of the paper. However, we do believe the code file lacks coherence between sections. The order of execution remains unclear in the read-me file when the R file tends to have generic names such as "main," "based main," and "based common utilities." The Readme file can benefit from having a more detailed description that clearly outlines how each code file relates to each other and relates to the paper.

Extension

One improvement the authors could make in their argument is to conduct additional validation by manipulating the keyword input to further demonstrate the robustness and limitations of the modeling approach.

As part of our extension of the authors' work, we decided to further examine the effect that the list of keywords had on KeyATM results. Using the original list of keywords, we created 5 new lists, each containing a random selection of 5 keywords per topic. We then fitted new KeyATM models based on each of the 5 lists and compared the results to our original findings by qualitatively assessing the keywords and comparing the ROC curve results. The 5 model results for the transportation topic are included below.

Table 3

Topic Transportation Topic Keyword

	Replication 1	Replication 2	Replication 3	Replication 4	Replication 5
KeyATM Keyword	**transportation**	military	transportation	transportation	transportation
	highway	force	highway	highway	highway
	safety	member	safety	safety	safety
	vehicle [8]	**air**	carrier	carrier	code [11]
	carrier	code	system	system	carrier
	system	authorization	airport	air	system
	air	authority	air	vehicle	air
	code	armed	code	airport	**airport**
	airport	**construction**	grant	code	strike
	grant	strike	motor	motor	motor
Selected keyword	transportation	deployment	waterway	rail	airport
	airport	air	maintenance	pilot	maritime
	aviation	construction	railroad	channel	inland
	maritime	ship	channel	traffic	freight
	infrastructure	inland	inland	travel	construction

In four of the five results, the model continued to associate the transportation topic with transportation-related words, even in instances when "transportation" was not a keyword used to inform the model. We also saw that the ROC curves for four replications remain very close to the original ROC curves (see Figure 7 in the Appendix) . However, for Replication 2, the combination of keywords skewed the results so that the topic results are more related to military topics rather than transportation and performed significantly worse in the classification exercise.

We infer from our model results that it is still possible to generate accurate topic results with distinguishable keywords predefined for each topic for semi-supervised learning. However, model performance will fluctuate when we select keywords that perfectly fit both the desired and latent topics. Therefore, selecting distinguishable terms is the key determinant for model performance, which can be guaranteed by increasing the number of selected keywords.

We believe keyword-assisted topic models are a very effective way of uncovering the latent structure of a corpus. This model could be applied when we have ample domain knowledge of the corpus of interest. For example, we could apply keyATM in the rallying speech of candidates with different party affiliations when examining perceptions of very specific topics of "Immigration," 'International Trade", "Foreign Aid", or "Taxation" in different political orientations.

Reference

Eshima, S., Imai, K. and Sasaki, T. (2024). Keyword-Assisted Topic Models. *American Journal of Political Science*, 68: 730-750. https://doi.org/10.1111/ajps.12779

Appendix

Figure 1: KeyATM model performance metrics

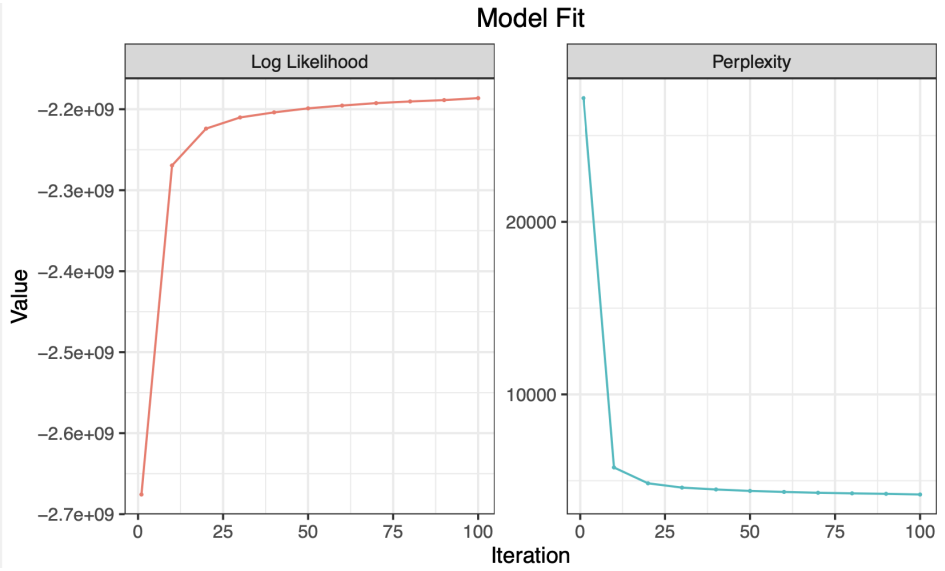


Figure 2: Key word distribution across the corpus

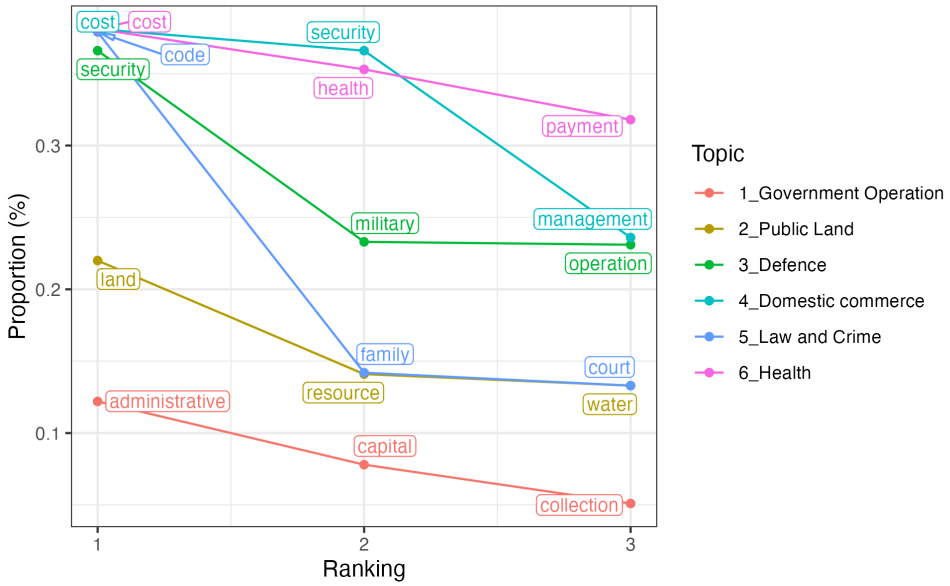


Figure 3: Original topic distribution

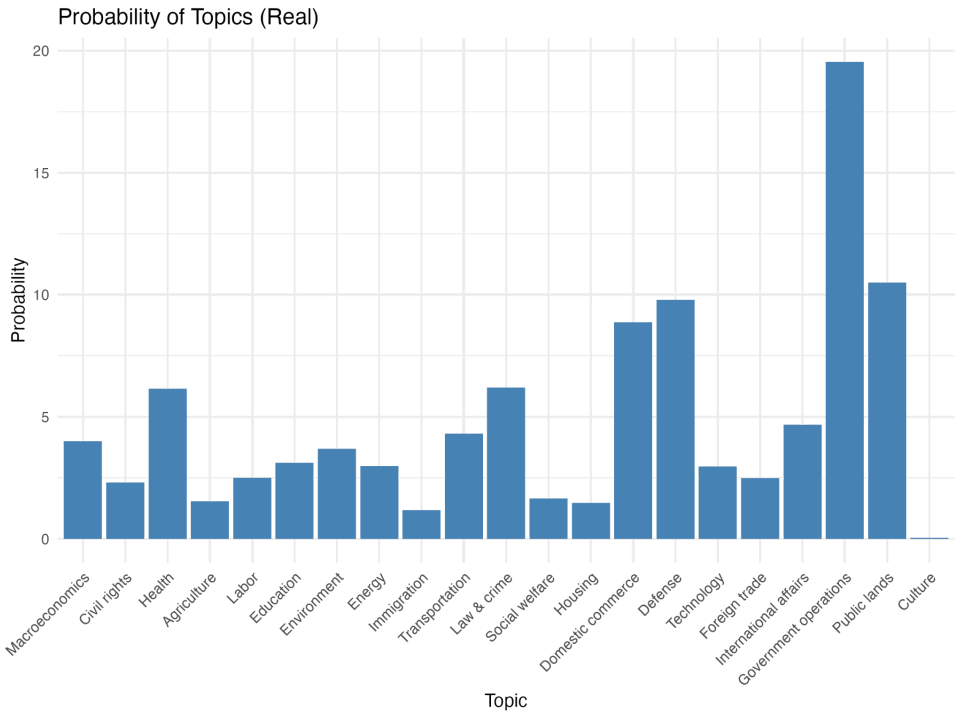


Figure 4: keyATM estimated model performance

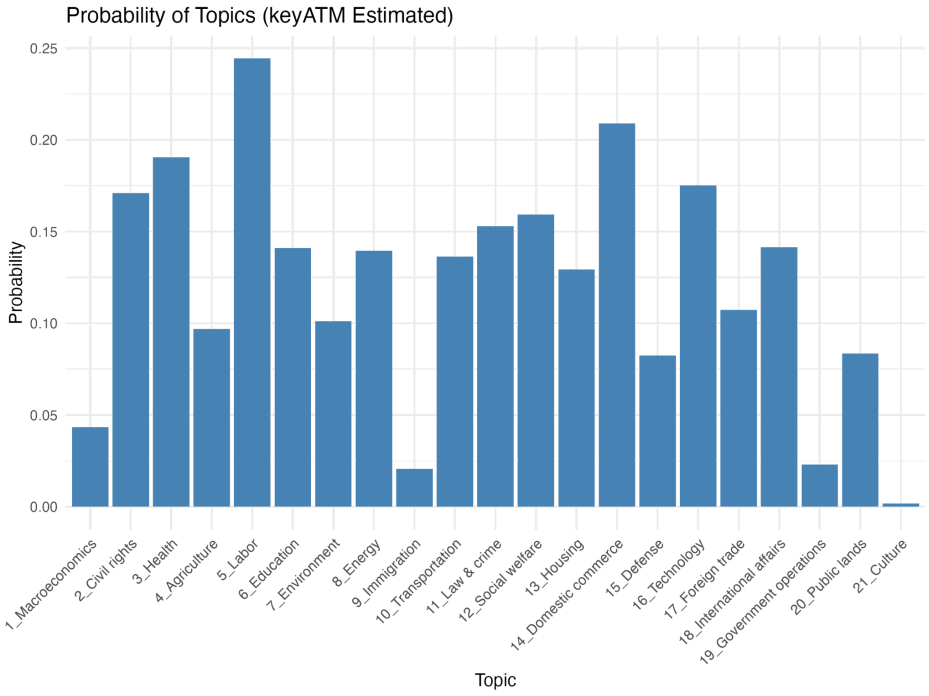


Figure 5: ROC curves for our implemented KeyATM and wLDA model for 6 topics

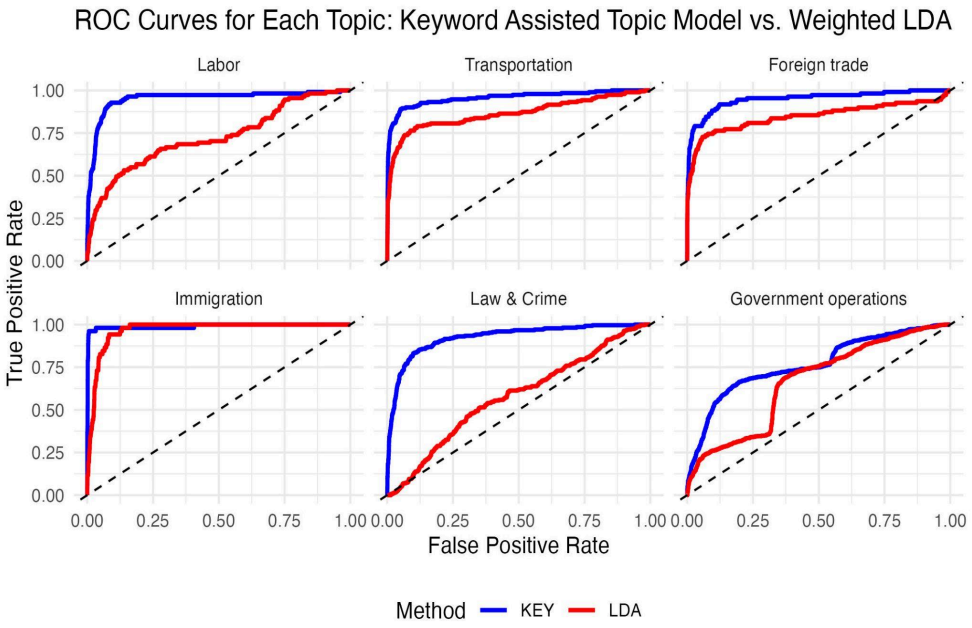


Figure 6: Comparison of keyATM and wLDA model performance (from original paper)

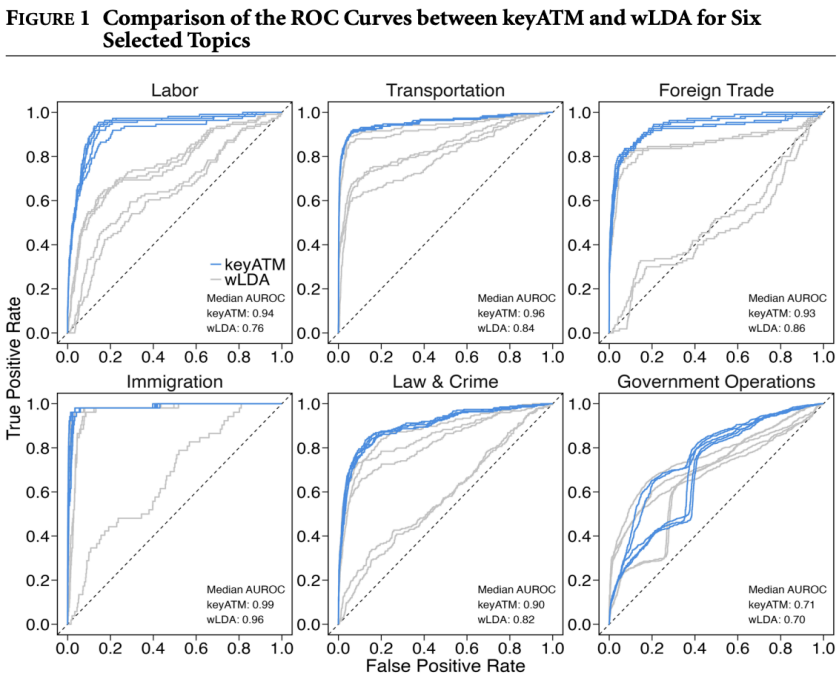


Figure 7: ROC curves comparing performance with full list of keywords vs. subsample

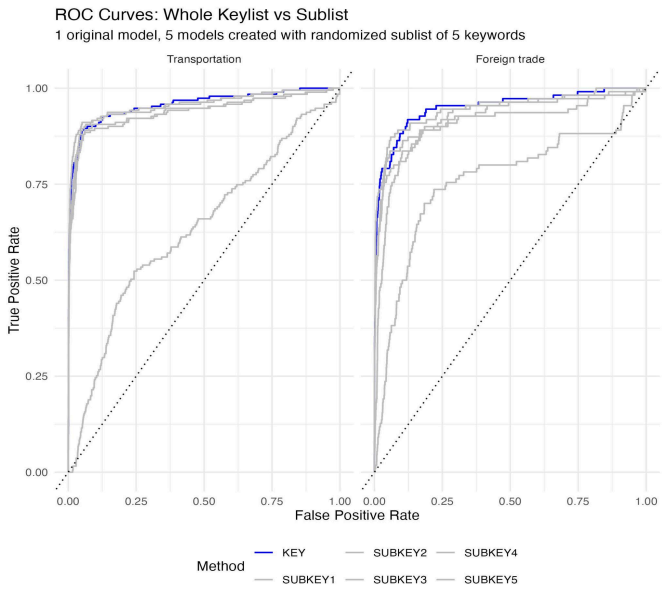


Table 1: Keyword results for Foreign Trade topic using shortened keyword list

Topic Foreign TradeTopic Keyword

	Replication 1	Replication 2	Replication 3	Replication 4	Replication 5
KeyATM Keyword	**trade**	**agreement**	**agreement**	**foreign**	**agreement**
	product	**foreign**	trade	**international**	**international**
	change	**international**	**export**	assistance	**trade**
	agreement	country	change	country	change
	tobacco	assistance	good	government	good
	good	president	head	president	article
	head	trade	chapter	development	head
	article	government	free	committee	chapter
	import	export	**foreign**	export	country
	chapter	change	article	organization	free
Selected keyword	international	import	dispute	international	aggreement
	barrier	aggrement	foreign	foreign	treaty
	trade	international	export	barrier	trade
	negotiation	balance	productivity	competitiveness	international
	import	foreign	aggrement	exchange	competitiveness