Replication II: Keyword-Assisted Topic Models

By Katharyn Loweth and Wendy Shi

Original author: Eshima, Imai, & Sasaki (2023)

Outline

Keyword Assisted Topic Models by Eshima, Imai, & Sasaki (2023)

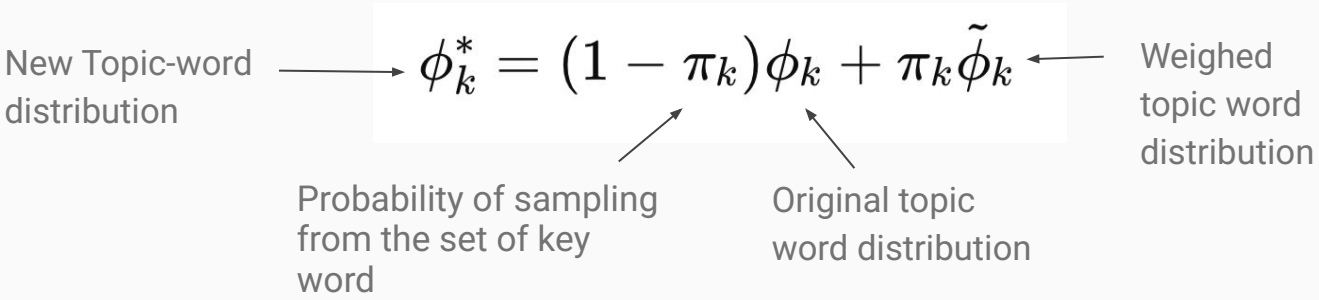
- Overview on Keyword Assisted Topic Models vs. LDA Topic Models
- Replicated Results based on their study of Congressional Bills
- Extension of study where we manipulate the keywords used to inform the keyword assisted topic models
- Autopsy of code

Keyword Assisted Topic Models by Eshima, Imai, & Sasaki (2023)

- Topic models have become a widely used tool to explore and understand text data.
- However, authors identify two key limitations with typical topic models (LDA):
  1. 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
  2. Researchers have to interpret and label the topics after model fitting
- Authors argue that keyword assisted topic models are one solution to these limitations

What are keyword assisted topic models (keyATM)?

- Semi-supervised approach that allows researchers to label topics via the specification of keywords prior to model fitting
- KeyATM allow one to predefined name of the Topic, and allow additional non-labeled topics to be classified



Comparing KeyATM & wLDA

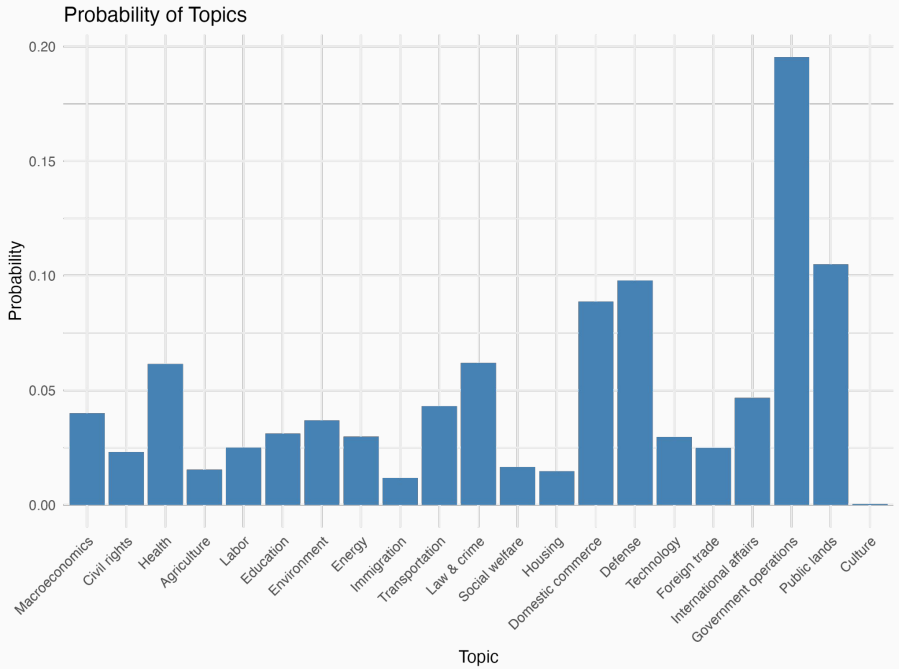
	KeyATM	wLDA
Fixed Word-topic distribution	No	Yes
Adaptive learning	Yes	No
Allow Co-variate	Yes	No
Interpretability	High	Lower

Replication Exercise

Congressional bills that were subject to floor votes during the 101st to 114th Sessions.

a. Text obtained from congress.gov.
b. The authors use 21 labels and associated keywords for these bills that were previously compiled by the Comparative Agenda Project (CAP) as part of their analysis. These labels are treated as the “true” label for the bill topic.
c. 4,421 such bills with an average of 316 bills
d. Pruning words that appear less than 11 times in the corpus, and lemmatizing the remaining words via the Python library NLTK
e. 5,537 words per bill and 7,776 unique words in the entire corpus.
f. The maximum document length is 152,624 and the minimum is 26.

Replication Exercise 1: Regenerate keyATM

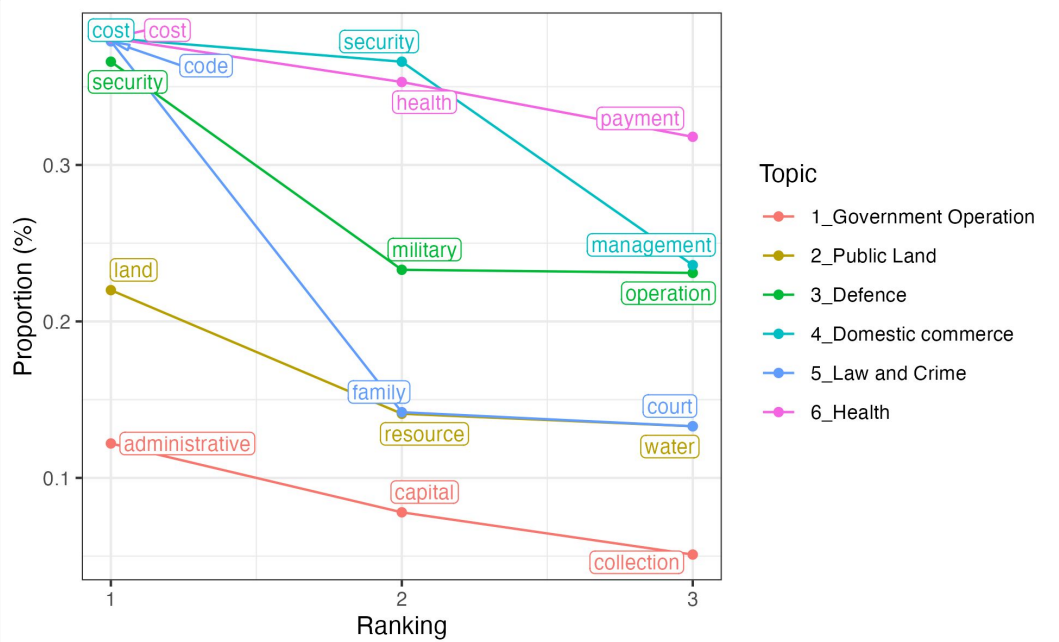


Top 6 Most common topic from the manual labeling result:

1. Government Operation
2. Public Lands
3. Defense
4. Domestic commerce
5. Law & Crime
6. Health

Sample Keywork Distribution

```
key_words_selected <-
  list('Government Operation' = c('administrative', 'capital', 'collection'),
       'Public Land' = c('land','resource', 'water'),
       'Defence' = c('security', 'military','operation'),
       'Domestic commerence' = c('cost','security','management'),
       'Law and Crime' = c('code', 'family', 'court'),
       'Health' = c('cost', 'health', 'payment')
       )
key_viz2 <- visualize_keywords(docs = keyATM_doc, keywords = key_words_selected)
key_viz2
```



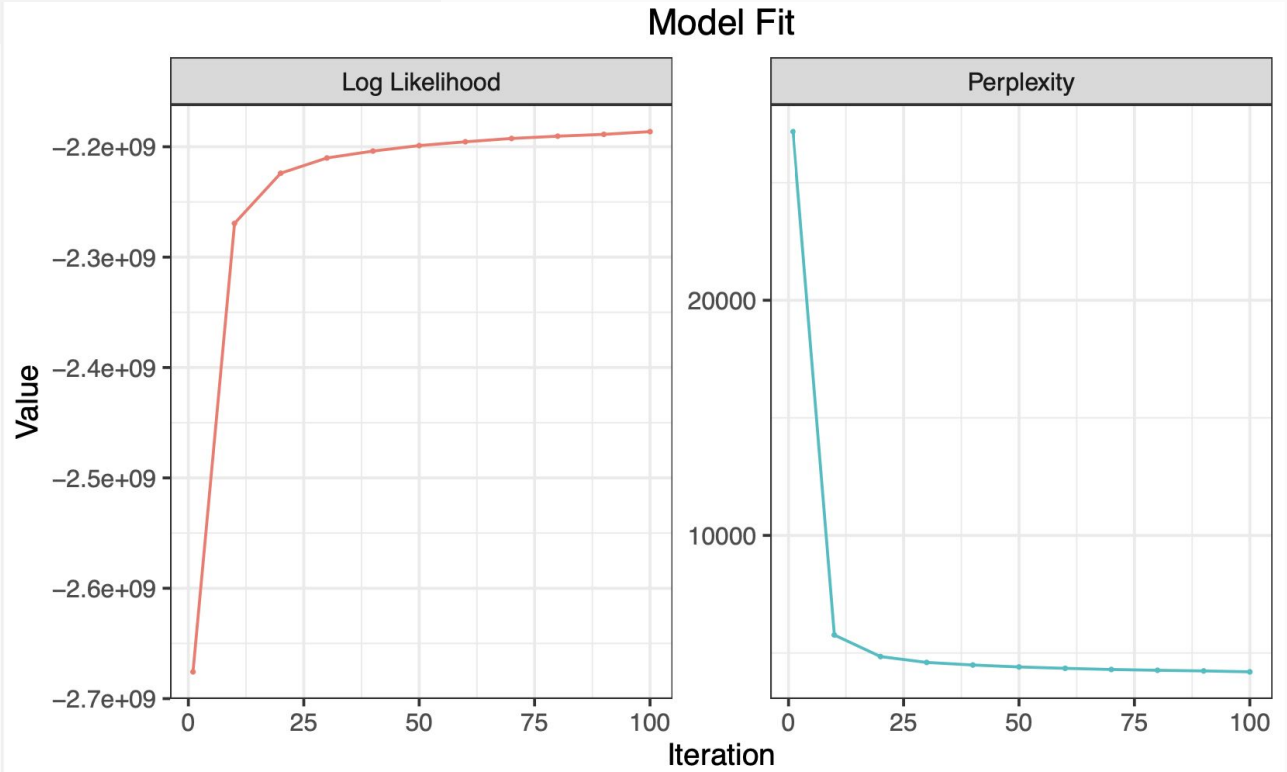
- On average 20 keywords per topics
- Minimum: 3
- Maximum: 25

KeyATM Generation

```{r}
#default is 1500
my_options <- list(
  seed           = 250, # Generating random seed
  iterations     = 100)

key_model <- keyATM(
  docs       = keyATM_doc, # text input
  no_keyword_topics = 3,
  keywords   = keys_list, # keywords
  model      = "base",      # select the model
  options    = my_options,  # use your own option list
)
```

KeyATM Model Fit

```
fig_modelfit <- plot_modelfit(out)
fig_modelfit
```



Topic 5: Labor (25 keywords)

Labor	
keyATM	wLDA
employee	apply
benefit	tax
individual	amendment
rate	end
compensation	taxable
period	respect
code*	period
payment*	individual
determine	case
agreement*	relate

Author's result

Labor	
KeyATM	wLDA
employee	apply
standard	tax
training	taxable
benefit	amendment
employment	relate
individual	end
compensation	respect
work	credit
employer	income
labor	rule

Our result

Keywords: bargaining, benefit, compensation, debt, employee, employer, employment, Fair, injury, insurance, job, labor, minimum, pension, protection, retirement, standard, training, unemployment, union, wage, work, worker, workforce, youth

Topic 10: Transportation (21 keywords)

Transportation	
keyATM	wLDA
transportation	transportation
highway	highway
safety	safety
carrier	vehicle
air	carrier
code*	motor
system	system
vehicle*	strike
airport	rail
motor	code

Author's result

Transportation	
KeyATM	wLDA
transportation	transportation
air	highway
cost [1]	safety
construction	carrier
maintenance	vehicle
safety	system
highway	grant
system	motor
carrier	code
commercial [14]	rail

Our result

keywords:
air, airport, aviation, channel, construction, deployment, freight, highway, infrastructure, inland, maintenance, maritime, mass, pilot, rail, railroad, ship, traffic, transportation, travel, waterway

Topic 17: Foreign Trade (17 keywords)

Foreign Trade

keyATM	wLDA
product*	air
trade	vessel
change	airport
agreement	transportation
good	aviation
tobacco*	administrator
head	aircraft
article	carrier
free	administration
chapter	coast

Author's result

Foreign Trade

KeyATM	wLDA
agreement	trade
trade	change
code [11]	agreement
payment [3]	good
export	head
change	chapter
good	article
commerce [14]	subchapter
chapter	free
head	new

Our result

Keywords: agreement, balance, barrier, competitiveness, dispute, exchange, export, foreign, import, international, negotiation, private, productivity, subsidy, tariff, trade, treaty

Topic 9: Immigration (3 keywords)

Immigration	
keyATM	wLDA
security*	alien
alien	attorney
immigration	child
homeland*	crime
border*	immigration
status	grant
nationality	enforcement
describe	person
individual	court
employer*	offense

Author's result

Immigration	
KeyATM	wLDA
alien	alien
immigration	attorney
homeland [15]	child
status	code
nationality	court
describe	crime
application	enforcement
visa	immigration
attorney	offense
period	person

Our result

Keywords: Citizenship, immigration, refugee

Topic 11: Law & Crime (25 keywords)

Law & Crime	
keyATM	wLDA
intelligence*	security
attorney	information
crime	intelligence
court	homeland
enforcement	committee
criminal	director
code	system
offense	foreign
person	government
justice	office

Author's result

Law & Crime	
KeyATM	wLDA
code	security
court	committee
enforcement	submit
attorney	stat
security [14]	intelligence
person	page
intelligence [15]	information
justice	director
crime	technology
grant	system

Our result

Keywords:
abuse, border, code, combat, court, crime, criminal, custom, cyber, drug, enforcement, family, fine, judiciary, justice, juvenile, legal, penalty, police, prison, release, representation, sexual, terrorism, violence

Topic 19: Government Operation (24 keywords)

Government Operations

keyATM	wLDA
expense	congress
appropriation	house
remain	senate
authorize	office
necessary	committee
transfer*	commission
expend	representative
exceed	congressional
office	strike
activity	bill

Author's result

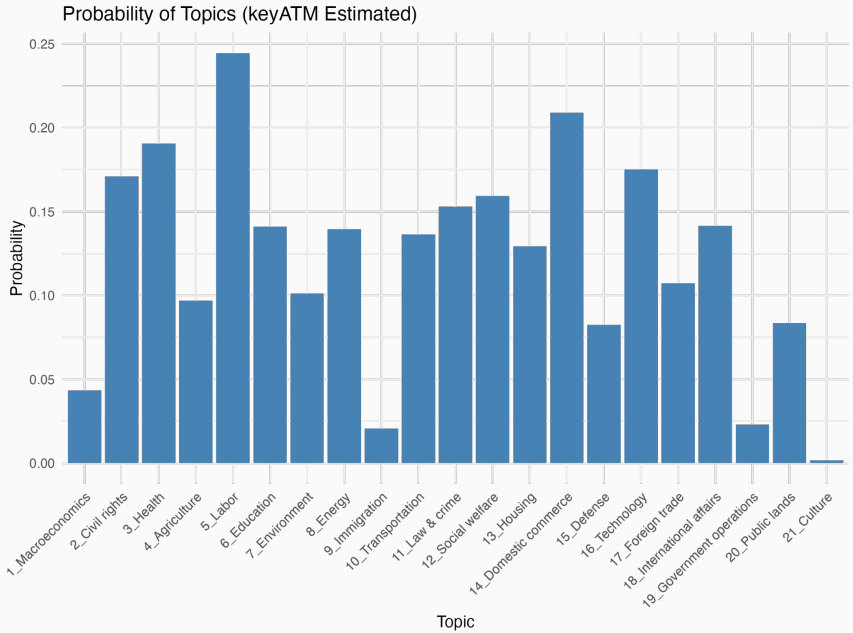
Government Operation	
KeyATM	wLDA
expense	congress
appropriation	house
remain	code
authorize	representative
necessary	senate
expend	congressional
office	committee
exceed	veteran
budget [1]	budget
assistance [12]	office

Our result

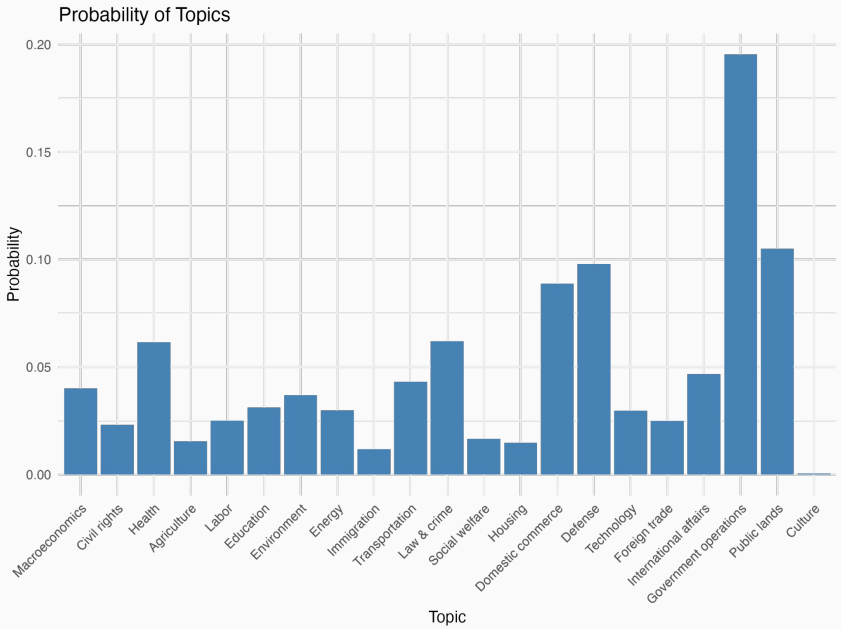
Keywords: administrative, advertising, appointment, attack, auditing, branch, campaign, capital, census, city, coin, collection, currency, mail, medal, mint, nomination, post, postal, registration, statistic, terrorist, victim, voter

Topic distribution generation

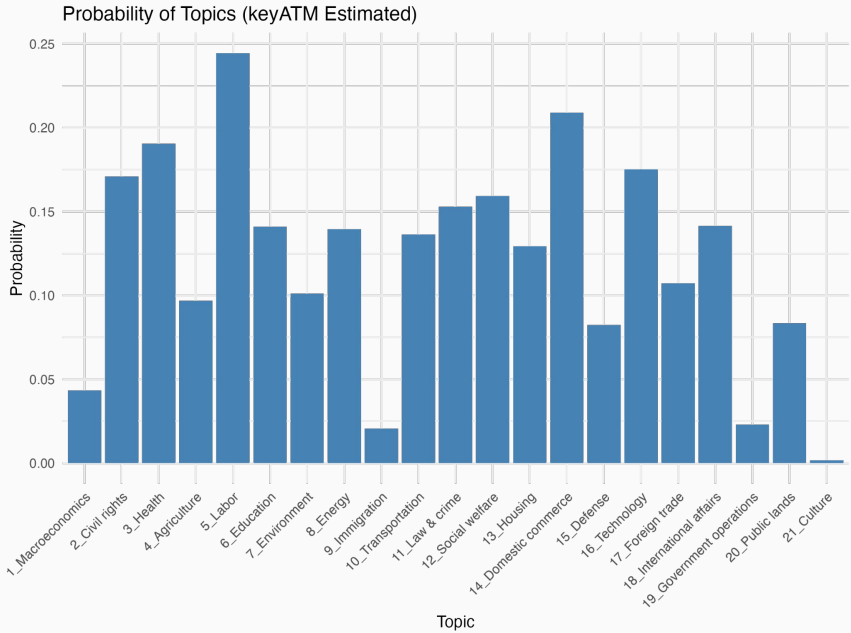
```r
{r}
library(ggplot2)

#transform visualizaton to table
topic_dis <- values_fig(plot_pi(out))

# Make sure Topic is ordered by Probability or by original order
topic_dis$Topic <- factor(topic_dis$Topic, levels = topic_dis$Topic)

# Plot
estimated_count_vish <- ggplot(topic_dis, aes(x = Topic, y = Probability)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  labs(title = "Probability of Topics (keyATM Estimated)", x = "Topic", y = "Probability") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels

estimated_count_vish

ggsave("estimated_count_vish.png", plot =estimated_count_vish, width = 8, height = 6, dpi = 300
)
```



True Topic distribution v.s. KeyATM estimated topic distribution



True Topic Distribution



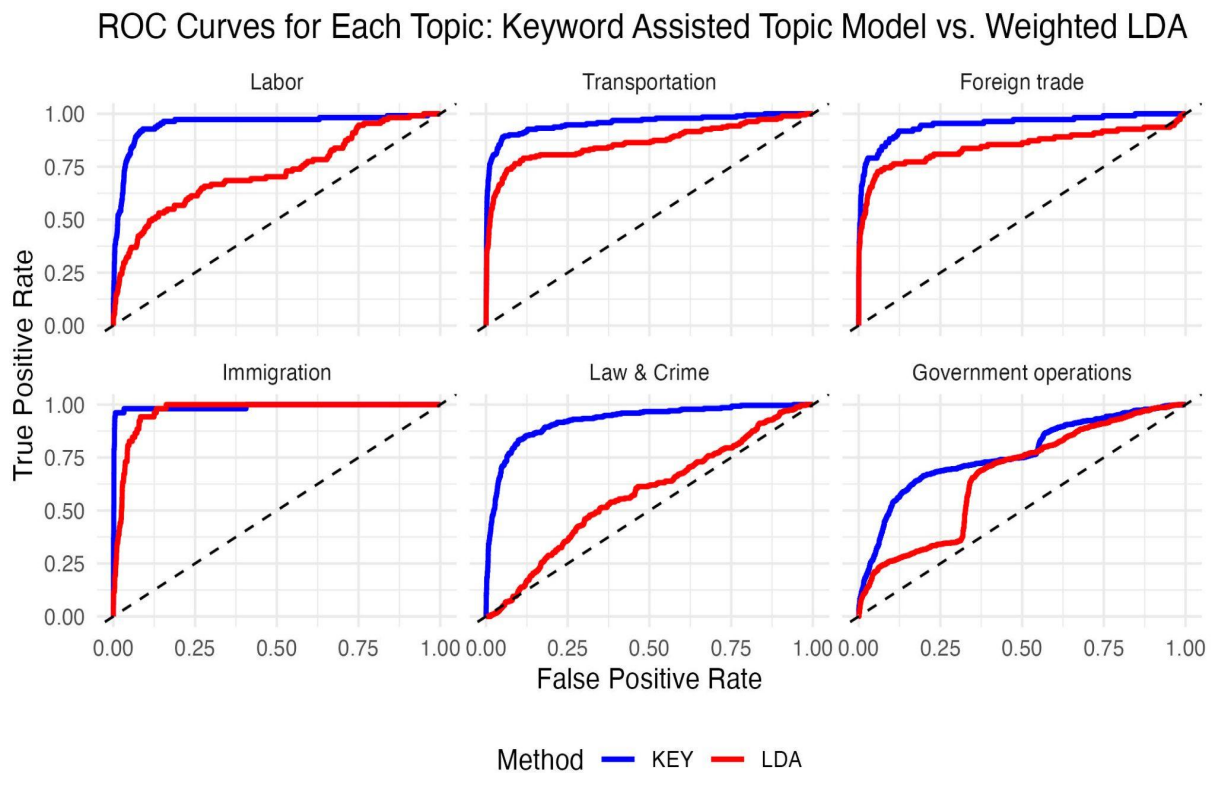
Estimated Topic Distribution

Replication Exercise #2: ROC curves

- The authors utilized ROC curves to evaluate model performance and compare classification results between KeyATM and wLDA.

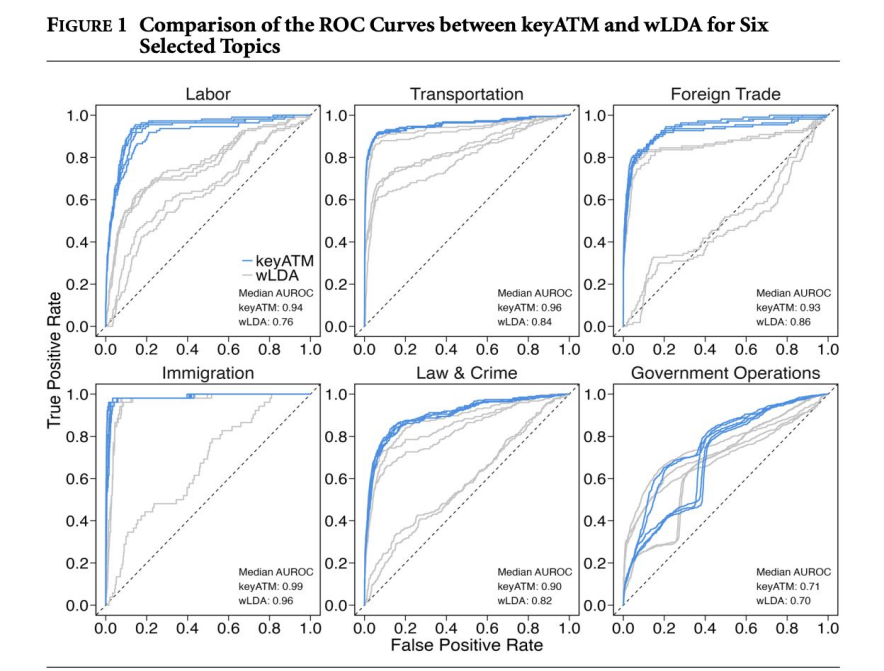
- While both KeyATM and wLDA allow documents to relate to multiple documents (as seen through theta output), the ROC curve uses the “true” topic label to make a binary comparison of whether the document was predicted correctly or not.

Replication Exercise #2: ROC curves



- We replicated the author's code using the multiROC package
- The key assisted topic model is better at classifying bills into the correct category than the weighted LDA model

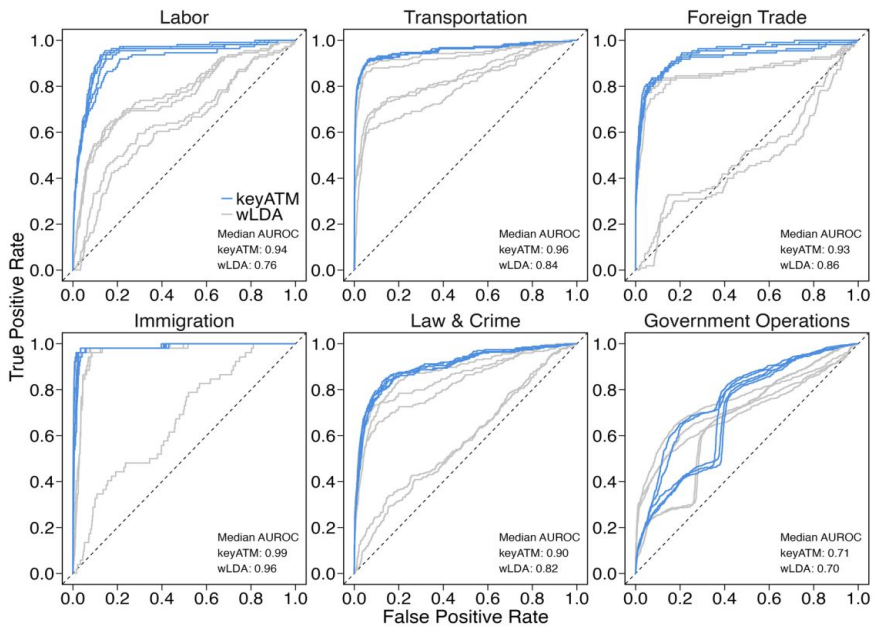
Replication Exercise #2: ROC curves

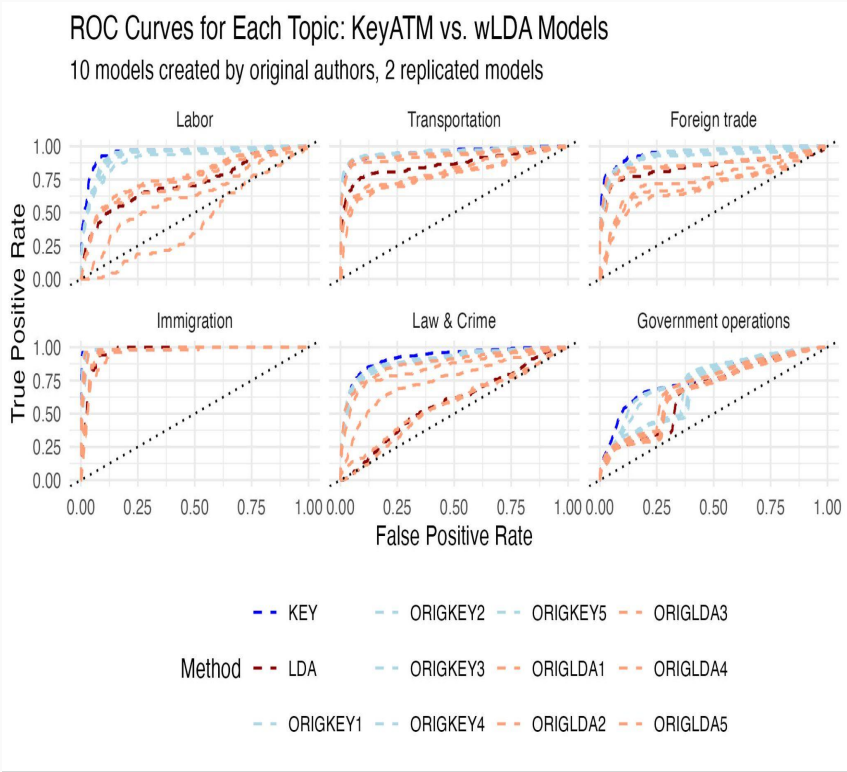


- The authors highlight that KeyATM results are more consistent/not as sensitive to starting values as wLDA through implementing and comparing results of multiple wLDA and KeyATM models with different starting values.

Replication Exercise #2: ROC Curves

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

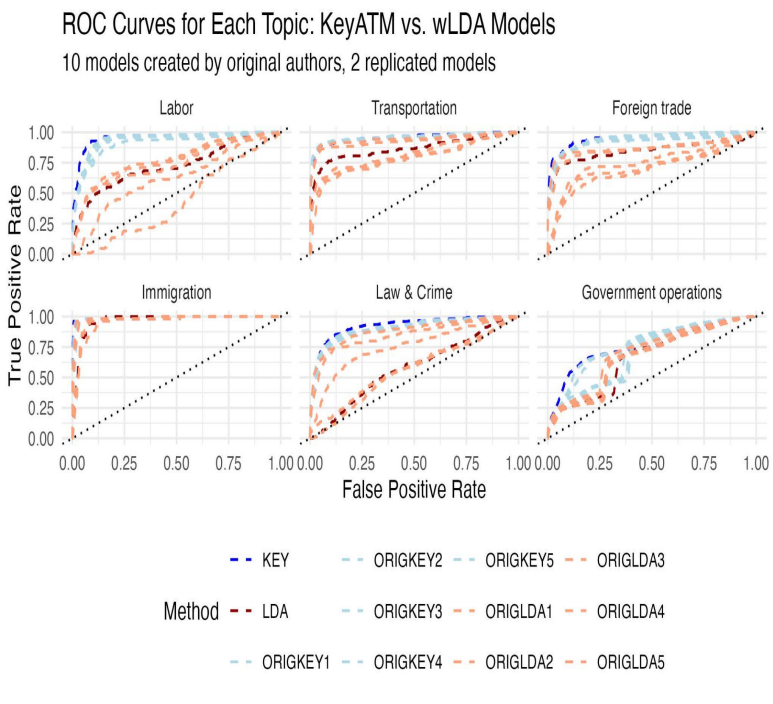




Replication Exercise #2: ROC curves

Discrepancies in our and the authors' may be due, especially with the wLDA results, likely due to how topics were formed in the LDA model.

- Single defined topic split across multiple topics
- Topic order switching between wLDA models.



Extension Activity

- One of the important assumptions of keyword assisted topic models is that the keywords align well with the identified topics.
- In this study, the majority of topics are informed by 15+ keywords. For our extension, we decided to randomly select a subsample of the keywords (5) and compare results.
- Focus on two topics for this exercise, transportation and foreign trade, as examples of topics with clear keywords (airport) and broader keywords (agreement), respectively.

Extension

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

Reduced keyword size can also generate accurate model performance!

Transportation

KeyATM	wLDA
transportation	transportation
air	highway
cost [1]	safety
construction	carrier
maintenance	vehicle
safety	system
highway	grant
system	motor
carrier	code
commercial [14]	rail

But can also lead to biased results like the topic becoming military focused

Extension

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

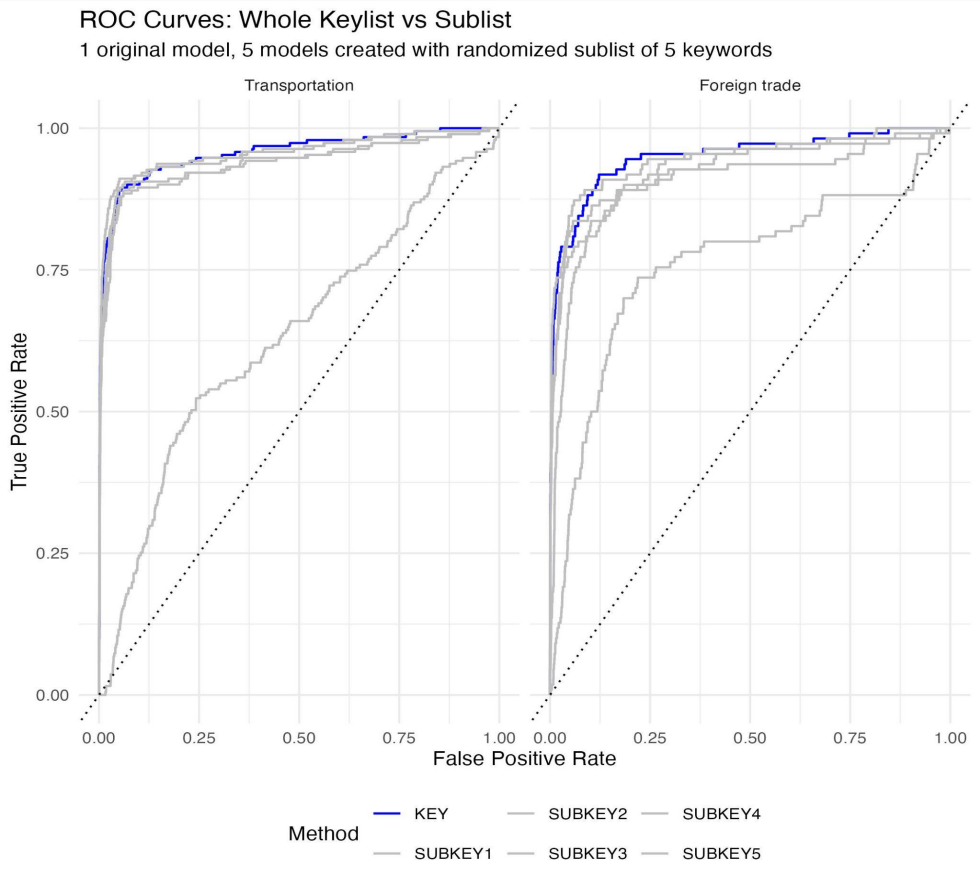
Could be “international aid”

We should avoid selecting generic keywords that could appears in other topics.

Foreign Trade

KeyATM	wLDA
agreement	trade
trade	change
code [11]	agreement
payment [3]	good
export	head
change	chapter
good	article
commerce [14]	subchapter
chapter	free
head	new

Extension



- Extension exercise demonstrates that results can fluctuate based on keywords
- Topics with clear keywords perform more consistently than those with general keywords.

Code Autopsy

- Pro: Detailed Replication Material (contains 100+ files)
  - The repository and code files were well commented and very clean
  - No random code chunks
- Con: Errors with the “common utility function” file
  - Could done better job explaining where they produce each visualization
  - Unclear that are the purposed for many detailed manually made functions
  - Could also do better job navigating reader through the file, Readme is ambiguous
  - Unclear how they set up the wLDA results to align with the keyATM results
  - Possible differences in keyATM R package since authors published paper

Overall Findings and Limitations

- KeyATM provides more interpretable and replicable results compared to standard LDA models.
- KeyATM works well only when we select distinguishable terms for each manually chosen topic; however, it may also cause biased results when we are unsure if similar words could appears in other topics.