Text as outcome: Topic models within a causal inference framework

Stanford CS224N Custom Project

Juliette Coly
Department of Economics
Stanford University
jcoly@stanford.edu

Abstract

Has 9/11 changed the focus of Congressmen on different topics? This project investigates the impact of the September 11, 2001 attacks on the focus of Congressional speeches, analyzing records before and after the event. It explores the evolution of topics in political discourse among Republicans and Democrats, employing a causal inference framework. Using a Difference-in-Difference approach, the study treats Congressional speeches as outcomes of an experiment, with 9/11 as the treatment. The analysis faces challenges due to the nature of text data, where inferred topics depend on the corpus. To address this, the study adopts a split-sample approach. Additionally, it examines the influence of different topic models (LDA and BERTopic) on the estimation of treatment effects. Results indicate that transformer-based BERTopic yields superior precision and coherence of topics compared to LDA models. Surprisingly, both models suggest that Democrats increased coverage of war topics more than Republicans post-9/11. BERTopic further reveals a significant increase in Republican coverage of terrorism, possibly substituting it for war topics. The findings underscore the critical role of textual representation in shaping the interpretation of causal estimates, highlighting the need for robust methodologies in analyzing text data within a causal framework.

1 Key Information to include

- Mentor: Shikhar Murty
- External Collaborators (if you have any): No
- Sharing project: No

2 Introduction

Has 9/11 changed the focus of Congressmen on different topics? The September 11, 2001 attacks were a major shock in the United States that had long-lasting political consequences. To name only a few, the US passed the USA PATRIOT Act in 2001, the Homeland Security Act in 2002, invaded Afghanistan in 2001, and Iraq in 2003. Using Congressional Speech records before and after the attacks, this project studies the evolution of topics in political speeches among Republicans and Democrats.

The main objective of this project is methodological. While the use of text data in social sciences has mostly been descriptive, this project incorporates text data in a causal setting. I treat the Congressmen speeches as outcomes of an experiment where the treatment is the 9/11 attacks. More precisely, I use a Difference-in-Difference (DiD) approach to study the evolution in speeches topics of the Republican congressmen compared to the evolution of the Democrat ones.

Using text data in a causal inference framework presents two difficulties. First, this complicates the inference process. Standard causal inference frameworks such as potential outcomes (Holland, 1986) assume that treatments and outcomes are known and do not depend on the data. With text data, the tone or the topic that is inferred from a document depends on the other documents in the corpus, i.e. *on the data itself*. This dependence on the data prevents proper identification of the estimands. Following the split-sample approach of Egami et al. (2022), which I will describe in more details in the next section, allows me to deal with this latent variable issue.

The second difficulty is that the topic models used (here, LDA and transformer-based BERTopic) affect the estimation of the treatment effect. Seeing how and why different models yield different estimates is important to ensure the robustness of the estimates. The goal of this project is thus to apply Egami et al. (2022) split-sample approach in a DiD setting (and is, to my knowledge, the first to do this) and to see how the representation of the texts (topic model using LDA or BERTopic) affects the results of the estimation.

This project shows that transformer-based BERTopic outperforms significantly LDA models when it comes to the precision and coherence of their topics (coherence of 0.35 vs. 0.26). Perhaps surprisingly, both LDA and BERTopic model text representations suggest that the Democrats have increased their coverage of war topics more than the Republicans after 9/11. BERTopic shows in addition that the Republicans have significantly increased their coverage of terrorism, suggesting that they may have substituted terrorism for war topics. The results demonstrate that the accuracy and caliber of the textual representation significantly influence the subsequent interpretation of causal estimates: what was a surprising result with LDA (which has only a war topic and not a terrorism one) could be better interpreted with BERTopic (which has both a war and a terrorism topic).

3 Related Work

This project is at the junction of two literatures. First, it relates to works in social science that use text data. A growing number of works in social sciences has leveraged text data (see Gentzkow et al. (2019) for a survey). A closely related work is Gentzkow and Shapiro (2019), which uses a bag-of-word model to represent Republicans' and Democrats' speeches in Congress. They specify a multinomial model of speech with choice probabilities that vary by party. This allows them to show that the party differences in speech that we observe today are a new phenomenon in US history. This project relates to this work because it uses text data from Congressional speeches to assess differences (here in the topics covered) between the Democrats and the Republicans. By leveraging advanced NLP methods such as transformers, I aimed at producing a richer and more precise analysis that is not limited to words count.

Second, this project is part of the literature that incorporates text data in a causal inference framework. While the use of text data in social sciences was mostly *descriptive*, a recent literature has started looking at text data in a *causal* setting: text as treatment (e.g what is the effect of news on investment) or as outcome (e.g how a political candidate speech affects votes). Egami et al. (2022) is the first (and only) article that proposes an approach that ensures proper causal identification when dealing with latent variables such as representation of text data. Indeed, standard causal inference (Holland, 1986) frameworks such as potential outcomes assume that treatments and outcomes are known and *do not depend on the data*. With text data, the tone or the topic that is inferred from a document depends on the other documents in the corpus, *i.e* on the data itself. This is what the authors call the Fundamental Problem of Causal Inference with Latent Variables (FPCILV). Their split-sample workflow addresses this issue and ensures valid causal inference with text data. To my knowledge, this project is the first that follows their five-step workflow and applies it with a transformer-based topic model (BERTopic). Moreover, this project applies their approach in a Difference-in-Difference setting with observational data, while the examples they give in their article are based on experimental data only.

4 Approaches

I follow a split-sample workflow first proposed by Egami et al. (2022) (see "Related Work" section). First, I split the corpus between train data and test data. This is key to prevent overfitting and identification issues.

Then, I discover the text representation mapping g and validate it using the training data. I fitted a topic model using Latent Dirichlet Allocation (LDA) and BERTopic model. The subsections will delve into the approaches of the models.

Finally, I apply g and estimate causal effect in the test set. The "Inference" subsection describes my approach.

Notation Throughout this project, the following notations will be used. Let $\mathcal{D}$ denote the set of documents, M denote the number of documents (speeches), and N_i the number of words in documents $d_i \in \mathcal{D}$ for $i = 1, \ldots, M$.

4.1 Topic model using Latent Dirichlet Allocation (Blei et al., 2003).

Description of the model LDA is a Bayesian hierarchical model. It is first supposed that when writing a text, the author draws a mixture of topics, a set of weights that will describe how prevalent the topics are. Then conditional on the topic, the actual words are drawn from a topic-specific distribution. This topic-specific distribution is common across documents and characterizes the rates at which the word appears when discussing a specific topic. Both the per-document topic distribution and the per-topic word distribution are assumed to follow a Dirichlet distribution. More formally, the data generative process is the following:

1. Choose the topic distribution in document i, θ_i, which follows a Dirichlet distribution with prior parameter α.
2. Choose ϕ_k, the word distribution of topic k for $k = 1, \ldots, K$, which follows a Dirichlet distribution with prior parameter β. Note once again that the topic distribution is the same across all the documents of the corpus.
3. For each of the word position i, j (word in position j in document d_i), for $i = 1, \ldots, M$ and $j = 1, \ldots, N_i$.
   (a) Choose a topic $z_{i,j} \sim Multinomial(\theta_i)$ (the topic distribution varies by document).
   (b) Choose a word $w_{i,j} \sim Multinomial(\phi_{z_{i,j}})$

Given a collection of documents, the goal is to infer the parameters θ_i's and ϕ_k's that best explain the observed documents. This typically involves techniques like variational inference or Gibbs sampling.

Implementation I coded a function that preprocesses the speeches. I then used `scikit-learn`'s implementation of LDA (Pedregosa et al., 2011). To tune the hyperparameter K (number of topics), I coded a function that implemented the topic coherence measure TC-W2V (O'Callaghan et al., 2015) (see the "Evaluation method section"). TC-W2V requires a word2vec model that I fitted using the `gensim` package (Řehůřek and Sojka, 2010) .

4.2 Topic model using BERTopic (Grootendorst, 2022).

Description of the model BERTopic is a five-step topic modeling technique that leverages transformers and c-TF-IDF. First, BERTopic maps the document to an embedding. I used the all-MiniLM-L6-v2 sentence-transformer[1] model, which is specifically trained for semantic similarity tasks.

Then, the dimension of the embeddings dataset is reduced. I used the Uniform Manifold Approximation and Projection (UMAP) [2], which is a non-linear technique that keeps some of a dataset's local and global structure when reducing its dimensionality.

The third step is the data clustering through a density-based cluster called HDBSCAN (Campello et al., 2013). It can find clusters of different shapes and identify outliers where possible.

Fourth, the algorithm creates a bag-of-word representation of a cluster. To do so, all the documents belonging to a cluster are concatenated into a single (big) document. The frequency of each word in this document is then computed.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[2] https://github.com/lmcinnes/umap

Next, it applies TF-IDF, considering the concatenated document of each cluster. The result is the importance scores for words within a cluster. The more important a word is within a cluster, the more it is representative of that topic. In other words, if we extract the most important words per cluster, we get descriptions of topics. The weight of word x in topic (or class) c is

$$W_{w,k} = tf_{w,k} \log\left(1 + \frac{A}{f_k}\right)$$

where $tf_{x,c}$ is the frequency of word w in topic c, A is the average number of words per topic, and f_w is the frequency of word w across all topics. The words with the highest $W_{w,k}$ in topic k are the representative words of topic k.

Implementation I removed very common words in politics. Then, I relied on the BERTopic package Grootendorst (2022) to fit the model. I computed the topic coherence for the 10 most important topics using the functions that I wrote for the LDA model.

4.3 Inference

Description I use a simple Difference-in-Difference (DiD) (Holland, 1986) framework. The estimate is computed as

$$\tau = \left(\frac{\sum_i I(Republican, post-9/11)_i g(Y_i)}{\sum_i I(Republican, post-9/11)_i} - \frac{\sum_i I(Republican, pre-9/11)_i g(Y_i)}{\sum_i I(Republican, pre-9/11)_i}\right) - \left(\frac{\sum_i I(Democrat, post-9/11)_i g(Y_i)}{\sum_i I(Democrat, post-9/11)_i} - \frac{\sum_i I(Democrat, pre-9/11)_i g(Y_i)}{\sum_i I(Democrat, pre-9/11)_i}\right),$$

which is the evolution in a topic proportion (from pre-9/11 to post 9/11) for the Republican Congressmen minus the difference for the Democrat Congressmen. This captures the difference in the evolution of topics between the two parties. I expect topics such as "War" and "Security" to have positive estimate τ, meaning that the evolution of the emphasis of the Republicans on these topics is greater than the Democrats one.

Implementation I coded the DiD estimator.

5 Experiments

5.1 Data

I use the speeches from the 106th (January, 1999-January, 2001) and 107th (January, 2001-January, 2003) Congress parsed by Gentzkow et al. (2018). The data contains 248,421 speeches (135,810 for the 106th Congress and 112,611 for the 107th Congress). For each speech, we have the transcript of the speech and the party of the Congressman who delivered it. Table 1 shows an example of an observation.

	speech_id	party	speech	Congress
90	1060000387	R	Mr. Speaker. the 106th Congress started the day with a nationwide consensus that the health of social security is in jeopardy. Millions of American seniors have come to depend on social security. and it is our responsibility to see that a solution is found to address this looming crisis. (...)	106

Table 1: Example of an observation from the Congressional record dataset Gentzkow et al. (2018)(the speech has been truncated for space sake and is displayed before pre-processing).

5.2 Evaluation method

In order to select the number of topics with the LDA approach, I computed the topic coherence using TC-W2V from O'Callaghan et al. (2015). The coherence score is the mean pairwise cosine similarity

of two term vectors generated with a Skip-gram model:

$$TC - W2V = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} cosine - similarity(wv_j, wvi)$$

where wv_i and wv_j are the word2vec representation of word i and j. N is the (chosen) number of representative words for the topic of interest. I then average the topic TC-W2V to get the overall coherence of a topic model.

5.3 Experimental details

LDA model

- I preprocessed the speech data. I lowered-case and stemmed the words, and removed usual stopwords, punctuation, and words that are very common in Congress speeches (e.g, "Speaker", "bill", ...).
- I created a document-term matrix, ignoring terms that appear in more than 80 % of the documents and less than 5% of them.
- I fitted a topic models using the Latent Dirichlet Allocation Blei et al. (2003) for a number of topic varying between 4 and 10.
- I then evaluated the coherence of each topic model. I first fitted a word2vec skip-gram model. With the vector representations of each words, I applied TC-W2V (see "Evaluation method" above) to compute the coherence of each topic model.

BERTopic model

- I removed stop words and words commonly used in politics ("Speaker", "bill", ...) to reduce the noise in the data. Then, I fitted a BERTopic model using the sentence-embedding `all-MiniLM-L6-v2`.
- The model had about 700 topics. I picked the 10 most important topics (excluding topics that deal only with Congress housekeeping) to estimate the DiD effect.

5.4 Results

Topic models The LDA model with the best average coherence is the one with four topics, for a coherence of 0.2635 (see Figure 3 in appendix A). Table 2 displays the most representative words of each topic.

Budget	Economics	Army	Health and education
tax	nation	unit	school
peopl	program	peopl	educ
get	support	nation	children
money	fund	american	health
budget	energi	countri	feder
secur	provid	war	care
american	import	right	law
need	servic	ask	provid
pay	develop	world	program
know	need	militari	need

Table 2: Representative words of the LDA topic model.

BERTopic model has an average coherence of about 0.3517. This is a great improvement relative to the best LDA model. Figure 1 displays the most-representative words of the BERTopic model for the first 10 topics [3].

[3] I have removed the topics that deal exclusively with Congress housekeeping

Moreover, we see that the topics from BERTopic are more precise than the LDA topics. The category "Health and education" in LDA is a coherent yet composite topic. By contrast, BERTopic has an "Education", an "Insurance", and a "Social Security" topic. It is therefore better to capture nuances in subjects.

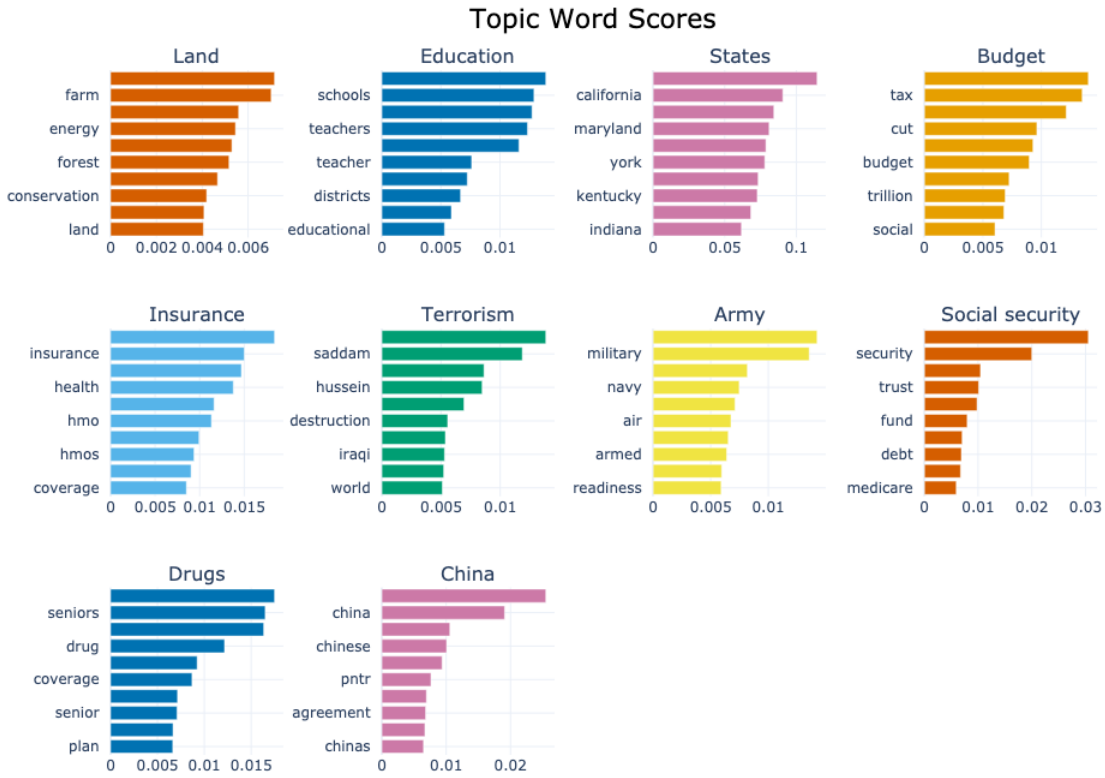
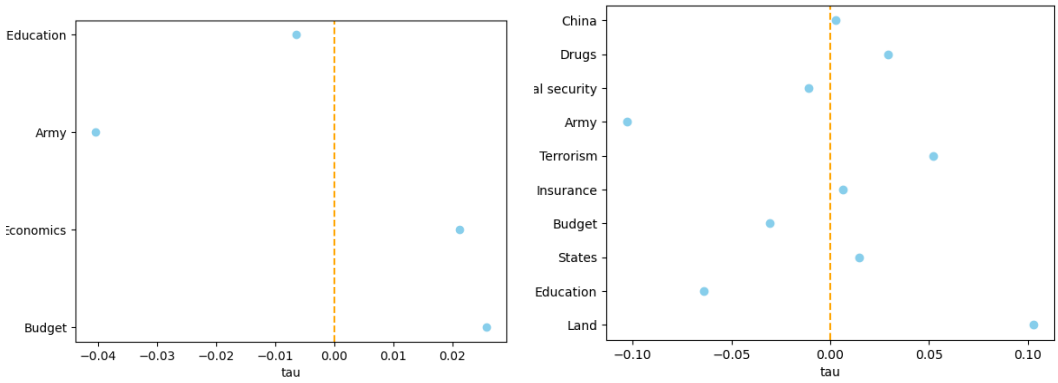


Figure 1: BERTopic representative words for the 10 most important topics

Difference-in-Difference estimates Figure 2 displays the results of the DiD estimates with the two models.

In both models, τ is negative for Education, meaning that the Democrats have increased their coverage of the Education topic more than the Republicans. Similarly, τ is negative for the Army topic, meaning that the Democrats have increased their coverage of this topic more than the Republicans. This is counterintuitive: I would have expected Republicans to turn more war-mongers than the Democrats after 9/11. This is because 9/11 has made Democrats talk more about this topic, while the Republicans were talking about it before. Morerover, for Republicans, some of the talk about Army might have been substituted by talk about "Terrorism", as the two topics are very close. This may explain the decrease in the share of the Army topic. BERTopic allows to see this substitution pattern while it is hidden in the LDA model.

The value of τ for the Economics and Budget topics goes in different directions. It is positive for the LDA estimates (the Republicans have increased their coverage relative to the Democrats) and negative with the BERTopic model. The negative effect in the LDA model is less that -0.05. I have not computed the standard errors of the estimator τ but it cannot be excluded that the (true) estimand values are actually 0 or weakly positive.



(a) LDA

(b) BERTopic

Figure 2: DiD estimates

6 Analysis

Table 3 contains two examples classified as "Army" by BERTopic. The first one is also classified as "Army" by LDA, with a proportion of 0.778 in this category. This is because the speech is mostly about arms. By contrast, the second speech describes a policy related to armed forces but is not limited to only army vocabulary. It is also about budget and health expenses (see the bolded words). As a result, LDA, which is based on bag-of-words gives a low army content to the speech (0.1482) and evenly splits across the other categories. The entire speeches are reproduced in appendix A.

This example shows how the two model work. LDA computes topic proportions according to certain key-words. BERTopic assigns the most relevant one topic to each speech. Because BERTopic output many precise topics, this one-hot encoding does not seem too problematic. BERT sentence-embedding provides a better classification than the LDA model.

7 Conclusion

The experimental results show that BERTopic is capable of significantly outperforming LDA (better coherence), demonstrating the capacity of transformer-sentence embeddings to provide performance improvements over more standard bag-of-words approaches. Moreover, BERTopic has more precise topics than the LDA models. Despite these differences, both LDA and BERTopic model text representations suggest that the Democrats have increased their coverage of war topics more than the Republicans after 9/11. This seemingly surprising result may be explained by two mechanisms. First, 9/11 has made Democrats talk more about this topic, while the Republicans were talking about it before. Second, for Republicans, the BERTOpic estimates suggest that some of the talk about Army might have been substituted by talk about "Terrorism", as the two topics are very close. This exercise thus shows that the precision and the quality of the text representation have a major impact on the downstream interpretation of the causal estimates.

The principal limitation of this work is that, sometimes, the two models give different estimates and there is no way to assess which one is right. Another limitation is the data used. The data is based on Congressmen that take the floor in Congress and therefore is not reflective of the stance of the parties as a whole.

Future work would involve to compute the standard errors of the estimate of the treatment effects, to ensure the validity of the causal interpretation. We have seen that the two models give different estimates. I'd like to find a way to aggregate or average the estimates from the models (through majority rule or other rules). This would ensure that the results are not too dependent on the text representation which is used.

	party	speech	congress	BERT-opic	Education (LDA)	Army (LDA)	Economics (LDA)	Budget (LDA)
56	R	I thank the Chair for recognition. (...) I am also pleased that the Appropriations Committee chose to specifically provide $90 million in the FY2002 Emergency Supplemental bill to accelerate the depot modernization period of the USS Scranton at the Norfolk Naval Shipyard from FY2002 to FY2003. (...)	107	Army	0.0021	0.778	0.2182	0.0021
811	D	Madam Speaker.(...) The $310 **billion** that this bill would authorize in the coming **fiscal** year represents the blueprint for defense policy and **spending** priorities as it does every year. Not only does it set the troop strength levels and extend expiring authorities. it goes to the heart of what our troops need to do the job. This bill will.. directly improve their quality of life. their readiness to fight. and the pace of the modernization of their equipment. I am especially pleased that this bill contains several important new initiatives. including a comprehensive package of military **- health care** reforms that would significantly improve access to quality **health care** for all military beneficiaries. particularly for over65 military **retirees**.(...)	106	Army	0.2569	0.1482	0.3440	0.2509

Table 3: Two examples of "Army" speeches according to BERT

8 Ethics Statement

Two ethical challenges are the political use of this work's results and the bias in the results. This project presents evidence that Republicans and Democrats talk about different topics in Congress. The results could be used by one of the two parties as a basis to attack the other one (for instance, "Your party stopped caring about education"!). While this work provides evidence about changes and differences, politicians may use the results in an un-nuanced way. This could trigger unhealthy political conflict. A mitigation strategy would be to be very clear and nuanced about the results in the abstract and the introduction. Since most newspapers and politicians won't take the time to read the entire article, avoiding "flashy" and simplistic claims in the at the beginning of the article should help limit the political use of this project.

The bias in the results come from the fact that the data is based on Congressmen that take the floor in Congress. The speeches thus reflects their opinions. The data don't tell anything about the opinions of other members of the GOP or the Democratic Party. As a result, the inference I make about the favorite topics of one party or the other is based on the spokeperson of these parties. There's a strong correlation between the view of a party and the view of its member but it is not a perfect one. To mitigate this problem, I can make clear in the article that my work does not cover the universe of Congressmen opinions in both parties.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. 2019. Text as data. *Journal of Economic Literature*, 57(3):535–74.

Matthew Gentzkow and Jesse M. Shapiro. 2019. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2018. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Paul W. Holland. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

D. O'Callaghan, D. Greene, J. Carthy, and P Cunningham. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

A Coherence of the LDA topic models

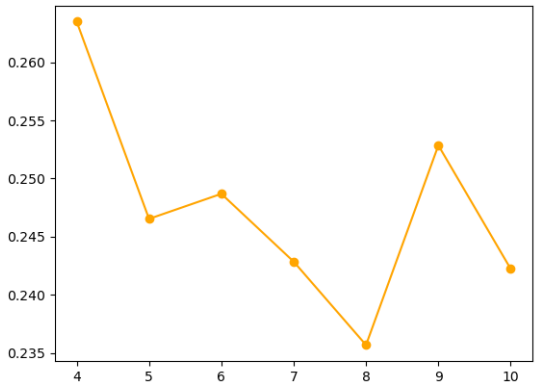


Figure 3: Average topic coherence for LDa models with different numbers of topics

B Entire speeches of the analysis section

B.1 Speech 1

I thank the Chair for recognition. I would like to express my appreciation to Mr. INOUYE. The Chair of the Senate Appropriations Subcommittee on Defense. and to Mr. STEVENS. the Ranking Member of the Subcommittee. for the fine work they have accomplished in crafting this important FY2003 Department of Defense Appropriations Bill. It has been my pleasure. as a member of the Appropriations Subcommittee on Defense. to work with them on this bill. as well as on the defense portions of the recently passed FY2002 Emergency Supplemental Bill. H.R. 4775. They certainly do a masterful job of setting priorities and balancing competing needs. I am also pleased that the Appropriations Committee chose to specifically provide $90 million in the FY2002 Emergency Supplemental bill to accelerate the depot modernization period of the USS Scranton at the Norfolk Naval Shipyard from FY2002 to FY2003. as it will result in dramatically improved fleet readiness. In addition. it will free up $90 million in FY2003. which had been programmed for the USS Scranton to be used for other U.S. Navy critical submarine requirements. This could include returning back to FY2003 the important USS Annapolis depot modernization period at the Portsmouth Naval Shipyard. which the Navy was recently forced to slip from FY2003 to FY2004. because of a Navy funding shortfall. I would like to direct a question to my friends. the chair and the ranking member of the Defense Appropriations Subcommittee. Is it the Subcommittees understanding that the appropriation of the additional $90 million to accomplish the USS Scrantondepot modernization period in FY2002. now gives the U.S. Navy flexibility to allocate the FY2003 USS Scranton funds to meet other critical submarine requirements?

B.2 Speech 2

Madam Speaker. each year. the legislative process consistently yields a particularly important authorization bill. and each and every year that authorization bill is signed into law by the President. I am speaking of the annual Defense authorization bill. A month ago on May 18. the Floyd D. Spence National Defense Authorization Act for fiscal year 2001. aptly named for our distinguished chairman in his last year at the helm of the committee. passed the House by a strong bipartisan margin of 353 to 63. The $310 billion that this bill would authorize in the coming fiscal year represents the blueprint for defense policy and spending priorities as it does every year. Not only does it set the troop strength levels and extend expiring authorities. it goes to the heart of what our troops need to do the job. This bill will.. directly improve their quality of life. their readiness to fight. and the pace of the modernization of their equipment. I am especially pleased that this bill contains several important new initiatives. including a comprehensive package of military -health care reforms that

would significantly improve access to quality health care for all military beneficiaries. particularly for over65 military retirees. But. Mr. Speaker. I am sorry to note that progress on the Defense Authorization bill. after passage in the House. has come to a sudden standstill in the other body. As I look about the legislative landscape. I see no other issue that I believe should take precedence over the authorization of the funds that our -troops need. I hope that this situation can be dealt with quickly. and that we can get about the business of going to conference on a Senate bill .and a House bill in the very near future. The Congress needs this bill. The troops need this bill. The country needs this bill.