Topic Modeling Analysis with the Japanese Parliament Debates Dataset

Ayako Watanabe[1], Taishi Okano[1]

[1]The University of Chicago
[1]ayakow@uchicago.edu, [2]taishiokano@uchicago.edu

Abstract

This study aims to visualize the discussions in the Japanese national parliament (Japanese Diet) using topic modeling techniques. We collect the Japanese parliament speeches made by congressmen from 2002 to 2023. Various topic modeling approaches, including Latent Dirichlet Allocation (LDA), Contextualized Topic Models (CTM), Top2vec, and BERTopic, are compared. Their performances are evaluated based on coherence scores and computational efficiency, and BERTopic is selected as the primary model for further analysis. Hierarchical clustering is performed to visualize the relationships between major topics, and committee labels are used to assess the effectiveness of our model. Speaker clustering is employed to identify ideological relationships among congressmen. Additionally, dynamic topic modeling is applied to analyze the temporal evolution of topics discussed by major political parties. The findings provide insights into the political, social, and economic landscape of Japan and highlight the potential of unsupervised learning techniques for analyzing parliamentary debates.

Index Terms: natural language processing, topic modelling

1. Introduction

1.1. Motivations

Can we visualize the Japanese national parliament by using unsupervised learning techniques? What topics have been under discussion by the Japanese congressmen? This question attracts us because it helps us to understand how Japan has evolved politically, socially, and economically, provides insights into the relationship between the government and its citizens, and has broader implications for understanding Japan's historical context.

Parties have significant roles in Japanese politics, and the analysis based on political parties would be useful to understand their priorities. Also, by focusing on each congressman, we can reveal their ideological relationships. Applying unsupervised learning to group them would provide a different view from the traditional groupings where each congressman belongs.

1.2. Literature Review

Since the works of [1] and [2], the task of extracting topics from large text corpora has been actively investigated, including in political science fields ([3], [4] and [5]). However, few studies have been conducted on Japanese political data due to the limited availability of Japanese NLP resources until recently ([6] and [7]). In this study, we apply cutting-edge topic models to our Japanese parliament speech data.

2. Data

We scrape the Japanese parliament debates data using an API [8]. Our data includes all speeches made on the Japanese parliament from 2002 to 2023 (901,851 records). The dataset includes the following features:

- speech date
- speech (text)
- speaker name
- house (House of Representatives or House of Councillors)
- committee name

To tokenize the texts into words, we use a Morphological Analyzer called mecab-ipadic-NEologd [9]. For text processing (only used in required cases), we extract nouns and exclude stopwords [10] commonly used in Japanese NLP.

For presentation purpose, we use Googletrans [11] to translate Japanese to English.

3. Model

We employ and compare Latent Dirichlet Allocation (LDA), Contextualized Topic Models (CTM), Top2vec, and BERTopic based on [12].

3.1. LDA

Blei et al. [1] introduced LDA to topic modeling, which is a generative probabilistic model that represents documents as a combination of latent topics using a bag-of-words approach. Processed data is used as input, and the number of topics is determined as a hyperparameter. A limitation is that it disregards the semantic structure of the text. We utilize the gensim package [13] for LDA implementation.

3.2. CTA

CTM (Correlated Topic Model) is a neural topic modeling approach that combines Neural ProdLDA with SBERT embedded representations, as introduced by Bianchi et al. [14]. It partially incorporates the semantic structure of the text to overcome LDA's drawback, but it still requires processed data. It can be implemented with package [15].

3.3. Top2vec

Top2Vec is an innovative topic modeling algorithm introduced by Angelov [16]. It utilizes word embeddings, dimensional reduction, clustering and centroid calculation to extract topics. Unlike traditional methods like LDA and CTM, Top2Vec operates on unprocessed data and automatically determines the number of topics. Additionally, Top2Vec incorporates the semantic structure of the text, enhancing the quality of topic modeling. The implementation of Top2Vec can be found in the package [17].

3.4. BERTopic

BERTopic [12] is an extension of the Top2Vec approach. Grootendorst [12] points out that Top2Vec assumes that words and topics lie within a sphere around a centroid, which may not always be the case. In contrast, BERTopic calculates TF-IDF for each cluster to derive topic representations (see Fig. 1 for the entire structure of model).

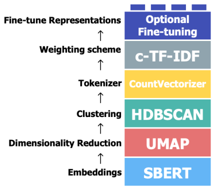

Figure 1: Structure of BERTopic [18]

4. Methodology

4.1. Evaluation

We evaluate the performance of our models using two metrics: perplexity and coherence scores. Perplexity measures the uncertainty of the model using the following formula [1]:

$$perplexity(D) = exp(-\frac{\sum_{d=1}^{M} \log p(\mathrm{w}_d)}{\sum_{d=1}^{M} N_d})$$

where $D = \{\mathrm{w}_1, \dots, \mathrm{w}_M\}$ is a corpus collection of M documents and each document has a sequence of N_d words.

Another metric we employ is topic coherence score, which measures how the top-k words of a topic are related to each other [19]. We utilize the Normalized Pointwise Mutual Information (NPMI) coherence score, as proposed by [20].

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\ln{(p(w_i, w_j))}}$$

$$PMI(w_i, w_j) = \ln{(\frac{p(w_i, w_j)}{p(w_i)p(w_j)})}$$

. where w_i, w_j are the representative words in topics.

For example, since lower perplexity and higher coherence are better, Fig. 2 suggests using 10 to 30 topics for LDA.

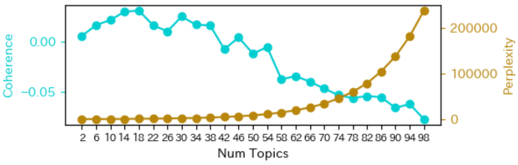

Figure 2: Hyperparameter tuning on LDA (Regarding the number of topics)

4.2. Model Selection

We compare the performance of different models using OCTIS [21] and Grootendorst's implementation [22]. Since perplexity could not be calculated for all models, we primarily rely on the NPMI metric for evaluation. The models are mainly trained with default hyperparameters, except for the choice of embedding models. For CTM, we use the "colorfulscoop/sbert-base-ja" embedding model, for Top2vec we use "universal-sentence-encoder-multilingual", and for BERTopic we use "paraphrase-multilingual-mpnet-base-v2". These embedding models are selected based on their ability to handle Japanese text.

4.3. Speaker Clustering

Using the results of the topic modeling, we calculate the similarities between congressmen. First, we generate vectors for each congressman by aggregating the number of topics in which their speeches are categorized and mapping them onto vectors. Then, we apply t-SNE to reduce the dimensionality of the vectors to 2D, considering the nonlinearity of the data.

4.4. Dynamic Topic Modeling

To analyze the temporal evolution of topics, we implement Dynamic Topic Modeling (DTM). Generally, the static topic modeling approaches are unable to capture how topics change over time. However, BERTopic can be easily extended to DTM without retraining by modifying the TF-IDF calculation.

5. Result

5.1. Word Cloud

We construct word clouds to assess the model performance qualitatively. The data used for evaluation ranges from 2022.01.01 to 2023.04.30. The models are evaluated using both unprocessed (UP) and processed (P) data.

Fig. 3 shows several major topics discussed one year ago in Japan, including medical matters for the COVID-19 and establishing the new digital agency. Although most topics make sense, a few topics seem unreasonable, like Top2Vec's second major topic and the third major topic of BERTopic (not in this paper). This suggests using unprocessed data might produce some unreasonable topics.

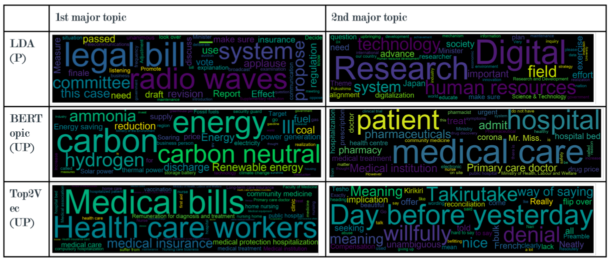

Figure 3: Comparison of word clouds. Note: The phrase of 1-3 words in pictures is one word in Japanese; for example, "medical care" = "iryo"

5.2. Model Performance

Table 1 shows the quantitative comparison of models. The data used in this part is the same as Section 5.1. BERTopic achieves the highest coherence score among the models, indicating better topic quality. On the other hand, LDA with processed data demonstrates the lowest computational time, making it the most efficient model. These findings align with the results reported by Grootendorst in their experiment [12]. Based on the observations from Section 5.1 and Section 5.2, we decide to focus on BERTopic for subsequent experiments. Furthermore, to reduce the computational intensity, we narrow down our dataset to focus on the House of Representatives.

	LDA (UP)	LDA (P)	CTM (UP + P)	Top2Vec (UP)	BERTopic (UP)
Coherence Score (NPMI)	-0.029	0.021	-0.019	-0.11	0.09
Wall Time (sec)	254	96	2614	153	487

Table 1: Comparison of model performance

5.3. Hierarchical Clustering

In order to further assess the BERTopic model qualitatively, we perform hierarchical clustering using the Ward linkage method and cosine similarity. The analysis is conducted on data spanning from 2020.01.01 to 2023.04.30, encompassing the COVID-19 era.

The resulting visualization in Fig. 4 displays the top 20 major topics and their hierarchical relationships. Topic labels are made from the top 3 words of each topic. The green clusters at the top represent infrastructure-related topics. The red clusters indicate discussions on the digitalization of schools in response to the school closure during the COVID-19 pandemic. The light blue clusters depict discussions related to the Russian-Ukraine war and its impact on the energy sector. The wine red and yellow clusters are associated with healthcare topics during the pandemic. Finally, the black cluster represents economic-related topics.

It should be noted that a few irrelevant topics ("5_thank you_thank you_Better" and "0_please give me_right_All right") are included within the major topics. This is because the dataset is unprocessed so that it includes speeches about the proceedings. We will remove such topics as necessary in subsequent experiments.

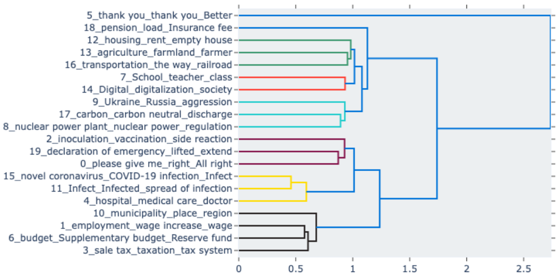

Figure 4: Hierarchy Clustering of top 10 topics

5.4. Committee Label

Our data does not include topic labels, but we can identify which committees the discussions took place in. While the discussions may not always directly align with the committee's purpose, comparing the most discussed topics in each committee can help us assess whether the learning process was conducted effectively. The data used in this analysis is the same as Section 5.3.

Table 2 presents the major committees in the House of Representatives and their corresponding topics extracted from our model. The results align with our expectations; for example, the Health Committee frequently discussed topics related to COVID-19 vaccination, and the Foreign Affairs Committee focused on the Russian-Ukraine war.

Name of Committee	Name of Topic (top 4 words)	Count
Budget Committee	inoculation/vaccination/side reaction/senior citizen	314
Health and Labor Committee	inoculation/vaccination/side reaction/senior citizen	570
Cabinet Committee	employment/wage/wage increase/minimum wage	167
General Affairs Committe	nhk/reception fee/broadcast/chairman	214
Judiciary Committee	trial/judgment/judge/hearing	144
Economy, Trade, and Industry Committee	nuclear power plant/nuclear power/regulation/operation	143
Land, Infrastructure, Transport , and Tourism Committee	housing/rent/empty house/apartment	143
Finance and Financial Affairs Committee	sales tax/taxation/tax system/corporate tax	390
Agriculture, Forestry, and Fisheries Committee	agriculture/farmland/farmer/bearer	227
Foreign Affairs Commitee	Ukraine/Russia/aggression/invasion	173
Others	School/teacher/class/training	478

Table 2: Major topics in committees

5.5. Speaker Clustering

In this section, we create a visualization of each congressman based on the BERTopic topic modeling results (Fig. 5) . The dataset used is the same as described in Section 5.3, with a focus on 60 topics and 200 congressmen. Each congressman is assigned a color based on the committee they have attended the most. By examining this visualization, we can observe the similarities between congressmen, not only within the same committee as expected, but also among those participating in different committees. This allows us to uncover connections that may have previously gone unnoticed.

5.6. Dynamic Topic Modeling

In this part, using Dynamic Topic Modeling with BERTopic, we analyze the speeches of four major political parties in Japan: the Liberal Democratic Party (LDP), New Komeito (NKP), the Democratic Party of Japan (DPJ), and the Japanese Communist Party (JCP). Since the DPJ was dissolved in 2016, we include the Con-

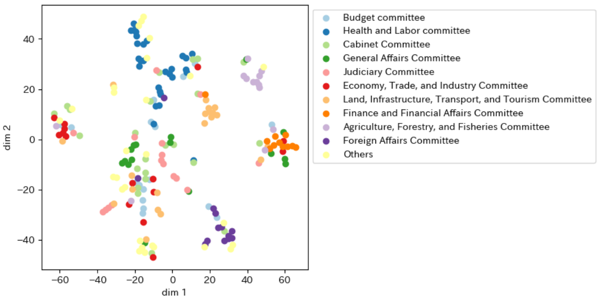

Figure 5: Structure of Congressmen's Ideology

stitutional Democratic Party of Japan as a significant successor within our DPJ dataset.

Fig. 6 illustrates the temporal evolution of the top 10 topics identified by our models. Topic labels are extracted from most frequent word in the topic. The upper-left panel represents the LDP, which has predominantly held power, except for the period from 2009 to 2012 when the DPJ was in control (lower-left panel). Notably, the LDP graph shows a significant spike in the "bank" and "nuclear power" topics in 2011, corresponding to the Great East Japan Earthquake and Fukushima Daiichi nuclear disaster. There is also a peak in the "pension" debate in 2012, when major reforms were being discussed, and an increase in the "infection" topic in 2020, reflecting the impact of the COVID-19 pandemic.

The upper-right panel represents the NKP, which is a ruling coalition partner with the LDP, except for the period from 2009 to 2012. Consequently, we observe a similar pattern of topics as the LDP, indicating their close alignment and shared focus on key issues.

The lower-left represents the DPJ. We observe that it has prioritized "police" and "agriculture" topics over the year. This finding is unexpected because the DPJ was not traditionally associated with a strong emphasis on these topics.

The lower-right panel represents the JCP, which has consistently prioritized the "labor" topic throughout the analyzed period. In 2012, there is a notable peak in the "sales tax" topic, which corresponds to the time when the JCP actively opposed the proposed tax increase.

6. Discussion

The BERTopic model successfully generated coherent and interpretable topics from our Japanese parliament debates dataset. The speaker clustering and dynamic topic modeling techniques provided insights into the political landscape of Japan, revealing both expected and unexpected patterns. These methods have the potential to assist citizens in visualizing and understanding the underlying structure of political debates in Japan.

In future work, it is crucial to enhance our computational capacity to fully leverage the power of BERTopic. Currently, memory limitations restrict us to utilizing only 50% of the available data. Additionally, our analysis of speakers and parties focused solely on topics and did not consider speech sentiment. Incorporating sentiment analysis into the analysis can provide a more nuanced speech clustering.

7. Acknowledgements

We would like to express our gratitude to Professor Karen Livescu and David Yunis for their support throughout the quarter. We are also thankful to our classmates for providing insightful feedback.

In this project, Ayako is responsible for data collection and model selection. Taishi is responsible for model evaluation and visualization. Both Ayako and Taishi collaborate on model training.

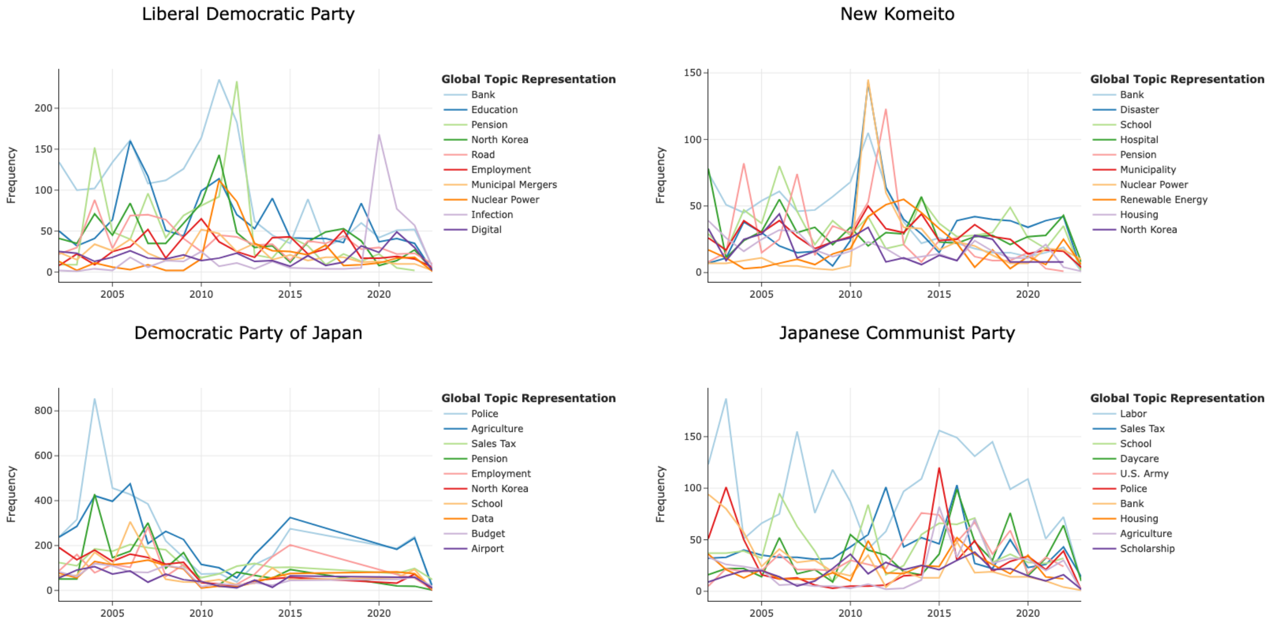

Figure 6: Time series of top 10 topics for major parties

8. References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[2] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National academy of Sciences, vol. 101, no. suppl_1, pp. 5228–5235, 2004.

[3] J. Grimmer, "A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases," Political Analysis, vol. 18, no. 1, pp. 1–35, 2010.

[4] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," Political analysis, vol. 21, no. 3, pp. 267–297, 2013.

[5] C. Lucas, R. A. Nielsen, M. E. Roberts, B. M. Stewart, A. Storer, and D. Tingley, "Computer-assisted text analysis for comparative politics," Political Analysis, vol. 23, no. 2, pp. 254–277, 2015.

[6] T. Sakamoto and H. Takikawa, "Cross-national measurement of polarization in political discourse: Analyzing floor debate in the us the japanese legislatures," in 2017 IEEE international conference on big data (Big Data). IEEE, 2017, pp. 3104–3110.

[7] A. Catalinac and K. Watanabe, "Quantitative text analysis in japanese," Research Bulletin of Waseda Institute for Advanced Study, vol. 11, pp. 133–143, 2019.

[8] "Kokkai api," https://kokkai.ndl.go.jp/api.html, [accessed 20-May-2023].

[9] "mecab-ipadic-neologd : Neologism dictionary for mecab," https://github.com/neologd/mecab-ipadic-neologd, [accessed 20-May-2023].

[10] "Slothlib," http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt, [accessed 20-May-2023].

[11] "Googletrans," https://py-googletrans.readthedocs.io/en/latest/, [accessed 20-May-2023].

[12] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," arXiv preprint arXiv:2203.05794, 2022.

[13] "gensim lda," https://radimrehurek.com/gensim/models/ldamodel.html, [accessed 20-May-2023].

[14] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a hot topic: Contextualized document embeddings improve topic coherence," arXiv preprint arXiv:2004.03974, 2020.

[15] "Cta," https://github.com/MilaNLProc/contextualized-topic-models, [accessed 20-May-2023].

[16] D. Angelov, "Top2vec: Distributed representations of topics," arXiv preprint arXiv:2008.09470, 2020.

[17] "Top2vec," https://top2vec.readthedocs.io/en/stable/index.html, [accessed 20-May-2023].

[18] "Bertopic," https://maartengr.github.io/BERTopic/algorithm/algorithm.html, [accessed 20-May-2023].

[19] S. Terragni, E. Fersini, B. G. Galuzzi, P. Tropeano, and A. Candelieri, "Octis: comparing and optimizing topic models is simple!" in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2021, pp. 263–270.

[20] N. Aletras and M. Stevenson, "Evaluating topic coherence using distributional semantics," in Proceedings of the 10th international conference on computational semantics (IWCS 2013)–Long Papers, 2013, pp. 13–22.

[21] "Octis," https://github.com/MIND-Lab/OCTIS, [accessed 20-May-2023].

[22] "Bertopic_evaluation," https://github.com/MaartenGr/BERTopic_evaluation, [accessed 20-May-2023].