Topic Modeling of Discussions in the Standing Committees of the Brazilian Chamber of Deputies

Matheus A. dos Santos, Nazareno Andrade, Fabio Morais

Universidade Federal de Campina Grande, Brazil
matheusalves@copin.ufcg.edu.br, {nazareno,fabio}@computacao.ufcg.edu.br

Abstract. In order to establish and reinforce democracy, civil society must have the ability to oversee and keep track of the actions of its representatives. Despite substantial progress in transparency, monitoring the committees of the Brazilian National Congress remains a challenge. Primarily, due to the large volume of activities in these committees and the lack of structured data on their discussions. This work provides two contributions to this context. The first is an open dataset containing structured speeches from the 25 standing committees of the Brazilian Chamber of Deputies over the past two decades, which we have created and made available. The second is the application of Natural Language Processing techniques, particularly Latent Dirichlet Allocation (LDA), to identify the topics addressed in these committees. Based on these latent topics, we studied the similarities and differences between the standing committees, their relations, and how their debates changed over time. We also explored how the speeches of parliamentarians from various political parties and states related differently to these latent topics. Our results demonstrate how the topics discussed in the standing committees reverberate external events and show that these committees accommodate conversations between their main topics and opposing agendas.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Text analysis; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: Brazilian Chamber of Deputies, Latent Dirichlet Allocation, Natural Language Processing, Politics

1. INTRODUCTION

According to the *trias politica* principle [de Secondat de Montesquieu et al. 1977], the Legislative Branch is responsible for reviewing and updating the laws that govern the lives of citizens and the operation of the State itself. Two legislative houses compose the Brazilian National Congress: the Federal Senate and the Chamber of Deputies. These houses generally hold debates and make collective decisions in plenaries and thematic committees. A plenary brings together all members of a legislative house and is responsible for making final decisions on some law proposals that require its approval. On the other hand, committees consist of smaller groups of parliamentarians who analyze the technical and legal aspects of law proposals, whose proceedings will always involve at least one committee.

Civil society needs to stay informed about what happens in the plenaries and committees of the National Congress in order to be a part of the legislative process. Although plenaries are responsible for the approval and final modifications of a fair amount of propositions, most of the debates take place in committees, where technical discussions have more space and some events (such as public hearings) may include the participation of non-parliamentary players.

Researchers and monitoring projects have traditionally paid more attention to the plenaries of the Brazilian Chamber of Deputies and Federal Senate when overseeing the Legislative Branch. The availability of open structured data regarding the activities and speeches of parliamentarians, the centralization of debates in a single space, and the greater media attention that plenaries receive are

This work was partially supported by *Paraíba* State Research Foundation (FAPESQ), Grant 07/2021.

some of the reasons that ease studying them. In contrast, there is much less available structured data about the committees and monitoring them requires methods to track dozens of spaces. There is not even open data to analyze committee speeches in the Chamber or in the Senate.

In this context, we present two contributions. The first consists of making available an open dataset on the discussions held by the committees of the Brazilian Chamber of Deputies from 1995 to 2021. This dataset was compiled from unstructured data provided by this legislative house and allows researchers – as well as civil society – to analyze new aspects of the activities of our representatives. The second contribution is the modeling and analysis of patterns and dynamics in the topics debated in the standing committees of the Brazilian Chamber of Deputies. Using the topic modeling results, we investigated how these committees concentrate on specific topics, how their focus on these topics changes over time, and how certain parliamentarians influence these behaviors.

The present article extends upon our previous work [dos Santos et al. 2021] with two main additions. The first is the inclusion of data on speeches in events held by the standing committees of the Brazilian Chamber of Deputies in 2021. We used this data to analyze the results of our Latent Dirichlet Allocation model when given data that was not part of its training dataset. The second addition is the use of our model to explore the deputies' speeches in standing committees' events during the 55th legislature – which spanned from 2015 to 2018 – and how they relate to Brazil's political parties and states.

The following sections are structured as follows. In the next section, we discuss the theoretical background and related work. Section 3 describes how we collected the data used in the topic modeling. Section 4 outlines the topic modeling method, including text preprocessing and the criteria used to evaluate the interpretability of the latent topics. Sections 5, 6, and 7 discuss the topic model by presenting the topics discussed in the standing committees, how they changed over time, and how the speeches of parliamentarians from different political parties and states are related to these latent topics. Finally, Section 8 summarizes the paper's contributions, main conclusions, and future work.

2. BACKGROUND AND RELATED WORK

Topic modeling is a Natural Language Processing task that aims to uncover implicit semantic structures in document collections [Blei and Lafferty 2009]. The Latent Dirichlet Allocation (LDA) [Blei et al. 2003] is one of the most traditional algorithms for this task. This probabilistic model, which has a three-level hierarchy, describes the documents in a corpus as combinations of latent topics present in the collection. Considering topics as probabilistic distributions of a fixed vocabulary of terms, this model assumes that a given number of topics is associated with the corpus and that these topics, in turn, are represented in the documents in different proportions.

Formally, LDA is a latent variable model in which the words in each document are the observable data and the latent variables are the topics and their respective proportions in the documents [Blei and Lafferty 2009]. Based on two probabilistic Dirichlet distributions (denoted by θ and ϕ) controlled, respectively, through the parameters α and β, the model assumes the existence of K topics, M documents, and N words per document. Thus, the n-th word of the m-th document (denoted by w_{nm}) will have its association z_{nm} to the topics defined by the distribution of topics per document θ_n and the distribution of words per topic ϕ_k. By applying this process to every word in the corpus, LDA models the probabilistic relationship between the latent topics and the documents.

In recent years, Natural Language Processing techniques have seen widespread use in political contexts. For instance, LDA was used to classify the law proposals presented to the Brazilian Chamber of Deputies between 1995 and 2014 into seven main topics, supporting the analysis and measurement of the parliamentarians' thematic emphasis [Batista 2020]. Another topic modeling algorithm, the Express Agenda model, was used to identify the latent topics in the Brazilian Chamber of Deputies' plenary speeches and, afterward, to compare the emphasis given by each parliamentarian to social or economic topics during the 1999 to 2014 legislatures [Moreira 2020].

Similar studies have also been conducted in the international scenario. For example, dynamic topic modeling based on two layers of non-negative matrix factorization was employed to analyze the political agenda of the European Parliament from 1999 to 2014. This approach enabled the researchers to track its evolution over time and to identify how internal and external events can impact the plenary speeches of the European parliamentarians [Greene and Cross 2017]. However, we have not identified any studies that use topic modeling to monitor or explore the activities of the Brazilian Chamber of Deputies committees.

Besides Latent Dirichlet Allocation, there are many other algorithms for topic modeling. Some are based on LDA, while others employ different methods to perform this task. Over the past few years, even some Neural Probabilistic Language Models (NPMLs) have arisen in this field. In this work, we opted to use LDA instead of more recent models, such as Topic2Vec [Niu et al. 2015] and BERTopic [Grootendorst 2020]. As our study is exploratory and aims to pave the way for political scientists to use topic analysis in the context of speeches in the Brazilian Chamber of Deputies, we understand that using a more widely known and traditional model – which is also easier to compute and analyze – is a more advisable approach.

3. DATA COLLECTION

The Brazilian Chamber of Deputies website[1] is the channel used by this legislative house to guarantee civil society access to information about its activities. This website publicizes data regarding the Chamber of Deputies' context, such as the law proposals and the plenary speeches. However, there is no open data on transcripts of the events held by the committees. In fact, these transcripts exist and are available on the Brazilian Chamber of Deputies website, but in disagreement with several of the open data principles listed by Open Definition[2]. Moreover, they are only available in HTML format, so monitoring and analyzing these contents through automated methods becomes much more complex.

In order to make this data open and available to the public, we developed the necessary tools to extract and structure it in a non-proprietary format. Our toolset includes two data crawlers: one for the metadata about committee events and another for the transcripts themselves. The extracted metadata allows us to associate each transcribed event with, for instance, a specific date or one of the Brazilian Chamber of Deputies' committees. Since the pages in which this data is available are not well-structured and do not make good use of the HTML semantics, the extraction process needs to extensively use regular expressions to identify the speeches and their speakers.

Overall, we extracted transcripts of 19,339 events held by Brazilian Chamber of Deputies committees between 1995 and 2021. The Internal Regulation of this legislative house states that transcribing committee events is not mandatory and only occurs upon request of a committee president. Therefore, even containing all the transcripts available on the Brazilian Chamber of Deputies, this dataset represents less than 30% of the events held by its committees. The resulting dataset is publicly available at `bit.ly/transcricoes-comissoes`. Furthermore, the toolset for data extraction was developed as open-source code and is available at `github.com/alvesmatheus/fala-camarada`.

4. TOPIC MODELING IN COMMITTEE DISCUSSIONS

The source code required to conduct the experiments described throughout the following sections, from text preprocessing to the analysis of the model results, is available at `github.com/beabaparlamentar/topics-chamber-std-committees`.

[1] `www.camara.leg.br/`
[2] `http://opendefinition.org/`

Table I. Standing committees of the Brazilian Chamber of Deputies in 2022.

Acronym	Committee	Acronym	Committee
CAPADR	Committee on Agriculture, Animal Industry, Supply, and Rural Development	**CFFC**	Committee on Financial Oversight and Control
CC	Committee on Culture	**CFT**	Committee on Finances and Taxation
CCJC	Committee on the Constitution and Justice and Citizenship	**CINDRA**	Committee on National Integration, Regional Development, and the Amazon
CCTCI	Committee on Science and Technology, Communications and Computer Sciences	**CLP**	Committee for Participatory Legislation
CDC	Committee on Consumer Protection	**CMADS**	Committee on the Environment and Sustainable Development
CDEICS	Committee on Economic Development, Commerce and Industry	**CME**	Committee on Mines and Energy
CDHM	Committee on Human Rights and Minorities	**CREDN**	Committee on Foreign Relations and National Defense
CDM	Committee on Defense of the Rights of Women	**CSPCCO**	Committee on Public Security and Fight against Organized Crime
CDPD	Committee on Defense of the Rights of People with Disabilities	**CSSF**	Committee on Social Security and Family
CDPI	Committee on Defense of the Rights of Senior People	**CT**	Committee on Tourism
CDU	Committee on Urban Development	**CTASP**	Committee on Labor, Administration and Civil Service
CEdu	Committee on Education	**CVT**	Committee on Transportation
CEsp	Committee on Sports		

4.1 Corpus and text preprocessing

Typically, the first step in topic modeling is defining the corpus to be analyzed. Our sample was selected based on three characteristics of the events. Firstly, we only considered events held by the standing committees of the Brazilian Chamber of Deputies (as shown in Table I). Moreover, we only included events classified as Debate, Forum, Extraordinary Meeting, Ordinary Meeting, Public Hearing with Guest or Minister, Seminar, or Technical Meeting. This filter excludes events of less relevant debates, such as honors and solemnities. Finally, due to computational constraints, we limited ourselves to events held between 2008 and 2019. Based on these criteria, we selected the transcripts of 4,140 events.

During the data preprocessing stage, we removed the names of the speakers from the documents and then applied capitalization standardization, punctuation removal, tokenization in unigrams and bigrams, stopword removal, and stemming. For the latter, we used the *Removedor de Sufixos da Língua Portuguesa (RSLP)* [Huyck and Orengo 2001], an algorithm specifically designed for Brazilian Portuguese that has a higher accuracy rate compared to traditional stemming algorithms. Finally, we applied the Term Frequency (TF) vectorization, excluding stems that appeared in less than 5% or more than 80% of the documents, in order to reduce the dimensionality of the term-document frequency matrix.

4.2 Finding the topics

The value of K, the number of latent topics associated with the corpus, was determined through experiments. Using 25 standing committees of the Brazilian Chamber of Deputies as a reference, we explored values in the range of 20 to 30 for this parameter. Additionally, we examined the influence of the learning rate decay on the results. To ensure an equal probability among the topics, we set the Dirichlet distribution hyperparameters (denoted by α and β) as the reciprocal of K $(1/K)$. For instance, if $K = 20$, both hyperparameters' values will be 0.05. Furthermore, we used the log-likelihood as the model evaluation metric so that the ideal model is the one that maximizes this function [Blei et al. 2003; Arora and Ravindran 2008].

Figure 1 presents the log-likelihood results for different numbers of topics. As can be observed, the lowest learning rate decay value consistently produced higher likelihood values. Furthermore, there is a decline in log-likelihood when the number of topics exceeds the threshold of $K = 22$. Based on these results, the chosen values for the learning rate decay and the number of topics in this work were 0.60 and 22, respectively. It's important to note that likelihood metrics only evaluate the K-dimensional spaces defined by the models and, therefore, do not evaluate the coherence or interpretability by humans of these latent topics [Chang et al. 2009]. To validate both the coherence and the interpretability of these latent topics, we followed two steps: i) we selected and analyzed the ten stems most related to each topic, and ii) we assigned descriptive labels to the latent topics, as presented in Table II.

Our analysis of the stems suggests that the topics are coherent and often closely related to the

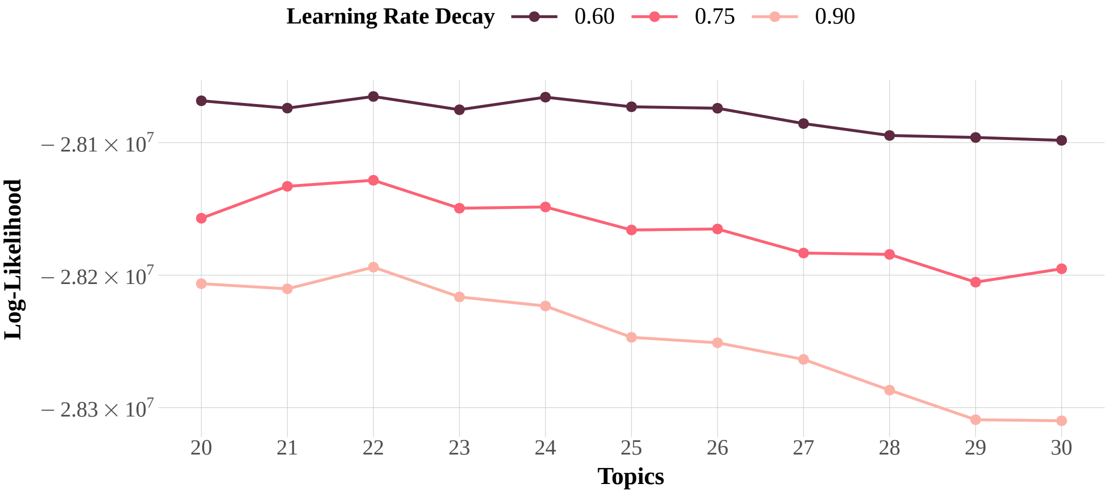

Fig. 1. Log-likelihood results for Latent Dirichlet Allocation models.

Table II. Most associated stems with each latent topic identified by the Latent Dirichlet Allocation model.

Topic	Label	Stems most associated
0	**Investigations**	“document”, “empr”, “animal”, “denúnc”, “mor”, “filh”, “investig”, “jan”, “dinh”, and “min”.
1	**Public Security**	“políci”, “crim”, “polic”, “milit”, “arm”, “seguranç públic”, “guard”, “políci feder”, “penal”, and “justiç’.’
2	**Education and Culture**	“cult”, “educ”, “municípi”, “plan”, “livr”, “orç”, “municip”, “estad”, “plan nacion”, and “emend”.
3	**Taxation**	“previd”, “reform”, “bilhã”, “tribut”, “fiscal”, “receit”, “rend”, “crédit”, “impost”, and “dinh”.
4	**Agriculture**	“produt”, “agricult”, “produç”, “rural”, “aliment”, “agricul”, “produz”, “leit”, “famili”, and “sul”.
5	**Legislation**	“paut”, “it”, “matér”, “parec”, “projet lei”, “retir”, “emend”, “constitucional”, “mérit”, and “propos”.
6	**Sports**	“esport”, “atlet”, “tur”, “jog”, “cop”, “futebol”, “olímp”, “club”, “estádi”, and “confeder”.
7	**Regulation and Control**	“consum”, “empr”, “comunic”, “internet”, “oper”, “agênc”, “regul”, “red”, “merc”, and “preç”.
8	**Amazon**	“amazôn”, “pesc”, “mar”, “forç”, “fronteir”, “arm”, “ama”, “forç arm”, “milit”, and “nort”.
9	**Human Rights**	“human”, “direit human”, “lut”, “mov”, “negr”, “viol”, “mulh”, “companh”, “civil”, and “comunidad”.
10	**Traffic and Transport**	“transport”, “obr”, “aeroport”, “contrat”, “veícul”, “empr”, “invest”, “avi”, “infraestrut”, and “oper”.
11	**Labor**	“empr”, “empreg”, “entidad”, “administr”, “contrat”, “profiss”, “sindicat”, “companh”, “regulament”, and “fiscal”.
12	**Environment**	“ambient”, “ambi”, “florest”, “conserv”, “áre”, “sustent”, “amazôn”, “unidad”, “ibam”, and “desmat”.
13	**Mines and Energy**	“energ”, “empr”, “elétr”, “miner”, “consum”, “distribu”, “min”, “gás”, “usin”, and “invest”.
14	**Health Research**	“defici”, “medic”, “produt”, “pesso defici”, “anvis”, “pesquis”, “drog”, “aliment”, “agrotóx”, and “control”.
15	**Justice**	“tribun”, “justiç”, “supr”, “julg”, “repúbl”, “defend”, “democrac”, “advog”, “parl”, and “judici”.
16	**Industry and Economy**	“empr”, “indústr”, “merc”, “tecnolog”, “invest”, “internac”, “cresc”, “econom”, “produt”, and “unid”.
17	**Social Security**	“mulh”, “crianç”, “viol”, “adolesc”, “famíl”, “filh”, “crianç adolesc”, “sex”, “menin”, and “hom”.
18	**Education**	“educ”, “escol”, “univers”, “profes”, “ensin”, “alun”, “curs”, “profiss”, “prof”, and “superi”.
19	**Health Care**	“médic”, “paci”, “hospit”, “doenç”, “profiss”, “su”, “idos”, “assist”, “diagnóst”, and “medicin”.
20	**Indigenous and *Quilombolas***	“indígen”, “terr”, “comunidad”, “pov”, “índi”, “incr”, “fun”, “assent”, “quilombol”, and “pov indígen”.
21	**Urban Development**	“municípi”, “águ”, “prefeit”, “urban”, “municip”, “nord”, “habit”, “resídu”, “sane”, and “plan”.

themes of the standing committees of the Brazilian Chamber of Deputies. Besides these committee-aligned topics, some focus on more specific aspects. For example, topics 14 and 19 address different aspects of health, while topics 8 and 12 explore environmental themes at distinct levels of granularity. We further assessed coherence and interpretability by reading a sample of documents associated with some of the latent topics, a subjective evaluation that corroborated our previous analysis of the stems.

5. TOPICS DISCUSSED IN THE STANDING COMMITTEES

After evaluating and understanding the latent topics, we explored their relations with the standing committees of the Brazilian Chamber of Deputies. Figure 2 presents the distribution of events for each of the 25 committees in a two-dimensional space created by applying the Uniform Manifold Approximation and Projection (UMAP) [McInnes et al. 2018] on the 22 dimensions of association between the event transcripts and the latent topics identified by the Latent Dirichlet Allocation

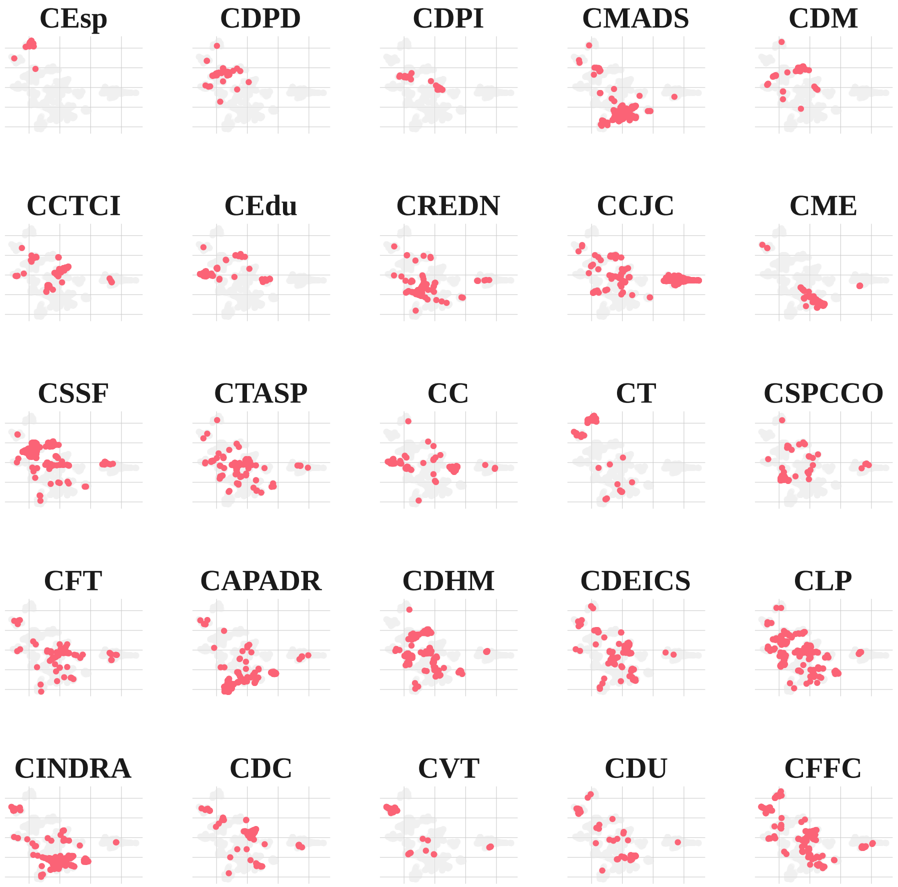

Fig. 2. Distribution of transcripts of the Brazilian Chamber of Deputies' standing committees in a two-dimensional space created by Latent Dirichlet Allocation and Uniform Manifold Approximation and Projection for Dimension Reduction.

model.

This distribution illustrates how most committees concentrate on well-defined regions of the two-dimensional space and, consequently, how they emphasize specific topics in their discussions. Examples include the Committee on Sports (CEsp) and the Committee on Defense of the Rights of Senior People (CDPI). However, it also reveals transcripts that differ from their respective committee regions, which indicates some diversity in the topics discussed. For instance, this diversity is prominent in the Committee for Participatory Legislation (CLP). Moreover, we observed substantial overlap in the transcripts of committees with intersecting themes, such as the Committee on the Environment and Sustainable Development (CMADS), the Committee on National Integration, Regional Development, and the Amazon (CINDRA), and the Committee on Agriculture, Animal Industry, Supply, and Rural Development (CAPADR).

Figure 3 presents the average associations between each standing committee of the Brazilian Chamber of Deputies and all the 22 latent topics identified by the LDA model. These associations support the hypothesis that each standing committee primarily discusses just a small number of topics, typically between 2 and 4. The only exception to this pattern is the Committee on Participatory

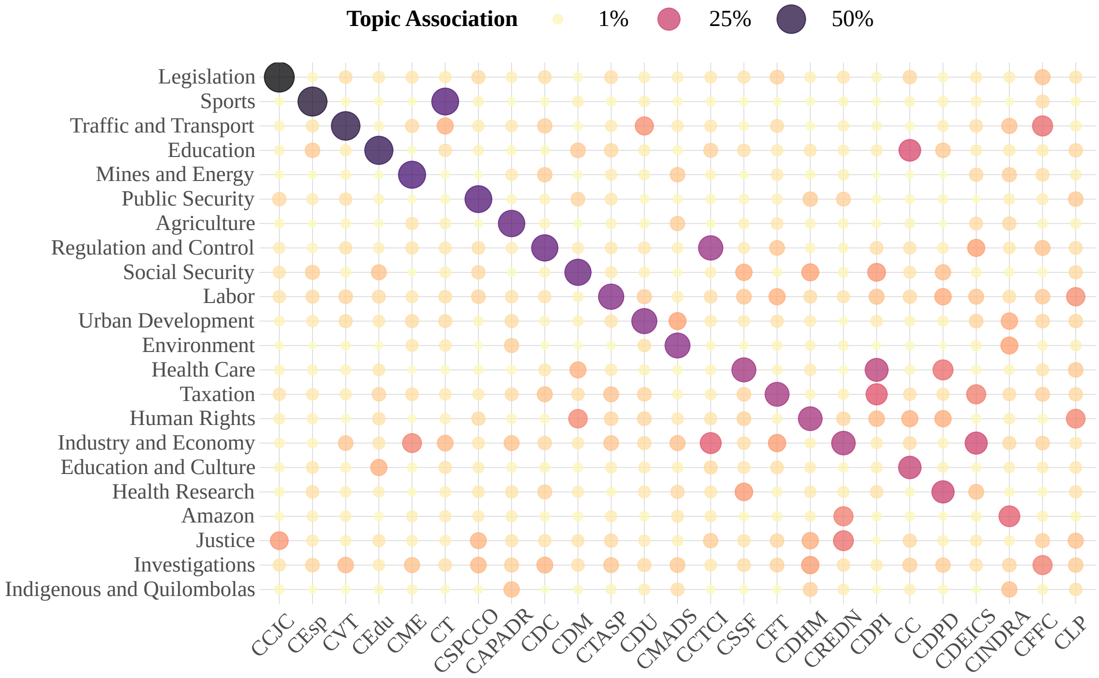

Fig. 3. Average association between the standing committees of the Brazilian Chamber of Deputies and the latent topics identified by Latent Dirichlet Allocation in the period from 2008 to 2019.

Legislation (CLP). Unlike the other committees, the CLP is responsible for receiving and discussing proposals from civil society for the Chamber. As a result, while promoting society's participation in the legislative process, this committee also encounters the need to discuss a wide variety of topics.

Profiling the standing committees allows us to observe more subtle dynamics within the Brazilian Chamber of Deputies. In Brazil, groups related to the agricultural and mining sectors often come into conflict with environmentalists and, sometimes, with human rights activists. The committee-topic associations we found reveal that this tension also occurs within the committees. In the Committee on the Environment and Sustainable Development (CMADS), the topics of Agriculture (topic 4) and Mines and Energy (topic 13) seem to be frequently discussed. Similarly, in the Committee on Agriculture, Animal Industry, Supply, and Rural Development (CAPADR), the topics Environment (topic 12) and Indigenous and *Quilombolas* (topic 20) are regularly present in the discussions. This emphasis on subjects that commonly oppose their primary themes suggests that the committees work not only as a forum for discussion of these primary themes, but also as a venue for conflicts and debates involving interests that antagonize those typically expected from them.

To assess the performance of our LDA model on transcripts that were not in its training dataset, we applied the model to the transcripts from 2020 and 2021 available on the Brazilian Chamber of Deputies website. The standing committees suspended their activities and did not release transcripts in 2020 due to the COVID-19 pandemic. Thus, our unseen dataset consists of 238 transcripts from events held by these committees in 2021. It is worth mentioning that eight standing committees did not publish transcripts that year and that, between 2008 and 2019, these committees made an average of 345 transcripts available per year (nearly 45% more documents). The exact text preprocessing operations described earlier in this article were applied to these 238 transcripts before being given to

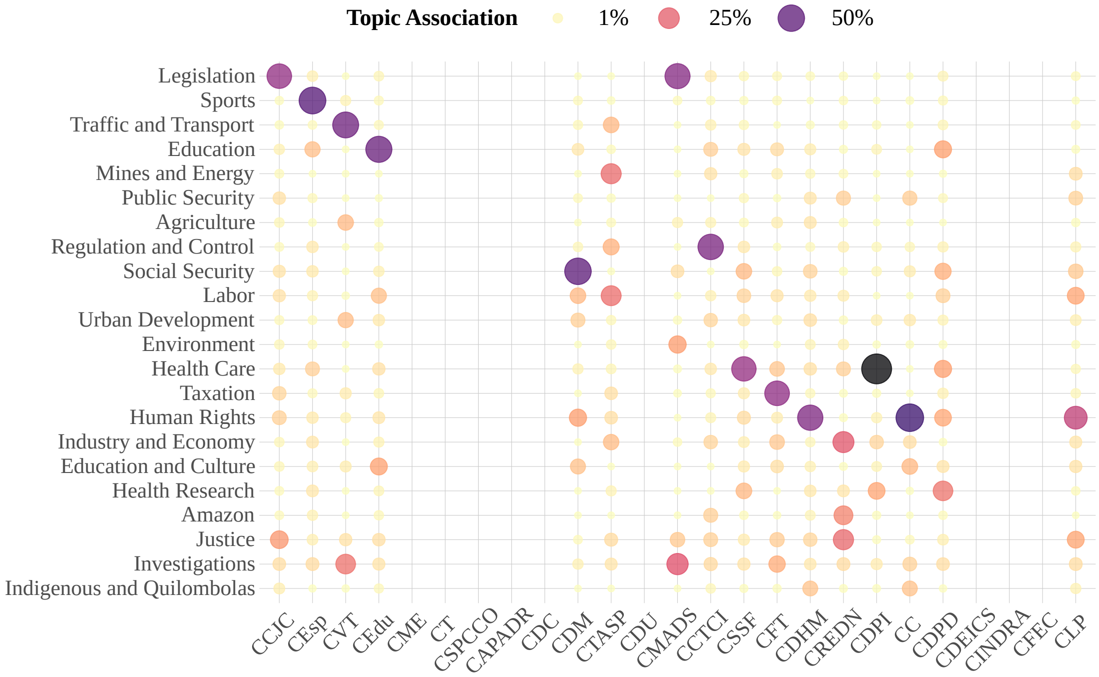

Fig. 4. Average association between the standing committees of the Brazilian Chamber of Deputies and the latent topics identified by Latent Dirichlet Allocation in 2021.

our model. Figure 4 presents the average associations in 2021 between the standing committees of the Brazilian Chamber of Deputies and the latent topics identified by the LDA model.

Overall, the committee-topic associations in the 2020-2021 biennium were similar to the long-term average from 2008 to 2019. Even with some slight variations in the vocabulary of the parliamentarians, this suggests that our model can still be applied to future time periods. One notable difference was the higher association of the Committee on Environment and Sustainable Development (CMADS) with the topics of Investigations (topic 0) and Legislation (topic 5), which aligns with the increased public interest and debates on environmental disasters and legislation changes in Brazil during this period. Furthermore, the Committee on Defense of the Rights of Women (CDM) showed a higher association with the topics of Labor (topic 11) and Urban Development (topic 21), reflecting the increased attention to these topics in discussions about women's rights in comparison to the 2000s decade.

6. TOPICS OVER TIME

Based on the insights of a political expert, we analyzed the evolution of topics over time from two perspectives. First, Figure 5 presents the average topic association for all events in months with at least ten events held in committees. We noticed that Social Security (topic 17), Human Rights (topic 9), and Health Care (topic 19) have become more prominent throughout the analyzed period. On the other hand, Labor (topic 11) and Industry and Economy (topic 16) have become less prevalent in committee discussions. This dynamic suggests that the Brazilian Chamber of Deputies has dedicated more attention to social issues as they have gained greater importance in Brazil's public debate.

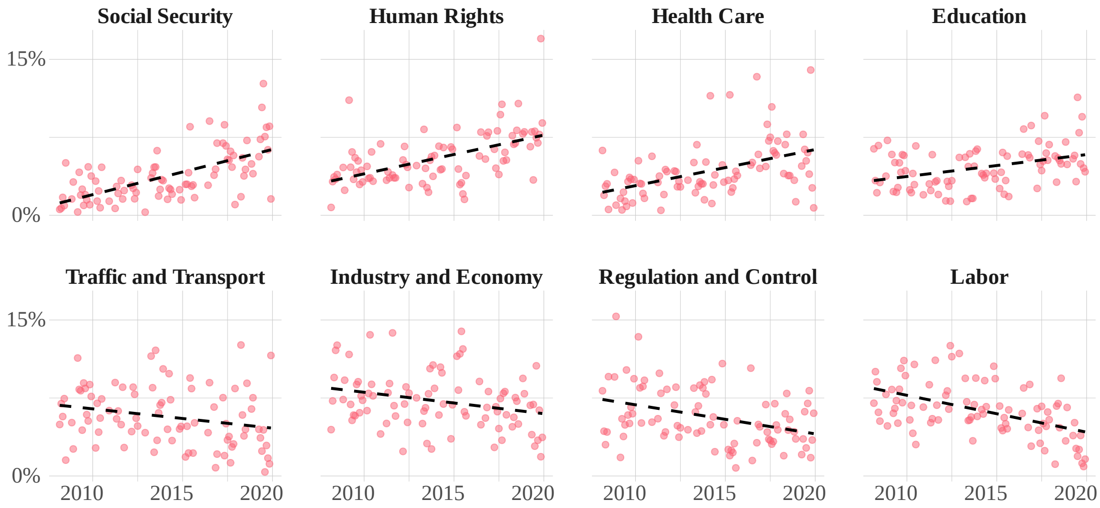

Fig. 5. Monthly association between events of the standing committees and the latent topics identified by the Latent Dirichlet Allocation model. The dashed line represents a linear regression of the association with time, and the eight topics with higher absolute coefficients in this regression are displayed. Also, only months with at least ten events held were included.

We also analyzed the evolution of latent topics from the perspective of the main committee of the Brazilian Chamber of Deputies: the Committee on the Constitution and Justice and Citizenship (CCJC). Figure 6 presents the monthly committee-topic association for the ten most discussed topics in this committee between 2015 and 2019. Over these five years, the CCJC remained predominantly associated with Legislation (topic 5), as expected given its responsibilities. However, in certain months, its discussions placed great emphasis on other latent topics, particularly Justice (topic 15).

After reviewing transcripts and news articles from the time, we discovered that the increase in discussions about Justice (topic 15) in the early second half of 2016 was related to the impeachment process of President Dilma Rousseff. Based on a charge of *"crime de responsabilidade"*, this process fell outside the purview of the CCJC and had not previously impacted its discussions. However, this changed on the eve of the final votes. Of the five transcripts made available by the CCJC in July 2016, only one did not mention the impeachment process. In general, these mentions occurred in contexts that were relevant to the events but led parliamentarians to engage in discussions about the merits and motivations of the process. It is also worth mentioning that in the month right after the end of the impeachment process (October 2016), the discussions held by the CCJC went back to being associated almost exclusively with the Legislation (topic 5).

In turn, 2017 was marked by two peaks of the CCJC's association with Justice (topic 15) in July and October. These months immediately followed complaints from the Brazilian Federal Prosecution Office against President Michel Temer. Both complaints were related to common crimes – such as passive corruption – rather than political crimes and required approval from the CCJC to proceed. Although the committee rejected them, the complaints still had some impact on its discussions. Unlike the previous year, the increase and decrease of Justice (topic 15) were gradual and related to other complaints of the *Lava Jato* Operation, which were also widely discussed in the CCJC during this period.

Finally, the monthly committee-topic associations demonstrate how unusual the CCJC discussions were during 2019. That year, the association with Justice (topic 15) did not have peaks related

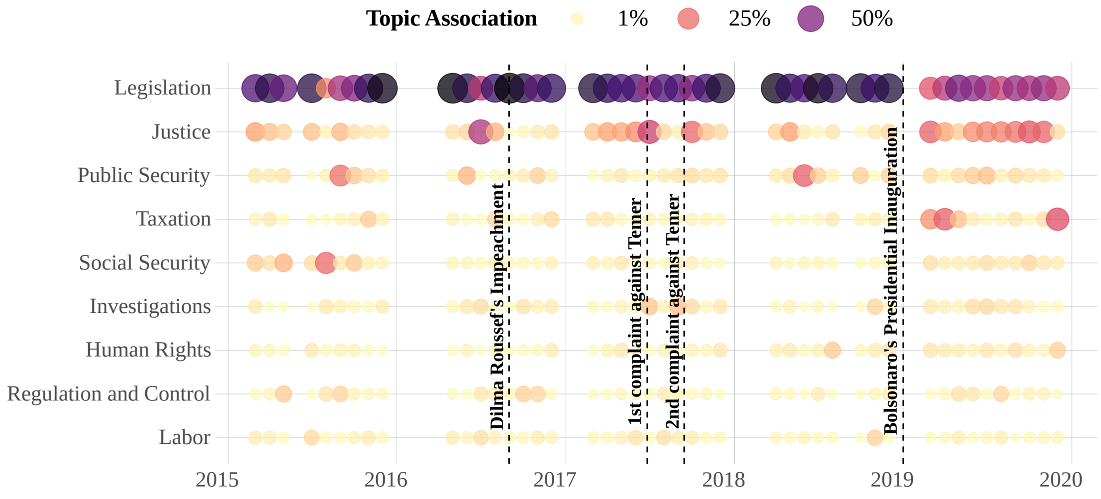

Fig. 6. Most associated topics, month-on-month, with the Committee on the Constitution and Justice and Citizenship (CCJC) between 2015 and 2019.

to specific events but still remained high throughout the entire year. The transcripts reveal that two factors contributed to this long-term shift in the committee's focus. First, Jair Bolsonaro's government frequently used presidential decrees to alter legislation, which were commented on and sometimes overturned by CCJC parliamentarians. Examples include the decree related to loosening gun ownership rules, which was overturned in June 2019. The second factor that impacted the topic-committee association for this committee was the multiple public hearings to discuss Proposed Constitutional Amendment 410/2018 (PEC 410/2018), which would allow second-instance convictions.

7. TOPICS IN POLITICAL PARTIES AND STATES OF ORIGIN

Along with studying the standing committees, we analyzed the political parties and the Brazilian federative units (states) using the speeches of their elected deputies. Our analysis included 1,168 transcripts of events held by the standing committees of the Brazilian Chamber of Deputies during the 55th legislature (2015-2018). We focused on speeches given by deputies who attended at least five events of a single committee. Also, we omitted transcripts from three committees that address general matters: the Committee on the Constitution and Justice and Citizenship (CCJC), the Committee on Finances and Taxation (CFT), and the Committee for Participatory Legislation (CLP). The average associations between deputies from each Brazilian state and the 22 latent topics identified by the LDA model are shown in Figure 7.

Similar to the standing committees, each federative unit exhibits high association with just a few latent topics, typically between 1 and 3. Moreover, the main interests of the Brazilian states seem to be well-represented in the speeches of their elected deputies. For instance, states in the North region of Brazil that are close to the Amazon – such as *Rondônia* (RO), *Pará* (PA) and *Amazonas* (AM) – show higher associations with Amazon (topic 8). On the other hand, the more general and broad topic of Environment (topic 12) has higher associations with states from other regions, such as *Maranhão* (MA) and *Piauí* (PI). It is also noteworthy that deputies elected by states from the South region of Brazil – *Mato Grosso do Sul* (MS), *Paraná* (PR) and *Santa Catarina* (SC) – have Agriculture (topic 4) as the primary subject of their speeches, which can be attributed to the strong connection between the economies of these federative units and the agricultural and cattle-raising sectors.

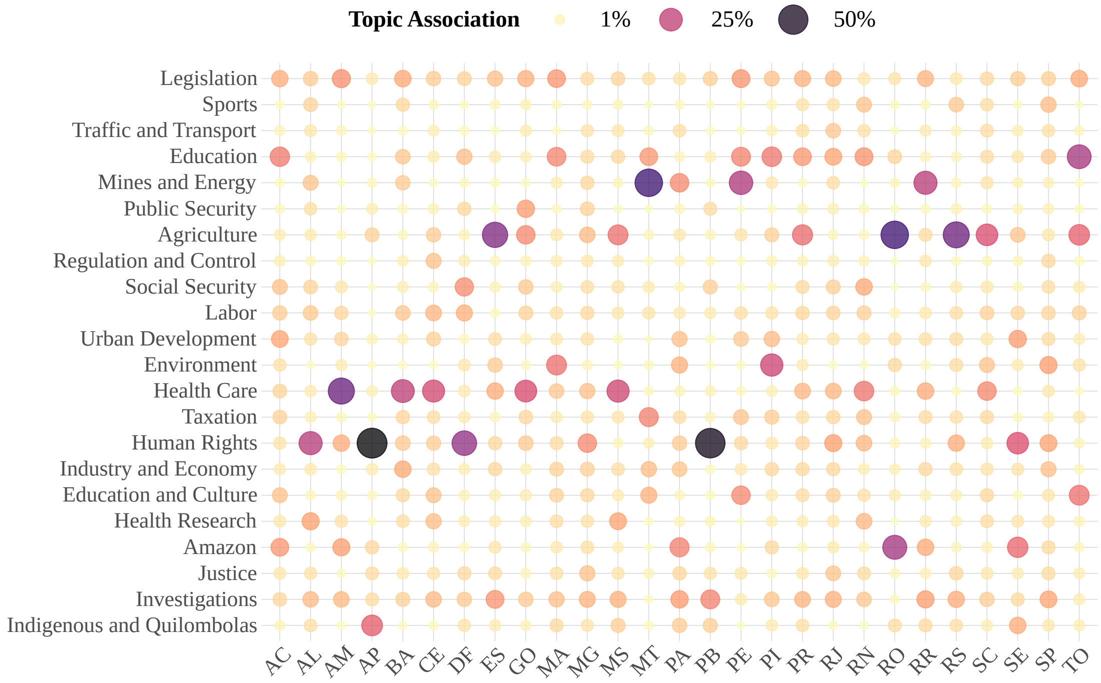

Fig. 7. Average association between the Brazilian states and the latent topics identified by Latent Dirichlet Allocation during the 55th legislature of the Brazilian Chamber of Deputies (2015-2018).

In turn, Figure 8 presents the average associations between political parties and the 22 latent topics identified by the LDA model. In contrast to the topic-state associations, political parties devote more space in their discussions to fewer topics. For example, higher associations with Agriculture (topic 4) are observed in political parties – such as MDB, PP, and SD – that compose a significant part of the so-called ruralist caucus: a group of parliamentarians who advocate for the agribusiness sector, often at the expense of the environment.

From a different perspective, right-wing parties with anti-environmentalist positions – such as PSC and PSL – show high association with Amazon (topic 8) but little or no association with Public Security (topic 1). Although this subject is one of the main political agendas of these parties, the highest association with Public Security (topic 1) was from a center-left party (PDT) during the period we analyzed. Similarly, it would be expected that Industry and Economy (topic 16) was more frequent in the speeches of deputies from center-right parties. However, this topic exhibits its highest average association with PCdoB, a left-wing party. These findings also support the idea that committees serve as spaces for conflict and debate between parliamentarians. Furthermore, they also indicate how our topic modeling approach can be employed for exploring and analyzing the activities of parliamentarians.

8. CONCLUSIONS AND FUTURE WORK

In this article, we explored the activities of the Brazilian Chamber of Deputies' standing committees from 2008 to 2019 based on the speeches of parliamentarians. The documents for this study were not initially available in an open structured format, so we had to extract and structure them. The resulting dataset includes 19,339 documents and describes the committees' discussions over the past

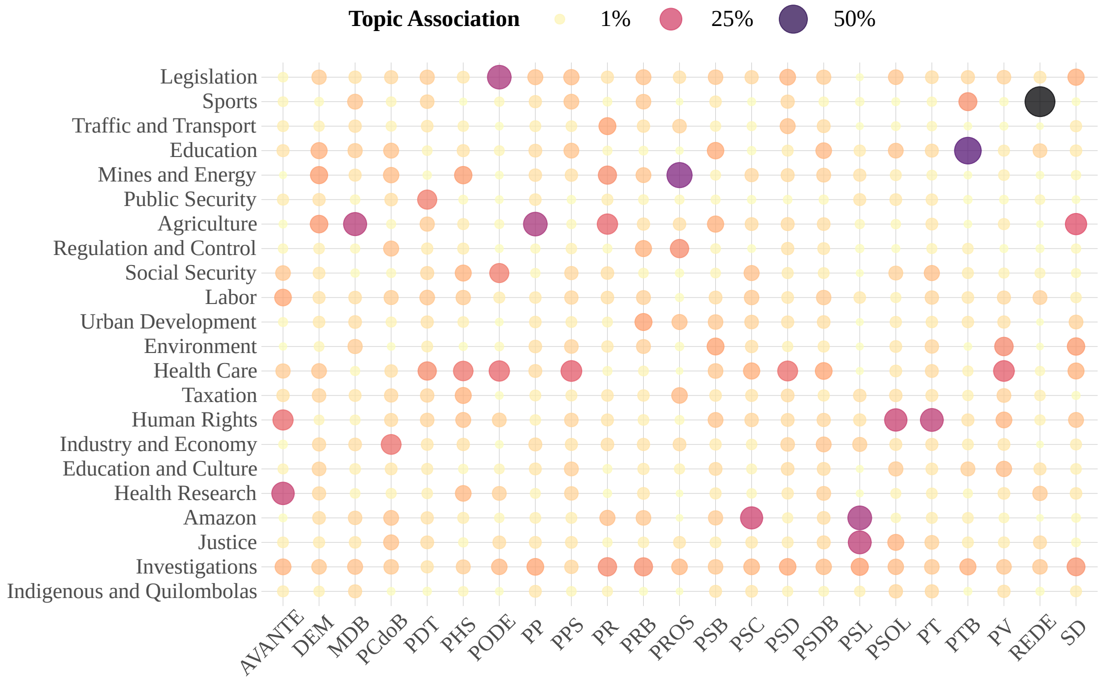

Fig. 8. Average association of the political parties and the latent topics identified by Latent Dirichlet Allocation during the 55th legislature of the Brazilian Chamber of Deputies (2015-2018).

two decades. Our topic modeling using Latent Dirichlet Allocation achieved good levels of likelihood, coherence, and interpretability. By linking the transcripts texts to the latent topics, we were able to identify characteristics and relations between the different committees and explore changes in their discussions over time.

In addition to providing insights into the Brazilian Chamber of Deputies, this approach points to a promising direction for further research and the social control of Brazilian parliamentary activities. In future studies, our Latent Dirichlet Allocation model (or a similarly trained model) can be used to analyze speeches in other spaces of the Brazilian Legislative Branch, such as temporary committees or the plenaries of the Chamber of Deputies or even of the Federal Senate. Furthermore, there is also potential for applying state-of-art topic modeling algorithms – such as Topic2Vec and BERTopic – on the dataset made available alongside this article.

REFERENCES

Arora, R. and Ravindran, B. Latent Dirichlet Allocation Based Multi-Document Summarization. In *Proceedings of the Second Workshop on Analytics for Noisy Unstructured Text Data*. Association for Computing Machinery, Singapore, pp. 91–97, 2008.

Batista, M. QUAIS POLÍTICAS IMPORTAM? Usando ênfases na agenda legislativa para mensurar saliência. *Revista Brasileira de Ciências Sociais* 35 (104): 1–20, 2020.

Blei, D. M. and Lafferty, J. D. Topic Models. In A. N. Srivastava and M. Sahami (Eds.), *Text Mining: Classification, Clustering, and Applications*. Chapman and Hall/CRC, New York, pp. 71–93, 2009.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3 (18): 993–1022, 2003.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. Reading Tea Leaves: How Humans

Interpret Topic Models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, New York, pp. 288–296, 2009.

de Secondat de Montesquieu, C.-L., Carrithers, D. W., and Nugent, T. *The Spirit of the Laws*. University of California Press, Berkeley, 1977.

dos Santos, M. A., Andrade, N., and Morais, F. Topic Modeling of Committee Discussions in the Brazilian Chamber of Deputies. In *Anais do IX Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2021)*. Sociedade Brasileira de Computação - SBC, Brazil, pp. 49–56, 2021.

Greene, D. and Cross, J. P. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis* 25 (1): 77–94, 2017.

Grootendorst, M. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics., 2020.

Huyck, C. and Orengo, V. M. A Stemming Algorithmm for the Portuguese Language. In *International Symposium on String Processing and Information Retrieval*. IEEE Computer Society, California, pp. 186–193, 2001.

McInnes, L., Healy, J., Saul, N., and Grossberger, L. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* 3 (29): 861, 2018.

Moreira, D. Com a Palavra os Nobres Deputados: Ênfase Temática dos Discursos dos Parlamentares Brasileiros. *Dados* 63 (1): 1–37, 2020.

Niu, L., Dai, X., Zhang, J., and Chen, J. Topic2Vec: Learning distributed representations of topics. In *2015 International Conference on Asian Language Processing (IALP)*. IEEE, Suzhou, China, pp. 193–196, 2015.