Topic Modeling of Committee Discussions in the Brazilian Chamber of Deputies

M. A. dos Santos, N. Andrade, F. Morais

Universidade Federal de Campina Grande, Brazil
matheus.santos@ccc.ufcg.edu.br, {nazareno,fabio}@computacao.ufcg.edu.br

Abstract. Ensuring that civil society can monitor and supervise the actions of its representatives is essential to build strong democracies. Despite significant advances in transparency, Brazilian National Congress committees are presently complex to follow and monitor due to the lack of open structured data about their discussions and the sheer volume of activity in these committees. This work presents two contributions to this context. First, we create and present an open dataset including structured speeches of the 25 Chamber of Deputies' standing committees over the last two decades. Second, we use Natural Language Processing techniques – especially Latent Dirichlet Allocation (LDA) – to identify themes addressed on these committees. Based on these latent topics, we explore similarities and differences between the standing committees, their relationships, and how their debates change over time. Our results show that committees accommodate conversations – including their main topic and opposing agendas – and describe how the topics discussed in the committees reverberate external events.

CCS Concepts: • **Computing methodologies** → **Natural language processing**.

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

1. INTRODUÇÃO

Segundo o princípio da *trias politica* [de Secondat de Montesquieu et al. 1977], o Legislativo é o poder do Estado responsável pela revisão do ordenamento jurídico que rege a vida das pessoas e o funcionamento do Estado. No Brasil, o Congresso Nacional é composto pelo Senado Federal e pela Câmara dos Deputados. De maneira geral, o debate e as decisões coletivas de ambas as casas acontecem em um plenário e em comissões temáticas. O plenário congrega todos os parlamentares de uma casa e é responsável pelas decisões finais a respeito das proposições cuja tramitação exige aprovação nesse espaço. Em contraste, as comissões são grupos menores de parlamentares que analisam aspectos técnicos e legais das proposições de lei, cuja tramitação sempre envolverá ao menos uma comissão.

Para participar do processo legislativo, a sociedade civil precisa acompanhar o que acontece no plenário e nas comissões do Congresso Nacional. Ainda que o plenário seja responsável pela aprovação e por alterações finais em um volume considerável de proposições, é nas comissões que acontece a maior parte dos debates, onde as discussões técnicas têm mais espaço e onde alguns dos eventos – como as audiências públicas – podem incluir a participação de atores não parlamentares.

O acompanhamento dos plenários da Câmara dos Deputados e do Senado Federal tem tido, historicamente, mais atenção de pesquisadores e de projetos de monitoramento do Poder Legislativo. Entre os fatores que facilitam o acompanhamento do plenário estão a disponibilização de dados abertos estruturados sobre ações e discursos dos parlamentares, a centralização dos debates em um só espaço e a maior atenção que o plenário tipicamente recebe da mídia. Em contrapartida, o acompanhamento das comissões necessita de métodos que observem dezenas de espaços e tem muito menos dados estruturados disponíveis. Afinal, não há dados abertos na Câmara ou no Senado que permitam, sequer, a análise das falas de suas comissões.

Nesse contexto, este trabalho apresenta duas contribuições principais. A primeira é a modelagem e análise de padrões e dinâmicas nos temas debatidos nas comissões permanentes da Câmara dos

Deputados. Através dos resultados da modelagem de tópicos, examinamos como as comissões se atêm a tópicos específicos e como sua atenção a esses tópicos varia ao longo do tempo. A segunda consiste na publicação de um conjunto de dados abertos referente às discussões realizadas pelas comissões dessa casa legislativa no período de 1995 a 2020. Esse conjunto de dados foi construído a partir dos dados não estruturados disponibilizados pela Câmara dos Deputados e permite que pesquisadores – e a sociedade civil em geral – possam examinar novos aspectos da atuação de nossos representantes.

2. CONTEXTO E TRABALHOS RELACIONADOS

A modelagem de tópicos é uma tarefa de Processamento de Linguagem Natural que visa identificar estruturas semânticas implícitas em coleções de documentos [Blei and Lafferty 2009]. Entre os algoritmos dedicados a essa tarefa está o *Latent Dirichlet Allocation* (LDA), um modelo generativo probabilístico que utiliza uma hierarquia de três níveis para descrever os documentos de um *corpus* como combinações finitas de tópicos implícitos a essa coleção [Blei et al. 2003]. Considerando os tópicos como distribuições probabilísticas de um vocabulário fixo de termos, esse modelo assume que um dado número de tópicos está associado ao *corpus* e que esses tópicos, por sua vez, estão representados nos documentos em diferentes proporções.

Formalmente, o LDA é um modelo de variáveis latentes em que as palavras de cada documento são os dados observáveis e as variáveis latentes são os tópicos e suas respectivas proporções nos documentos [Blei and Lafferty 2009]. Considerando θ e ϕ como distribuições probabilísticas de Dirichlet controladas, respectivamente, através dos parâmetros α e β, são definidos K tópicos, M documentos e N palavras por documento. Desse modo, a n-ésima palavra do documento m (denominada w_{nm}) terá sua associação z_{nm} aos tópicos definida a partir da distribuição θ_n de tópicos por documento e da distribuição ϕ_k de palavras por tópico. Realizando esse processo para todas as palavras do *corpus*, o LDA modela a relação probabilística dos temas implícitos com os documentos.

As técnicas de Processamento de Linguagem Natural têm sido consistentemente aplicadas no âmbito político nos últimos anos. Num contexto semelhante ao deste trabalho, o LDA já foi adotado para classificar as proposições de lei apresentadas à Câmara dos Deputados entre 1995 e 2014 em sete temas principais, dando suporte à exploração e à mensuração das ênfases temáticas dos parlamentares da época [Batista 2020]. O *Express Agenda Model*, outro algoritmo para modelagem de tópicos, também já foi utilizado para identificar os temas de pronunciamentos no plenário da Câmara dos Deputados e para, posteriormente, contrastar a ênfase dedicada por cada parlamentar aos temas sociais ou econômicos durante as legislaturas que se estendem de 1999 a 2014 [Moreira 2020].

Estudos semelhantes também têm sido conduzidos no cenário internacional. Considerando o período de 1999 a 2014, a agenda política do Parlamento Europeu foi examinada através da modelagem dinâmica de tópicos baseada em duas camadas de fatoração de matrizes não negativas. Dessa forma, foi possível não somente acompanhar sua evolução ao longo do tempo, mas também identificar os impactos de eventos internos e externos sobre os discursos em plenário dos parlamentares europeus [Greene and Cross 2017]. Não identificamos, contudo, quaisquer trabalhos que se dediquem a acompanhar e/ou explorar as atividades das comissões da Câmara dos Deputados brasileira através da modelagem de tópicos.

3. COLETA DOS DADOS

O Portal da Câmara dos Deputados[1] é o meio utilizado por essa casa legislativa para garantir o acesso da sociedade civil às informações sobre suas atividades. É através dele que são publicados dados abertos referentes ao contexto da Câmara dos Deputados como, por exemplo, as proposições de lei tramitadas e a íntegra dos discursos realizados em plenário. Não estão presentes, no entanto, dados

[1] `https://www.camara.leg.br/`

abertos de transcrições dos eventos realizados pelas comissões. Esses registros taquigráficos existem e são disponibilizados pelo Portal da Câmara dos Deputados, mas em desacordo com diversos princípios de dados abertos, elencados pela *Open Definition*[2]. Ademais, por serem disponibilizados em formato HTML, a automação do acompanhamento e da análise desses conteúdos se torna mais complexa.

Visando contribuir com a libertação desses dados, implementamos o ferramental necessário para extraí-los, estruturá-los em formato não proprietário e, posteriormente, disponibilizá-los ao público. Esse ferramental é composto por dois raspadores de dados (*crawlers*): um dedicado aos metadados de eventos das comissões e outro às transcrições em si. A extração desse conjunto de metadados permite que cada evento transcrito possa ser associado, por exemplo, a uma das comissões da Câmara dos Deputados ou a uma data específica. A raspagem dos textos das transcrições faz amplo uso de expressões regulares para identificar as falas e seus autores, uma vez que as páginas em que esses dados estão disponíveis não são claramente estruturadas e tampouco fazem bom uso da semântica HTML.

Ao todo, foram extraídas transcrições de 18.839 eventos realizados por comissões da Câmara dos Deputados entre 1995 e 2020. Segundo Regimento Interno da Câmara dos Deputados, o registro taquigráfico dos eventos das comissões não é obrigatório e ocorre apenas sob determinação de seus respectivos presidentes. Assim, ainda que possua todas as transcrições disponibilizadas pelo Portal da Câmara dos Deputados, esse conjunto de dados se refere a cerca de 30% dos eventos realizados pelas comissões. A base de dados construída nesse processo está publicamente disponível em `https://bit.ly/transcricoes-comissoes`. Ainda, o ferramental para a libertação dos dados foi desenvolvido em formato de código aberto e está disponível em `https://github.com/alvesmatheus/fala-camarada`.

4. TÓPICOS NAS DISCUSSÕES DAS COMISSÕES

4.1 *Corpus* e pré-processamento do texto

O primeiro passo para a modelagem dos tópicos é, tipicamente, a definição do *corpus* a ser adotado. Nosso recorte para análise se baseia em três características dos eventos. Primeiro, focamos em eventos das comissões permanentes da Câmara dos Deputados (apresentadas na Tabela I). Também nos restringimos a eventos categorizados como Audiência Pública com Convidado(a) ou Ministro(a), Debate, Fórum, Reunião Extraordinária, Reunião Ordinária, Reunião Técnica ou Seminário. Esse filtro exclui eventos de debate pouco relevante, como homenagens e solenidades. Por fim, considerando os recursos computacionais disponíveis, limitamo-nos aos eventos realizados no período de 2008 a 2019. A partir dessas características, selecionamos as notas taquigráficas de 4.140 eventos.

Tabela I. Comissões permanentes da Câmara dos Deputados em 2021.

Sigla	Comissão	Sigla	Comissão
CAPADR	Comissão de Agricultura, Pecuária, Abastecimento e Desenvolvimento Rural	**CFFC**	Comissão de Fiscalização Financeira e Controle
CC	Comissão de Cultura	**CFT**	Comissão de Finanças e Tributação
CCJC	Comissão de Constituição e Justiça e de Cidadania	**CINDRA**	Comissão de Integração Nacional, Desenvolvimento Regional e Amazônia
CCTCI	Comissão de Ciência e Tecnologia, Comunicação e Informática	**CLP**	Comissão de Legislação Participativa
CDC	Comissão de Defesa do Consumidor	**CMADS**	Comissão de Meio Ambiente e Desenvolvimento Sustentável
CDEICS	Comissão de Desenvolvimento Econômico, Indústria, Comércio e Serviços	**CME**	Comissão de Minas e Energia
CDHM	Comissão de Direitos Humanos e Minorias	**CREDN**	Comissão de Relações Exteriores e de Defesa Nacional
CDM	Comissão dos Direitos da Mulher	**CSPCCO**	Comissão de Segurança Pública e Combate ao Crime Organizado
CDPD	Comissão dos Direitos das Pessoas com Deficiência	**CSSF**	Comissão de Seguridade Social e Família
CDPI	Comissão dos Direitos da Pessoa Idosa	**CT**	Comissão de Turismo
CDU	Comissão de Desenvolvimento Urbano	**CTASP**	Comissão de Trabalho, Administração e Serviço Público
Cedu	Comissão de Educação	**CVT**	Comissão de Viação e Transportes
Cesp	Comissão do Esporte		

Durante o pré-processamento dos dados, removemos os nomes de oradores dos documentos e aplicamos operações de padronização de capitalização, remoção da pontuação, tokenização em unigramas e bigramas, remoção de *stopwords* e stemização. Nesta última, adotamos o Removedor de Sufixos da Língua Portuguesa (RSLP) [Huyck and Orengo 2001], um algoritmo desenvolvido especificamente

[2] `http://opendefinition.org/`

para o português brasileiro e com taxa de acerto superior à de algoritmos tradicionais. Por fim, aplicamos a vetorização *Term Frequency* (TF) desconsiderando os *stems* presentes em menos de 5% ou em mais de 80% dos documentos, de modo a reduzir a dimensionalidade da matriz de frequências termo-documento gerada.

4.2 Definição dos tópicos

O número K de tópicos associados ao *corpus* foi determinado experimentalmente. Tomando a quantidade de 25 comissões permanentes na Câmara dos Deputados como referência, investigamos valores no intervalo de 20 a 30 para esse parâmetro. De forma semelhante, também avaliamos o impacto do decaimento da taxa de aprendizado nos resultados obtidos. Além disso, os parâmetros de controle das distribuições Dirichlet (denotados por α e β) foram fixados com valor inverso a K, tornando os tópicos equiprováveis. Se $K = 20$, por exemplo, essas variáveis assumem o valor 0,05. Ademais, adotamos a log-verossimilhança como métrica de avaliação para os modelos produzidos, de forma que o modelo ideal é aquele que maximiza o valor dessa função [Blei et al. 2003][Arora and Ravindran 2008].

A Figura 1 apresenta os resultados do log-verossimilhança para diferentes números de tópicos. É possível observar que o menor valor de decaimento da taxa de aprendizado foi aquele que esteve, consistentemente, associado às maiores verossimilhanças. Ainda, há uma tendência de perda da verossimilhança a partir do limiar $K = 22$. Dessa forma, os valores do decaimento da taxa de aprendizado e do número de tópicos adotados neste trabalho foram de 0,60 e 22, respectivamente. Vale ressaltar, no entanto, que as métricas de verossimilhança avaliam exclusivamente os espaços K-dimensionais definidos pelos modelos e que, portanto, não permitem qualificar a coerência dos temas ou sua interpretabilidade por humanos [Chang et al. 2009].

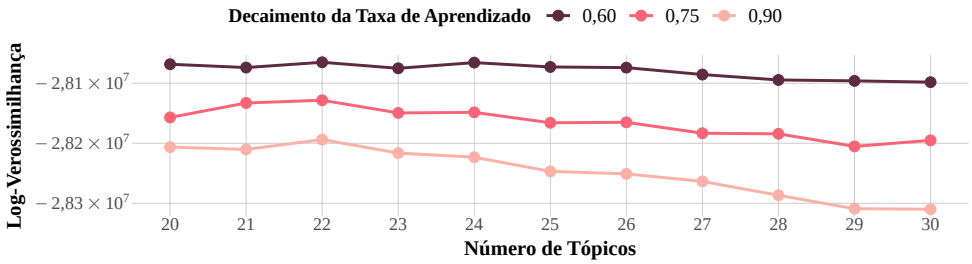

Fig. 1. Log-Verossimilhança dos modelos *Latent Dirichlet Allocation* produzidos.

A validação da coerência e da interpretabilidade dos tópicos foi feita em duas etapas. Primeiro foram selecionados e analisados os 10 *stems* mais associados a cada tópico e, a partir deles, definimos rótulos descritivos para os tópicos, conforme apresentado na Tabela II. A análise desse conjunto de *stems* aponta para tópicos coerentes e frequentemente semelhantes aos temas mencionados nos nomes das comissões permanentes da Câmara dos Deputados. Além dos tópicos alinhados às comissões, há aqueles que parecem ser mais específicos. Por exemplo, tanto o tópico 14 quanto o tópico 19 estão relacionados com a área de Saúde, mas com aspectos distintos da mesma. Já os tópicos 8 e 12 exploram temas relacionados ao meio ambiente, mas seus níveis de granularidade diferem bastante. Nossa segunda checagem de coerência e interpretabilidade envolveu a leitura de uma fração dos documentos associados a alguns dos tópicos, em uma avaliação subjetiva que corroborou a análise dos *stems*.

5. TÓPICOS DISCUTIDOS NAS COMISSÕES

Com os tópicos latentes avaliados e compreendidos, nos atemos agora a sua relação com as comissões permanentes da Câmara dos Deputados. A Figura 2 apresenta a distribuição dos eventos de cada

Tabela II. *Stems* mais associados aos temas definidos pelo modelo *Latent Dirichlet Allocation* adotado.

Tópico	Rótulo	*Stems* mais associados
0	**Investigações**	"document", "empr", "animal", "denúnc", "mor", "filh", "investig", "jan", "dinh" e "min".
1	**Segurança Pública**	"políci", "crim", "polic", "milit", "arm", "seguranç públic", "guard", "políci feder", "penal" e "justiç'.'
2	**Educação e Cultura**	"cult", "educ", "municípi", "plan", "livr", "orç", "municip", "estad", "plan nacion" e "emend".
3	**Tributação**	"previd", "reform", "bilhã", "tribut", "fiscal", "receit", "rend", "crédit", "impost" e "dinh".
4	**Agricultura**	"produt", "agricult", "produç", "rural", "aliment", "agricul", "produz", "leit", "famili" e "sul".
5	**Legislação**	"paut", "it", "matér", "parec", "projet lei", "retir", "emend", "constitucional", "mérit" e "propos".
6	**Esporte**	"esport", "atlet", "tur", "jog", "cop", "futebol", "olímp", "club", "estádi" e "confeder".
7	**Regulação e Controle**	"consum", "empr", "comunic", "internet", "oper", "agênc", "regul", "red", "merc" e "preç".
8	**Amazônia**	"amazôn", "pesc", "mar", "forç", "fronteir", "arm", "ama", "forç arm", "milit" e "nort".
9	**Direitos Humanos**	"human", "direit human", "lut", "mov", "negr", "viol", "mulh", "companh", "civil" e "comunidad".
10	**Viação e Transporte**	"transport", "obr", "aeroport", "contrat", "veícul", "empr", "invest", "avi", "infraestrut" e "oper".
11	**Trabalho**	"empr", "empreg", "entidad", "administr", "contrat", "profiss", "sindicat", "companh", "regulament" e "fiscal".
12	**Meio Ambiente**	"ambient", "ambi", "florest", "conserv", "áre", "sustent", "amazôn", "unidad", "ibam" e "desmat".
13	**Minas e Energia**	"energ", "empr", "elétr", "miner", "consum", "distribu", "min", "gás", "usin" e "invest".
14	**Pesquisa em Saúde**	"defici", "medic", "produt", "pesso defici", "anvis", "pesquis", "drog", "aliment", "agrotóx" e "control".
15	**Jurídico**	"tribun", "justiç", "supr", "julg", "repúbl", "defend", "democrac", "advog", "parl" e "judici".
16	**Indústria e Economia**	"empr", "indústr", "merc", "tecnolog", "invest", "internac", "cresc", "econom", "produt" e "unid".
17	**Seguridade Social**	"mulh", "crianç", "viol", "adolesc", "famíl", "filh", "crianç adolesc", "sex", "menin" e "hom".
18	**Educação**	"educ", "escol", "univers", "profes", "ensin", "alun", "curs", "profiss", "prof" e "superi".
19	**Assistência Médica**	"médic", "paci", "hospit", "doenç", "profiss", "su", "idos", "assist", "diagnóst" e "medicin".
20	**Indígenas e Quilombolas**	"indígen", "terr", "comunidad", "pov", "índi", "incr", "fun", "assent", "quilombol" e "pov indígen".
21	**Desenvolvimento Urbano**	"municípi", "águ", "prefeit", "urban", "municip", "nord", "habit", "resídu", "sane" e "plan".

uma das 25 comissões em um espaço bidimensional gerado aplicando o algoritmo *Uniform Manifold Approximation and Projection* (UMAP) [McInnes et al. 2018] às 22 dimensões de associação entre as notas taquigráficas e os tópicos latentes definidos pelo LDA.

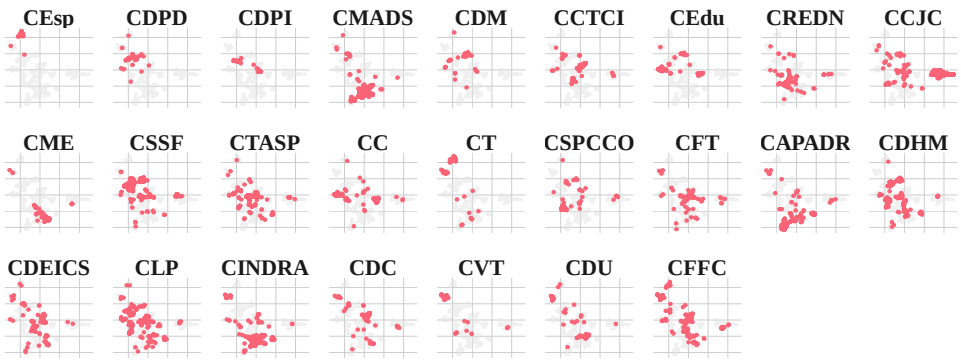

Fig. 2. Notas taquigráficas das comissões permanentes da Câmara dos Deputados distribuídas em espaço bidimensional gerado com *Latent Dirichlet Allocation* e *Uniform Manifold Approximation and Projection for Dimension Reduction.*

Essa análise mostra comissões cujas notas taquigráficas se concentram em regiões bem específicas. São exemplos a Comissão do Esporte (CEsp) e a Comissão dos Direitos da Pessoa Idosa (CDPI). Entretanto, todas as comissões possuem transcrições que desviam dessa tendência, indicando alguma diversidade nos tópicos discutidos. Na Comissão de Legislação Participativa (CLP), por exemplo, essa variação é evidente. Ademais, percebemos considerável semelhança entre a distribuição das notas taquigráficas de comissões cujos temas têm notória interseção. É o caso das comissões de Meio Ambiente e Desenvolvimento Sustentável (CMADS), de Integração Nacional, Desenvolvimento Regional e Amazônia (CINDRA) e de Agricultura, Pecuária e Desenvolvimento Regional (CAPADR).

A Figura 3 mostra as associações médias entre cada comissão permanente da Câmara dos Deputados e todos os 22 tópicos latentes definidos pelo LDA. As associações médias corroboram a hipótese de que cada comissão permanente se dedica a alguns poucos tópicos, tipicamente entre 2 e 4 deles. A única exceção a essa tendência é a Comissão de Legislação Participativa (CLP). Diferente das demais comissões, a principal função da CLP é acolher e discutir as sugestões encaminhadas pela sociedade

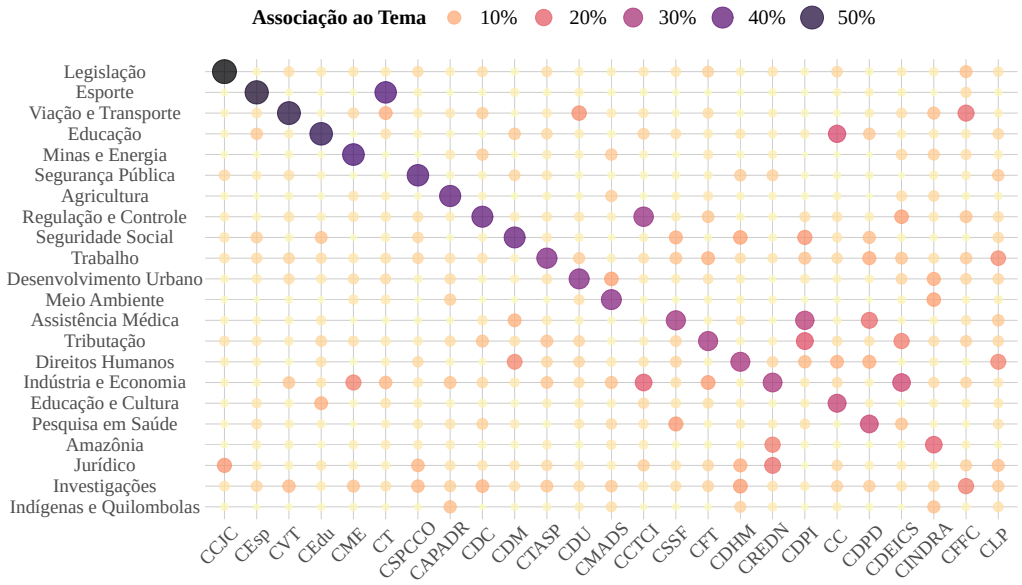

Fig. 3. Associação média das comissões permanentes da Câmara dos Deputados aos tópicos identificados pelo *Latent Dirichlet Allocation* no período de 2008 a 2019.

civil à Câmara. Assim, ao mesmo tempo em que amplia a participação popular no processo legislativo, essa comissão se depara com a necessidade de discutir temas muito diversos.

O perfilamento das comissões permanentes também nos permite captar dinâmicas mais sutis no contexto da Câmara dos Deputados. No Brasil, grupos relacionados aos setores agropecuário e de mineração frequentemente estão em conflito com ambientalistas e, por vezes, com ativistas dos direitos humanos. As associações comissão-tópico que encontramos mostram que essa tensão acontece também no âmbito das comissões. Na Comissão de Meio Ambiente e Desenvolvimento Sustentável (CMADS), os temas Agricultura (tópico 4) e Minas e Energia (tópico 13) parecem ser discutidos com frequência. De maneira complementar, na Comissão de Agricultura, Pecuária, Abastecimento e Desenvolvimento Rural (CAPADR), os temas Meio Ambiente (tópico 12) e Indígenas e Quilombolas (tópico 20) estão regularmente presentes nas discussões. Essa ênfase em assuntos que, com frequência, antagonizam o tema principal sugere que as comissões funcionam não apenas como espaço de discussão de um tema central, mas também como espaço de conflito que envolve o debate de interesses discordantes dos esperados na comissão.

6. TÓPICOS AO LONGO DO TEMPO

Com base em observações de um especialista político, investigamos a dinâmica dos tópicos no tempo por duas perspectivas. Primeiro, a Figura 4 mostra a associação média de todos os eventos nos meses em que houve, ao menos, 10 eventos nas comissões. Vemos que os temas de Seguridade Social, Direitos Humanos e Assistência Médica se tornaram consideravelmente mais presentes no decorrer do período analisado. Em contraste, temas como Trabalho e Indústria e Economia se tornaram menos presentes. Essa dinâmica sugere que a Câmara dos Deputados tem dedicado mais espaço aos temas sociais em seus debates à medida em que eles ganharam espaço no debate público brasileiro.

A segunda perspectiva em que analisamos a dinâmica dos temas é nos atendo aos tópicos mais discutidos na principal comissão da Câmara dos Deputados, a Comissão de Constituição e Justiça e de Cidadania (CCJC). A Figura 5 apresenta a associação comissão-tópico mensal, entre 2015 e 2019,

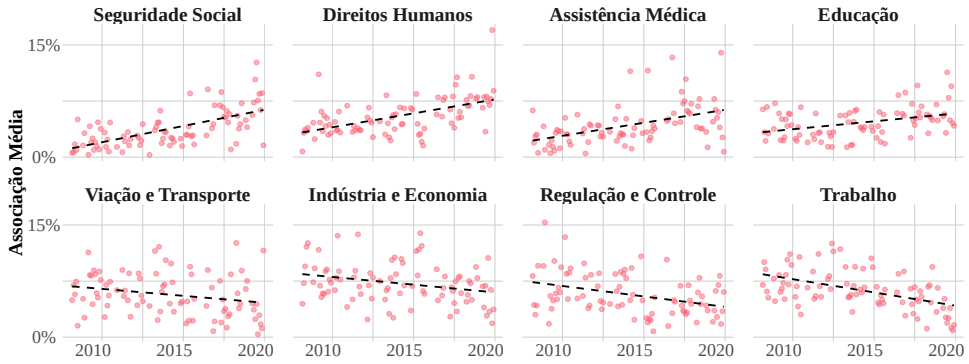

Fig. 4. Associação mensal de todos os eventos das comissões permanentes aos tópicos identificados através do *Latent Dirichlet Allocation*. A linha tracejada indica uma regressão linear da associação em função do tempo. Mostramos os 8 tópicos com maior coeficiente absoluto nessa regressão. Apenas meses com ao menos 10 eventos são incluídos.

para os 10 tópicos mais discutidos nessa comissão. No decorrer desses 5 anos, a CCJC se manteve majoritariamente associada ao tema Legislação (tópico 5), um comportamento esperado dadas as suas atribuições. Em alguns meses específicos, no entanto, as discussões dessa comissão passam a dar destaque maior a outros temas, especialmente ao Jurídico (tópico 15).

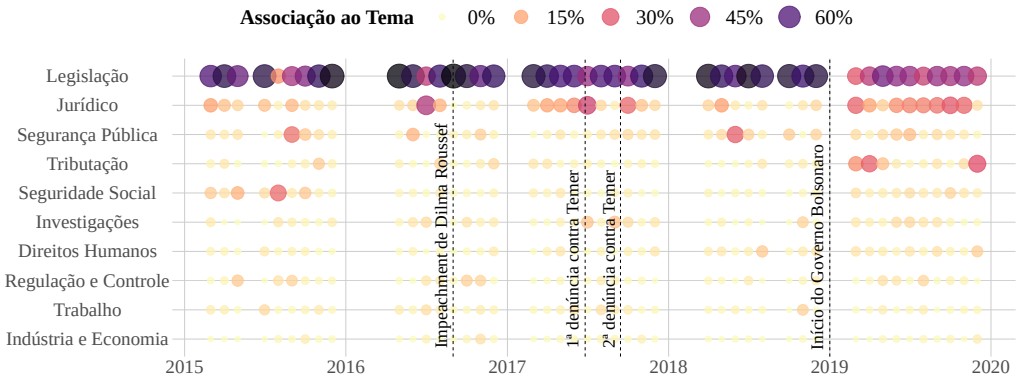

Fig. 5. Tópicos mais associados, mês a mês, à Comissão de Constituição e Justiça e de Cidadania entre 2015 e 2019.

A leitura das notas taquigráficas e registros jornalísticos da época mostram que o aumento do tema Jurídico no início do 2º semestre de 2016 está relacionado ao processo de *impeachment* da presidenta Dilma Rousseff. Por se basear em uma denúncia de crime de responsabilidade, esse processo foge às atribuições da CCJC e, até então, não tinha impactado suas discussões. Às vésperas das últimas votações, no entanto, essa separação se desfaz. Nas 5 notas taquigráficas disponibilizadas pela CCJC durante julho de 2016, apenas uma não menciona o *impeachment*. Tipicamente, as menções surgem em contextos pertinentes aos eventos, mas levam os parlamentares a se engajar em debates sobre os méritos e as motivações do processo. É notável ainda que no mês seguinte à conclusão do processo de *impeachment* (outubro de 2016), as discussões da CCJC voltam a estar associadas quase exclusivamente ao tema Legislação.

Já o ano de 2017 é marcado dois picos de associação da CCJC ao tema Jurídico, especificamente nos meses de julho e outubro. Esses períodos sucedem imediatamente as denúncias da Procuradoria Geral da República contra o presidente Michel Temer. Em ambos os casos, as acusações envolvem

apenas crimes comuns (como corrupção passiva) e não crimes políticos. Por esse motivo, o processo requeria aprovação da CCJC para prosseguir. As duas denúncias receberam parecer contrário dessa comissão, mas impactaram as discussões vigentes. Diferente do que aconteceu em 2016, o crescimento e decrescimento do tema Jurídico é gradativo, o que está associado a outras denúncias da Operação Lava Jato que também são extensivamente discutidas na CCJC nesse período.

Por fim, as associações comissão-tópico mensais mostram quão atípicas foram as discussões da CCJC durante 2019. Nesse ano, a associação ao tema Jurídico não apresentou picos relacionados a eventos específicos, mas se manteve alta durante todos os meses. As notas taquigráficas sugerem que dois fatores são responsáveis por esse longo deslocamento temático na comissão. Primeiro, o início do governo de Jair Bolsonaro teve uso frequente de decretos presidenciais para alterar a legislação. Essas práticas foram comentadas pelos parlamentares da CCJC e eles chegaram, inclusive, a derrubar alguns desses decretos. É o caso do decreto relacionado à flexibilização do porte de armas, derrubado pela CCJC em junho de 2019. O segundo fator que modificou a associação tópico-comissão dessa comissão foi a realização de diversas audiências públicas para discutir a PEC 410/2018, que permitiria a condenação em segunda instância.

7. CONCLUSÃO

Neste trabalho, investigamos as atividades desenvolvidas pelas comissões permanentes da Câmara dos Deputados, durante o período de 2008 a 2019, a partir das falas de seus integrantes. A princípio, os documentos que embasam esse estudo não estavam disponíveis em formato aberto, tornando necessária sua libertação. O conjunto de dados produzido contém 18.839 documentos que descrevem as discussões das comissões durante mais de duas décadas. Utilizando o *Latent Dirichlet Allocation*, nossa modelagem de tópicos apresentou bons níveis de verossimilhança, de coerência e de interpretabilidade. A partir da associação entre as notas taquigráficas e os tópicos latentes, fomos capazes de identificar características e relações entre as diferentes comissões, bem como examinar mudanças em suas discussões no decorrer do tempo. Além de prover *insights* ao acompanhamento da Câmara dos Deputados, essa abordagem sugere uma direção promissora para novos estudos e para o controle social da atividade parlamentar brasileira.

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.

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.

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.