UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MATHEUS ALVES DOS SANTOS

RECONHECIMENTO AUTOMÁTICO DE TEMAS ABORDADOS E DESVIOS TEMÁTICOS EM COMISSÕES DA CÂMARA DOS DEPUTADOS

CAMPINA GRANDE - PB

2021

MATHEUS ALVES DOS SANTOS

RECONHECIMENTO AUTOMÁTICO DE TEMAS ABORDADOS E DESVIOS TEMÁTICOS EM COMISSÕES DA CÂMARA DOS DEPUTADOS

Trabalho de Conclusão de Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Nazareno Ferreira de Andrade

CAMPINA GRANDE - PB

2021

S237r Santos, Matheus Alves dos.
Reconhecimento automático de temas abordados e desvios temáticos em comissões da câmara dos Deputados. / Matheus Alves dos Santos. – 2021.

10 f.

Orientador: Prof. Dr. Nazareno Ferreira de Andrade.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Processamento de linguagem natural. 2. Modelo estatístico generativo. 3. Latent Dirichlet Allocation - LDA. 4. Reconhecimento automático de temas. 5. Câmara dos deputados - comissões. 6. Poder legislativo – debates políticos. 7. Dados abertos. 8. Libertação de dados. 9. Dados textuais - processamento. 10. Modelagem de tópicos. 11. Transcrição de dados de libertação. I. Andrade, Nazareno Ferreira de. II. Título.

CDU:004.62(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

MATHEUS ALVES DOS SANTOS

RECONHECIMENTO AUTOMÁTICO DE TEMAS ABORDADOS E DESVIOS TEMÁTICOS EM COMISSÕES DA CÂMARA DOS DEPUTADOS

Trabalho de Conclusão de Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Nazareno Ferreira de Andrade

Orientador – UASC/CEEI/UFCG

Professora Dra. Patrícia Duarte de Lima Machado

Examinadora – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 25 de maio de 2021.

CAMPINA GRANDE - PB

ABSTRACT

Access to information is indispensable for creating a politically participative society. In Brazil, many initiatives aim to ensure transparency in the actions of the Legislative Branch. However, Brazilian National Congress committees still receive only a small fraction of the media attention dedicated to the plenary sessions. This scenario is harmful to civil society since the committees are the real stage for the political clashes and debates between the Brazilian parliamentarians. Using Natural Language Processing techniques, especially the generative statistical model Latent Dirichlet Allocation (LDA), this work presents an approach for automatic recognition of addressed themes and thematic deviations in events of the Brazilian Chamber of Deputies's permanent committees. The obtained results prove the applicability of this statistical model in the monitoring of current political debates, defining latent topics aligned with the themes of the committees and allowing the detection of the events whose debates were affected by thematic deviations.

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

Reconhecimento automático de temas abordados e desvios temáticos em comissões da Câmara dos Deputados

Matheus Alves dos Santos
matheus.santos@ccc.ufcg.edu.br
Universidade Federal de Campina Grande

Nazareno Andrade
nazareno@computacao.ufcg.edu.br
Universidade Federal de Campina Grande

Fábio Morais
fabio@computacao.ufcg.edu.br
Universidade Federal de Campina Grande

RESUMO

O acesso à informação é imprescindível para a construção de uma sociedade politicamente participativa. No Brasil, são numerosas as iniciativas que visam garantir transparência às ações do Poder Legislativo. Contudo, as comissões do Congresso Nacional ainda recebem apenas uma fração da atenção midiática dedicada aos plenários. Esse cenário é prejudicial à sociedade civil, uma vez que as comissões são o verdadeiro palco dos embates e discussões políticas dos parlamentares brasileiros. Utilizando técnicas de Processamento de Linguagem Natural, especialmente o modelo estatístico generativo *Latent Dirichlet Allocation* (LDA), este trabalho descreve uma abordagem para reconhecimento automático dos temas abordados e dos desvios temáticos em eventos das comissões permanentes da Câmara dos Deputados. Os resultados obtidos comprovam a aplicabilidade desse modelo estatístico no acompanhamento dos debates políticos vigentes, definindo tópicos latentes alinhados aos temas das comissões e permitindo a detecção do conjunto de eventos cujos debates foram afetados por desvios temáticos.

PALAVRAS-CHAVE

Processamento de Linguagem Natural, *Latent Dirichlet Allocation*, Câmara dos Deputados, Política.

ABSTRACT

Access to information is indispensable for creating a politically participative society. In Brazil, many initiatives aim to ensure transparency in the actions of the Legislative Branch. However, Brazilian National Congress committees still receive only a small fraction of the media attention dedicated to the plenary sessions. This scenario is harmful to civil society since the committees are the real stage for the political clashes and debates between the Brazilian parliamentarians. Using Natural Language Processing techniques, especially the generative statistical model Latent Dirichlet Allocation (LDA), this work presents an approach for automatic recognition of addressed themes and thematic deviations in events of the Brazilian Chamber of Deputies's permanent committees. The obtained results prove the applicability of this statistical model in the monitoring of current political debates, defining latent topics aligned with the themes of the committees and allowing the detection of the events whose debates were affected by thematic deviations.

KEYWORDS

Natural Language Processing, Latent Dirichlet Allocation, Chamber of Deputies, Politics.

1 INTRODUÇÃO

Segundo o princípio da *trias politica*, 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, a Constituição Federal de 1988 institui a adoção do bicameralismo na esfera nacional do Poder Legislativo. Assim, duas casas legislativas compõem o Congresso Nacional brasileiro: o Senado Federal e a Câmara dos Deputados. Esta última é composta por 513 deputados federais, eleitos quadrienalmente e de maneira proporcional à população das unidades federativas brasileiras. Os deputados federais são eleitos como representantes dos interesses do povo brasileiro para legislar sobre assuntos de interesse nacional e fiscalizar a aplicação de recursos públicos [7].

As comissões parlamentares são grupos formados por integrantes de uma casa legislativa com intuito de analisar aspectos técnicos e legais das proposições de lei. Seus poderes estão definidos no artigo 58 da Constituição Federal e em seus Regimentos Internos. O poder de apreciação conclusiva, por exemplo, permite que vários tipos de proposições possam ser aprovadas ou rejeitadas sem votação em plenário, sendo necessário apenas que as comissões envolvidas na tramitação estejam em consenso.

Atualmente, existem 25 comissões permanentes na Câmara dos Deputados. Todavia, mesmo com seu profundo impacto no processo legislativo brasileiro, as atividades desses grupos recebem menos atenção midiática do que o plenário e, consequentemente, são menos acompanhadas pela sociedade civil. Visto que as informações disponibilizadas acerca das atividades das comissões se resumem quase exclusivamente a notas taquigráficas e gravações de sessões, torná-las mais acessíveis e criar ferramentas que facilitem o acompanhamento do processo legislativo são maneiras de fomentar o controle social e a participação de cidadãos no debate político. Algumas ferramentas dessa natureza já existem no Brasil e vêm se mostrando efetivas. São exemplos o Parlametria[1], o Radar Legislativo[2] e o Elas no Congresso[3].

As técnicas de Processamento de Linguagem Natural têm sido consistentemente aplicadas no contexto político nos últimos anos, criando soluções automatizadas para atividades como a medição de proporções ideológicas em discursos políticos [17] e a identificação

[1] https://parlametria.org/
[2] https://radarlegislativo.org/
[3] https://www.elasnocongresso.com.br/

de opiniões políticas em *tweets* [15]. Entre as técnicas que têm se destacado no acompanhamento das ações do Poder Legislativo está a modelagem de tópicos, uma abordagem estatística não supervisionada que modela a relação de textos com temas abstratos. Seu uso já se mostrou viável tanto no Brasil quanto no exterior, com estudos dedicados às ênfases temáticas em discursos do plenário da Câmara dos Deputados [16] e ao acompanhamento da agenda política do Parlamento Europeu [9], por exemplo.

No cenário brasileiro, entretanto, um dos maiores desafios à aplicação dessas técnicas é a obtenção dos conjuntos de dados. Em contradição com a ótima classificação do Brasil no *Global Open Data Index*[4], as transcrições de sessões do Congresso Nacional não estão disponíveis em formato adequadamente estruturado em nenhuma das fontes governamentais existentes. Dessa forma, aqueles que desejem acessar esses dados em formato processável por máquina deverão solicitá-los através da Lei de Acesso à Informação ou obtê-los via técnicas de extração da informação, como a raspagem de dados.

Nesse contexto, o presente trabalho tem duas contribuições:

(1) Apresentamos um conjunto de dados estruturados contendo 18.839 transcrições de eventos realizados por comissões da Câmara dos Deputados entre 1995 e 2020. A construção dessa coleção de documentos foi feita por meio de raspagem de dados aplicada ao Portal da Câmara dos Deputados[5] e todo o ferramental desenvolvido para sua realização também foi disponibilizado publicamente. Dessa forma, buscamos contribuir com o surgimento de estudos retrospectivos e pesquisas correlatas que se beneficiem do uso desse conjunto de dados, especialmente na área de Ciência Política.
(2) Utilizando os dados extraídos, propomos uma abordagem de modelagem de tópicos para o reconhecimento automático de temas abordados e de desvios temáticos em sessões de comissões permanentes da Câmara dos Deputados. Para esse fim, aplicamos o *Latent Dirichlet Allocation*, um modelo generativo probabilístico para conjuntos de dados discretos. Apresentamos também os resultados de uma validação que demonstra a viabilidade da adoção desse modelo no reconhecimento dos temas discutidos nas comissões permanentes da Câmara dos Deputados e que nos permitiram identificar a ocorrência de desvios temáticos.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, estão apresentados os conceitos de Ciência de Dados e Processamento de Linguagem Natural necessários à adequada compreensão deste trabalho.

2.1 Dados abertos e libertação de dados

Segundo a *Open Definition*[6], dados abertos são aqueles que podem ser livremente acessados, utilizados, modificados e compartilhados por qualquer pessoa e para qualquer propósito. Além disso, seu uso não deve estar sujeito a exigências, exceto aquelas que busquem preservar sua proveniência e abertura. A *Open Government Data*[7] também define um conjunto de oito princípios a serem seguidos por dados abertos governamentais. Entre eles estão a completude, a processabilidade por máquina e a disponibilização em formato não proprietário. Assim, a ausência de limitações de privacidade torna um conjunto de dados público, mas não necessariamente aberto.

O processo de conversão de dados "fechados" em dados abertos é conhecido como libertação de dados. Por meio de operações de extração e correção, ele pode ser aplicado a conjuntos de dados disponibilizados publicamente para torná-los abertos ou, ao menos, aproximá-los desse estado. Nesse contexto, o Brasil.io[8] e o Dados-JusBR[9] são exemplos de iniciativas da sociedade civil brasileira que se dedicam a garantir e democratizar o acesso à informação através da libertação de dados.

A raspagem de dados é uma técnica de extração automatizada de informação que opera nas saídas de serviços ou aplicativos cuja legibilidade é potencialmente exclusiva para humanos. Por se beneficiar fortemente da estrutura semântica do *HyperText Markup Language* (HTML) para a identificação de informações úteis, essa técnica se tornou uma das principais soluções para a libertação de dados governamentais públicos disponibilizados por meio de páginas web.

2.2 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é a área de pesquisa dedicada ao estudo de técnicas computacionais para compreensão e manipulação de linguagens naturais. Adotando fundamentos de Estatística, Linguística e de outras áreas de conhecimento, os algoritmos e ferramentas desenvolvidas realizam tarefas como a tradução automática de texto, a modelagem de tópicos e a síntese de fala [6]. Os primeiros registros de estudos nessa área remontam ao final da década de 1940. Mesmo com resultados pouco satisfatórios na tradução automática de texto, esses estudos foram o ponto de partida para o surgimento de abordagens estatísticas, simbolistas e conexionistas ao longo dos anos [14]. Atualmente, as técnicas mais modernas (como Redes Neurais Recorrentes) apresentam bons resultados até mesmo em tarefas mais complexas, como o reconhecimento de voz [8].

2.3 Pré-processamento de dados textuais

Em conjuntos de dados textuais, os registros são chamados de documentos e a coleção de documentos é chamada de *corpus*. Nesse cenário, o pré-processamento de texto é o conjunto de operações realizadas sobre os documentos para adequá-los a análises ou outras tarefas de Processamento de Linguagem Natural. Essas operações são utilizadas para padronizar o conjunto de dados, reduzir sua dimensionalidade e/ou descartar informações desnecessárias [1]. Em função da diversidade de abordagens existentes para esses fins, as operações que compõem o *pipeline* de pré-processamento adotado neste trabalho estão resumidamente descritas a seguir.

- **Padronização de capitalização:** Conversão dos caracteres alfabéticos dos documentos para uma mesma caixa (alta ou baixa).
- **Remoção de pontuação:** Remoção dos caracteres não alfanuméricos presentes nos documentos.

[4] https://index.okfn.org/
[5] https://www.camara.leg.br/
[6] http://opendefinition.org/
[7] https://opengovdata.org/
[8] https://brasil.io/
[9] https://dadosjusbr.org/

- **Tokenização:** Decomposição de cada documento do *corpus* em um conjunto de unidades menores, denominadas *tokens*. Essa divisão costuma se basear no conceito de *n*-gramas, definindo os *tokens* como sequências de *n* palavras, sendo $n \in \mathbb{N}^+$ [10].
- **Remoção de *stopwords*:** Remoção das palavras com pouca ou nenhuma contribuição semântica para os documentos em que estão inseridas [13]. Essas palavras, denominadas *stopwords*, podem ser escolhidas a partir de suas classes gramaticais (como artigos e pronomes), mas também de características do próprio *corpus*.
- **Stemização:** Conversão das palavras flexionadas e/ou derivadas para suas bases, denominadas *stems* ou raízes [18]. Através dessa operação, as palavras "*legislação*" e "*legislativo*" podem ser reduzidas para o *stem* "*legisl*", por exemplo.

Nas tarefas relacionadas à Aprendizagem de Máquina, é comum que o treinamento dos modelos seja baseado exclusivamente em dados numéricos. Assim, após as operações de pré-processamento, os dados textuais ainda precisam ser convertidos em representações numéricas adequadas. Esse processo, denominado **vetorização**, transforma cada documento do *corpus* em um vetor contendo valores discretos ou contínuos. O *Term Frequency* (TF) é uma das estratégias mais simples de vetorização e consiste em transformar o *corpus* em uma matriz de frequência de termos. Essa matriz M possui formato $D \times S$ e é construída de maneira que D seja o número de documentos, S seja o número de *tokens* distintos no *corpus* e M_{ij} seja o número de ocorrências do j-ésimo *token* distinto no i-ésimo documento.

2.4 Modelagem de tópicos

A modelagem de tópicos é uma tarefa de Processamento de Linguagem Natural e Aprendizagem de Máquina que visa identificar estruturas semânticas implícitas em coleções de documentos [3]. Em outras palavras, essa técnica não supervisionada utiliza modelos probabilísticos de análise Bayesiana hierárquica para descrever a relação entre os textos dos documentos com temas abstratos. O *Latent Dirichlet Allocation* e o *Latent Semantic Analysis* são exemplos desse tipo de modelo. Com bons resultados em diversos domínios, incluindo artigos científicos e matérias jornalísticas, a modelagem de tópicos tem sido consistentemente adotada para automatizar a classificação de documentos em *corpora* modernos, cujas dimensões e taxa de crescimento inviabilizam o uso de métodos tradicionais de classificação.

O *Latent Dirichlet Allocation* (LDA) é um modelo generativo probabílistico que utiliza uma hierarquia de três níveis para descrever os documentos de um *corpus* como combinações finitas de tópicos implícitos [4]. 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. Dessa forma, o modelo torna-se capaz de representar a natural heterogeneidade de temas abordados em documentos.

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 [3]. Essa estrutura está apresentada na Figura 1 através da notação de placa, na qual os círculos representam as variáveis e as caixas denotam multiplicidade. Nela, θ e ϕ são distribuições probabilísticas de *Dirichlet* controladas através dos parâmetros α e β, respectivamente. Além disso, estão definidos K tópicos, M documentos e N palavras por documento. Dessa maneira, a j-ésima palavra do documento i (denominada w_{ij}) terá sua associação z_{ij} aos tópicos definida a partir da distribuição θ_i 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 dos documentos com os temas implícitos.

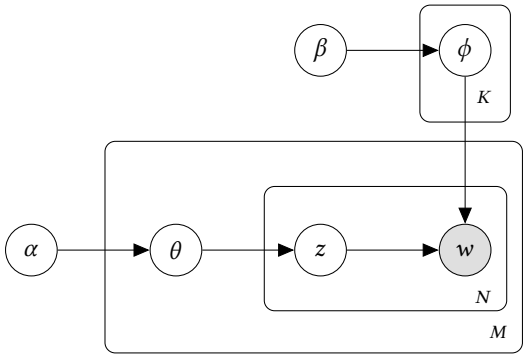

Figura 1: Representação do modelo *Latent Dirichlet Allocation* em notação de placa.

3 LIBERTAÇÃO DOS DADOS DE TRANSCRIÇÕES

O Portal da Câmara dos Deputados é o meio utilizado por essa casa legislativa para garantir o acesso da sociedade civil às informações sobre suas atividades. Além de serviços dedicados ao acompanhamento e à transparência das ações dos parlamentares, esse portal possui uma seção de dados abertos em que são publicados diversos conjuntos de dados referentes à Câmara dos Deputados. Esses dados estão disponíveis através de uma API RESTful bem documentada, mas os usuários também podem acessá-los através do download de arquivos em diferentes formatos, como CSV, JSON e XLSX. As proposições de lei e a íntegra dos discursos realizados em plenário são exemplos dos conjuntos de dados disponíveis. Não estão presentes, no entanto, dados abertos de transcrições dos eventos realizados pelas comissões. Mesmo assim, esses registros taquigráficos existem e são disponibilizados pelo Portal da Câmara dos Deputados em páginas HTML. Por isso, ainda que essa base de dados seja pública, seu uso é inviável para a maioria dos usuários.

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. A principal tecnologia utilizada durante esse processo foi o Scrapy[10], um *framework* para raspagem de dados em páginas web. Além dele, também foram utilizadas as bibliotecas BeautifulSoup[11]

[10] https://docs.scrapy.org/
[11] https://www.crummy.com/software/BeautifulSoup/

e Pandas[12] para a manipulação de conteúdo HTML e dos dados extraídos, respectivamente.

Esse ferramental inclui dois raspadores de dados distintos: um dedicado aos metadados de eventos das comissões e outro dedicado às transcrições em si. A obtenção do conjunto de metadados foi necessária para recuperar informações importantes que não estão presentes nas páginas das transcrições, como as comissões de origem e as datas de realização dos eventos aos quais esses registros taquigráficos se referem. Os metadados foram armazenados em um único arquivo, cuja estrutura está apresentada na Tabela 1, e constituem um índice das transcrições dos eventos de comissões que são disponibilizadas pelo Portal da Câmara dos Deputados. Durante essa etapa da raspagem de dados, foram incluídas operações de padronização e de correção de grafia. Além disso, também se criou um atributo para categorizar as comissões e diferenciá-las, por exemplo, entre permanentes e temporárias.

Tabela 1: Descrição dos metadados extraídos sobre os eventos das comissões da Câmara dos Deputados

Atributo	Descrição
`id_evento`	Identificador único do evento.
`categoria_evento`	Categoria do evento. Permite identificar, por exemplo, os eventos em que ocorrem audiências públicas ou eleições.
`comissao`	Nome da comissão da Câmara dos Deputados a que o evento está associado.
`categoria_comissao`	Categoria definida a partir do nome da comissão a que o evento está associado. Permite identificar, por exemplo, quais comissões são permanentes ou temporárias.
`data`	A data em que o evento foi realizado.
`horario`	O horário em que o evento foi realizado.

A raspagem dos textos das transcrições se mostrou mais desafiadora, uma vez que as páginas em que esses dados estão disponíveis não são bem estruturadas e tampouco fazem bom uso da semântica HTML. Por esse motivo, recorremos à criação de expressões regulares que auxiliassem o raspador na coleta dos dados. Utilizando-as, foi possível automatizar a divisão do texto das notas taquigráficas em falas específicas e, para cada uma delas, separar o conteúdo da fala e o nome do orador. Devido ao volume de dados extraído, as notas taquigráficas foram armazenadas em arquivos individuais, cuja estrutura está apresentada na Tabela 2.

Ao todo, foram extraídas as transcrições de 18.839 eventos realizados por comissões da Câmara dos Deputados entre 1995 e 2020. Essa base de dados foi disponibilizada publicamente através do

[12]https://pandas.pydata.org/

Tabela 2: Descrição dos dados extraídos sobre as transcrições de eventos das comissões da Câmara dos Deputados

Atributo	Descrição
`id_evento`	Identificador único do evento a que a transcrição se refere.
`ordem_discurso`	Atributo auxiliar para a ordenação das falas na transcrição.
`orador`	Identificador do orador de uma fala específica. Geralmente representado pelo nome do parlamentar, esse identificador inclui o partido e o estado de origem nas transcrições mais recentes.
`transcricao`	Transcrição de uma fala específica realizada durante o evento.

Google Drive[13]. Além disso, todo o ferramental utilizado para a libertação dos dados foi desenvolvido de forma *open source* e pode ser acessado em um repositório do GitHub[14]. Vale notar que, segundo o artigo 41 do Regimento Interno da Câmara dos Deputados, o registro taquigráfico dos eventos não é obrigatório e ocorre apenas sob determinação dos presidentes das comissões. Assim, ainda que possua todas as transcrições disponibilizadas pelo Portal da Câmara dos Deputados, esse conjunto de dados se refere a apenas 34,97% dos eventos realizados pelas comissões no período supracitado.

4 MODELAGEM DOS TEMAS ABORDADOS NAS COMISSÕES

Após a libertação dos dados de transcrições das comissões da Câmara dos Deputados, realizamos um experimento com o objetivo de validar qualitativamente o modelo *Latent Dirichlet Allocation* enquanto ferramenta de reconhecimento automático dos temas discutidos e dos eventuais desvios temáticos ocorridos nessas comissões.

A priori, selecionamos uma amostra dessas notas taquigráficas e as submetemos a um *pipeline* de operações de pré-processamento de texto. Os dados resultantes dessas operações foram utilizados para o treinamento de modelos LDA que, posteriormente, foram avaliados quanto à sua verossimilhança. Essa avaliação nos permitiu selecionar o modelo mais adequado à tarefa proposta e, a partir de seus resultados, validamos sua capacidade de identificar os temas abordados por comissões da Câmara dos Deputados, bem como as ocorrências de desvios temáticos nesses órgãos. A metodologia e os resultados desse experimento de validação são descritos e discutidos com maior profundidade ao longo desta seção.

4.1 Seleção da amostra

Devido à dimensão da base de dados e às limitações da capacidade computacional disponível, optamos por definir uma amostra mais

[13]https://bit.ly/transcricoes-comissoes
[14]https://github.com/alvesmatheus/fala-camarada

específica para ser utilizada nesse experimento. Inicialmente, foram selecionadas apenas as transcrições de eventos realizados por comissões permanentes entre 2009 e 2018. Dessa forma, além de assegurar uma cobertura parcial das legislaturas 53, 54 e 55 da Câmara dos Deputados, evitamos os possíveis ruídos gerados por categorias de comissão com escopo mais abrangente. A partir desse subconjunto dos dados, também descartamos as notas taquigráficas de eventos cujas categorias fossem pouco relacionadas às discussões sobre os temas das comissões como, por exemplo, as solenidades, as homenagens e as eleições. Por fim, com base na forte distinção entre seus temas e na quantidade de documentos disponíveis, filtramos as transcrições de três comissões específicas. São elas:

- Comissão de Agricultura, Pecuária, Abastecimento e Desenvolvimento Rural (CAPADR);
- Comissão de Constituição e Justiça e de Cidadania (CCJC);
- Comissão de Direitos Humanos e Minorias (CDHM).

Na Tabela 3, estão apresentadas as distribuições de frequência da amostra em relação às características que motivaram sua escolha. Ao todo, 930 notas taquigráficas de comissões permanentes da Câmara dos Deputados compõem a amostra utilizada no experimento.

Tabela 3: Distribuição das transcrições presentes na amostra em relação às legislaturas, às categorias de evento e às comissões de origem

Legislatura	**# Transcrições**
Legislatura 55 (2015 a 2018)	452
Legislatura 54 (2011 a 2014)	293
Legislatura 53 (apenas 2009 e 2010)	185
Categoria de Evento	**# Transcrições**
Audiência Pública com Convidado(a)	437
Reunião Órdinária	391
Seminário	71
Audiência Pública com Ministro(a)	20
Fórum	6
Reunião Extraordinária	3
Debate	1
Reunião Técnica	1
Comissão de Origem	**# Transcrições**
Comissão de Constituição e Justiça e de Cidadania (CCJC)	405
Comissão de Direitos Humanos e Minorias (CDHM)	270
Comissão de Agricultura, Pecuária, Abastecimento e Desenvolvimento Rural (CAPADR)	255

4.2 Pré-processamento dos dados

Com o objetivo de adequá-las às etapas seguintes do experimento, as notas taquigráficas da amostra selecionada foram submetidas a operações de pré-processamento de texto. A princípio, cada documento do *corpus* foi criado com base na concatenação das falas presentes em uma dessas transcrições. Em seguida, os documentos foram submetidos à padronização de capitalização em caixa baixa e à remoção da pontuação. Ainda, utilizando o ferramental da biblioteca NLTK[15], os textos resultantes dessas operações foram tokenizados em unigramas e bigramas, isto é, em *n*-gramas de tamanho 1 e 2.

Em posse dos *tokens* do *corpus*, iniciamos a criação do conjunto de *stopwords* a serem removidas. Para isso, após incluir as *stopwords* definidas pelas bibliotecas NLTK e SpaCy[16] para a língua portuguesa, investigamos os termos mais comuns em transcrições das comissões permanentes e, iterativamente, selecionamos aqueles que tivessem pouca contribuição semântica nesses documentos. Na Figura 2 estão apresentadas as 100 palavras mais comuns em transcrições de comissões permanentes da Câmara dos Deputados, excetuando aquelas que foram incluídas no conjunto de *stopwords* criado.

Sucedendo a remoção das *stopwords* selecionadas, submetemos os *tokens* restantes à stemização. Para isso, adotamos o Removedor de Sufixos da Língua Portuguesa (RSLP), um algoritmo implementado pela biblioteca NLTK, desenvolvido especificamente para a língua portuguesa e com taxa de acerto superior à de algoritmos de stemização mais tradicionais, como o Porter [12]. Além disso, visto que o LDA é um modelo probabilístico generativo, seu aprendizado se baseia em distribuições de frequência e, por isso, os *stems* criados precisaram ser convertidos em uma representação adequada. Através do ferramental da biblioteca Scikit-Learn[17], aplicamos a vetorização *Term Frequency* (TF) ao *corpus*, não incluindo os *stems* presentes em menos de 5% ou em mais de 80% dos documentos, uma vez que sua frequência os tornaria pouco úteis (e até prejudiciais) ao treinamento do modelo.

4.3 Treinamento do *Latent Dirichlet Allocation*

Após a preparação dos dados a serem utilizados, executamos o treinamento do LDA para o reconhecimento automático dos temas abordados nas comissões. A princípio, investigamos os parâmetros definidos pela biblioteca Scikit-Learn em sua implementação desse modelo. Nela, os parâmetros de controle das distribuições *Dirichlet* (geralmente denominados α e β) assumem, por padrão, o valor inverso ao número de tópicos, tornando-os equiprováveis. Assim, nós nos limitamos a estipular valores (descritos na Tabela 4) para outros dois parâmetros: o número de tópicos definidos pelo modelo e a taxa de decaimento de seu aprendizado. Esta última deve assumir valor entre 0,5 e 1,0 (não inclusos) para garantir a convergência assintótica do treinamento e, por isso, selecionamos valores distribuídos igualmente a partir do ponto médio desse intervalo. Já para o número de tópicos, optamos por uma faixa de

[15] https://www.nltk.org/
[16] https://spacy.io/
[17] https://scikit-learn.org/

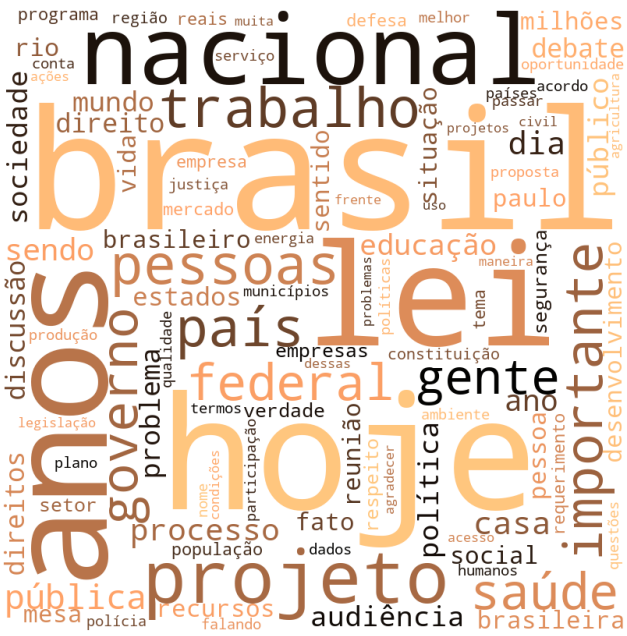

Figura 2: Palavras mais comuns em transcrições de comissões permanentes da Câmara dos Deputados.

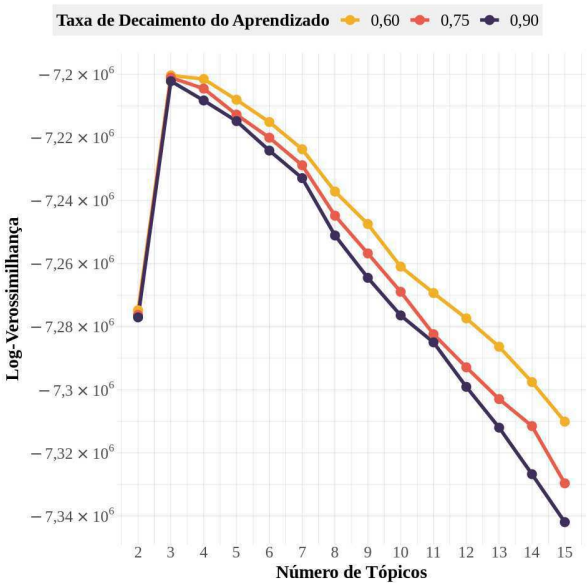

Figura 3: Log-Verossimilhança dos modelos *Latent Dirichlet Allocation* treinados durante a definição de parâmetros.

valores que garantisse, simultaneamente, a produção de modelos que mesclassem duas comissões num mesmo tópico e de modelos que encontrassem vários temas numa mesma comissão.

Tabela 4: Valores estipulados para os parâmetros de treinamento do *Latent Dirichlet Allocation*

Parâmetro	Valores
Número de Tópicos	De 2 a 15 com passo 1.
Decaimento do Aprendizado	De 0,60 a 0,90 com passo 0,15.

Para cada combinação possível entre esses valores de parâmetros, um modelo LDA distinto foi produzido (e treinado) usando a matriz de frequência de termos citada anteriormente. Em seguida, esses 42 modelos foram analisados para determinar quais valores seriam, de fato, adotados como parâmetros. Tipicamente, a avaliação de modelos *Latent Dirichlet Allocation* se baseia em funções de verossimilhança, tal que o modelo ideal é aquele que maximiza o valor dessas funções [2][4]. Neste experimento, adotamos especificamente o log-verossimilhança como métrica de avaliação e, a partir dele, examinamos os resultados alcançados por cada modelo, conforme apresentado na Figura 3.

Com impacto relativamente discreto, uma taxa de decaimento do aprendizado menor se mostrou benéfica ao aprendizado dos modelos. Já para o número de tópicos, observamos uma forte tendência de perda de verossimilhança para valores acima do número de comissões presentes na amostra. Assim, os valores do número de tópicos e da taxa de decaimento selecionados foram de 3 e 0,6, respectivamente. Vale notar, no entanto, que as métricas de verossimilhança avaliam exclusivamente os espaços N-dimensionais definidos pelos N tópicos dos modelos e, portanto, não indicam nenhum juízo de valor sobre a coerência dos temas ou sobre sua interpretabilidade por humanos [5].

4.4 Avaliação dos temas identificados

Escolhido o modelo LDA a ser adotado, nós nos dedicamos à avaliação qualitativa dos temas definidos por ele. Com base em abordagens já consolidadas para essa tarefa [11][16], selecionamos e analisamos os dez *stems* mais associados a cada um dos três temas (apresentados na Tabela 5). A análise desses conjuntos de *stems* permite inferir semelhanças e diferenças entre os contextos semânticos em que eles estão inseridos e, consequentemente, contribui com a compreensão dos temas definidos pelo modelo. Para complementar essa etapa de interpretação dos tópicos por humanos, também selecionamos os dez documentos mais associados a cada um dos temas para que fossem lidos integralmente.

Os três conjuntos de *stems* se mostraram disjuntos e com aparente origem em palavras de significados bem distintos. Para o tema 0, a associação de *stems* como "*produt*", "*agricult*", "*banc*" e "*merc*" indica uma forte ligação com o setor agropecuário, incluindo tanto termos relacionados às atividades da agricultura, quanto termos relacionados ao mercado e à visão econômica desse setor. Já para o tema 1, os principais *stems* se mostraram bastante alusivos às atividades e processos do Poder Legislativo, especialmente "*projet lei*", "*emend*" e "*propos*". Por fim, *stems* como "*direit human*", "*crianç*", "*viol*" e "*lut*" evidenciam a associação do tema 3 às questões sociais e à defesa dos direitos humanos.

Tabela 5: *Stems* mais associados aos temas definidos pelo modelo *Latent Dirichlet Allocation*

Tema	*Stems* mais associados
0	"produt", "agricult", "rural", "produç", "terr", "set", "milhã", "empr", "banc" e "merc".
1	"it", "matér", "parec", "projet lei", "retir", "emend", "constitucional", "constituc", "técn legisl" e "propos".
2	"human", "direit human", "crianç", "viol", "lut", "mulh", "gent", "polici", "crim" e "adolesc".

Corroborando com essas impressões iniciais, a leitura das notas taquigráficas selecionadas nos permitiu validar manualmente o enfoque desses documentos, bem como identificar um possível alinhamento entre os tópicos definidos pelo LDA e os temas aos quais as comissões presentes na amostra se dedicam. Buscando averiguar esse alinhamento, definimos um tema principal para cada documento, isto é, o tema ao qual a nota taquigráfica está mais associada. Em seguida, avaliamos a distribuição das comissões em relação aos temas principais, conforme apresentada na Figura 4. Essa distribuição explicita o alinhamento pressuposto e nos permite considerar que os temas 0, 1 e 2 estão, respectivamente, associados à CAPADR, à CCJC e à CDHM.

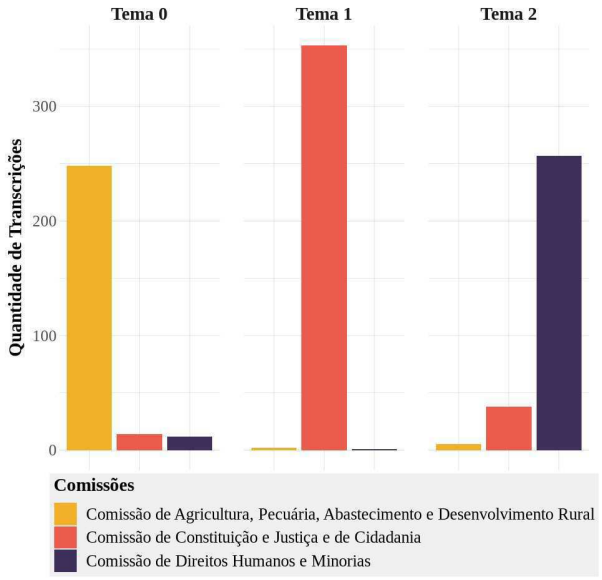

Figura 4: Distribuição das comissões da amostra em relação aos temas principais das transcrições segundo o *Latent Dirichlet Allocation*.

Com a etapa de interpretação dos tópicos devidamente concluída, investigamos a maneira como as notas taquigráficas presentes na amostra se associam a esses temas. Para isso, adotamos o diagrama ternário, um gráfico que representa as proporções de três variáveis como posições em um triângulo equilátero. Conforme apresentado na Figura 5, cada tema definido pelo modelo está graficamente representado por um dos vértices do triângulo e as transcrições, por sua vez, estão representadas pelos pontos contidos nesse triângulo. Dessa forma, a proximidade entre os pontos e os vértices descreve o percentual de associação das transcrições aos três temas.

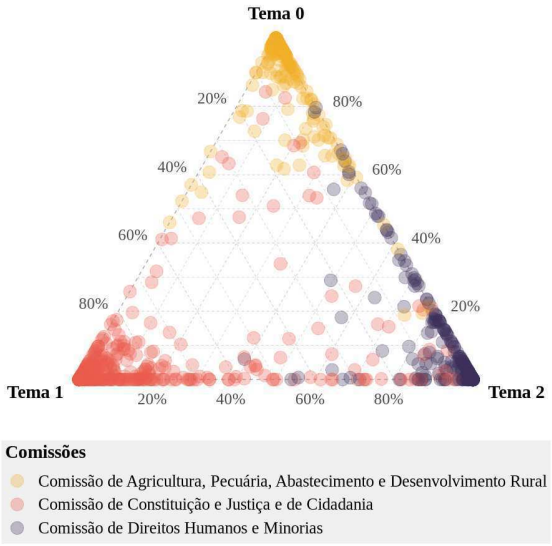

Figura 5: Associação das transcrições da amostra ao temas definidos pelo *Latent Dirichlet Allocation*.

É notável que, entre os documentos da amostra, houve uma tendência de associação quase exclusiva a apenas um dos tópicos, o que se reflete graficamente na alta concentração de pontos nas áreas mais próximas aos vértices do diagrama ternário. A distribuição das associações de documentos aos temas também denota uma nítida, mas não absoluta, separação entre as transcrições oriundas de comissões diferentes, reforçando a caracterização de alinhamento entre os resultados do LDA e os temas das comissões permanentes da Câmara dos Deputados. Considerando apenas os documentos que não estão alinhados a essas tendências, há poucos casos de transcrições significativamente associadas aos três temas, tornando a área central do triângulo quase vazia. Entretanto, visto que as discussões legislativas na Câmara dos Deputados podem envolver duas ou mais comissões, é esperado um certo nível de intersecção entre esses temas.

Convém ressaltar, ainda, que os resultados definidos por modelos LDA não são exaustivos em precisão e, por isso, subtópicos relevantes podem estar contidos nos temas apresentados. Dessa forma, dado o contexto de validação e os resultados apresentados ao longo desta seção, consideramos que o *Latent Dirichlet Allocation* é uma ferramenta útil ao reconhecimento automático de temas abordados nas comissões da Câmara dos Deputados.

4.5 Desvios temáticos

Apesar de possuírem temas bem definidos, é natural supor que as comissões permanentes da Câmara dos Deputados sejam influenciadas por eventos internos e externos à casa legislativa e que, por esse motivo, ocorram desvios temáticos em seus debates. Esses desvios não são necessariamente prejudiciais e podem ser um indicador da sensibilidade das discussões parlamentares à realidade brasileira. Contudo, monitorá-los é de suma importância para avaliar a atuação dos deputados federais nesses ambientes. Assim, considerando a comprovada aplicabilidade do *Latent Dirichlet Allocation* para o reconhecimento de temas abordados pelas comissões permanentes, adotamos os resultados obtidos e descritos anteriormente para identificar a ocorrência de desvios temáticos. Nesse processo, tomamos como base o alinhamento identificado entre os temas definidos pelo modelo LDA e os temas das comissões permanentes presentes na amostra. Dessa forma, selecionamos apenas os documentos que atendessem a uma das seguintes condições:

- A transcrição refere-se a um evento da CAPADR e está mais fortemente associada ao tema 1 ou ao tema 2.
- A transcrição refere-se a um evento da CCJC e está mais fortemente associada ao tema 0 ou ao tema 2.
- A transcrição refere-se a um evento da CDHM e está mais fortemente associada ao tema 0 ou ao tema 1.

O subconjunto de documentos selecionado está apresentado em destaque no diagrama ternário da Figura 6 e inclui 72 notas taquigráficas, o que representa 7,63% da amostra.

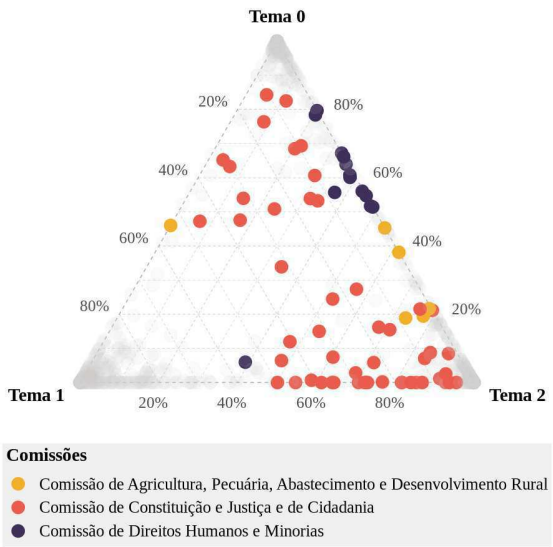

Figura 6: Possíveis desvios temáticos identificados através do *Latent Dirichlet Allocation*.

É evidente a predominância de notas taquigráficas oriundas da CCJC entre os possíveis desvios temáticos identificados: são 52 ocorrências dessa comissão contra apenas 13 da CDHM e 7 da CAPADR. No entanto, a distribuição desigual não implica necessariamente numa maior suscetibilidade aos desvios temáticos, especialmente considerando as atribuições de cada comissão. Na Câmara dos Deputados, a aceitação das proposições de lei deve ser precedida por avaliação favorável da CCJC quanto à constitucionalidade e a outros aspectos legais. Dessa forma, é compreensível que essa comissão aborde parcialmente os temas das demais, incluindo a CAPADR e a CDHM. Nessas condições, recorremos à leitura integral de vinte notas taquigráficas desse subconjunto, selecionadas aleatoriamente, para verificar a ocorrência de desvios temáticos. Durante essa etapa, também pudemos confirmar que os possíveis desvios temáticos identificados pelo modelo incluem os debates sobre assuntos que permeiam duas ou mais comissões. Assim, dada a subjetividade da definição e a ausência de métricas consolidadas para qualificar desvios temáticos, selecionamos alguns exemplos para comprovar sua ocorrência e os apresentamos na Tabela 6.

Provenientes de anos e comissões diferentes, os exemplos selecionados mostram que os desvios temáticos ocorrem em cenários bem variados. O evento 0800/14, por exemplo, é a segunda etapa de uma audiência pública realizada em 2014 pela CAPADR em que ocorreu uma acalorada discussão entre o convidado e um deputado federal acerca da consciência étnica indígena. Assim, de forma semelhante aos trechos apresentados, assuntos que são de competência direta da CDHM acabaram sendo abordados durante o evento inteiro, causando a associação ao tema 2. Já no evento 0994/11, uma audiência pública realizada em 2011 pela CCJC, o debate proposto referia-se às medidas de combate ao consumo de *crack* adotadas no Rio de Janeiro, mas a discussão ganhou foco exclusivo no abrigamento compulsório de usuários dessa substância, especialmente sob a ótica social da questão. Esse desvio temático também gerou uma forte associação ao tema 2. Por sua vez, o evento 53486 foi uma reunião ordinária realizada em 2018 pela CDHM e se dedicou ao debate de um novo traçado da BR-158. Apesar de envolver os direitos dos povos indígenas ao território que seria atravessado por essa estrada (um assunto natural à CDHM), a discussão se dedicou aos impactos que essas mudanças no projeto trariam aos agricultores e aos municípios da região, causando a associação do evento ao tema 0.

A partir dos resultados apresentados ao longo desta seção, consideramos que o *Latent Dirichlet Allocation* também pode ser utilizado para identificar as ocorrências de desvios temáticos nas comissões da Câmara dos Deputados. Para cumprir esse objetivo, no entanto, ainda se faz necessária uma investigação mais detalhada das transcrições ou a criação de novas abordagens que possam automatizar esse processo de maneira adequadamente precisa.

5 CONCLUSÃO

O acesso à informação pública é um direito fundamental para o exercício da cidadania e para a construção de democracias robustas. Ao longo da última década, diversos países têm se empenhado em garantir esse direito à sua população, mas ainda há um longo caminho a se percorrer. Mesmo no Brasil, uma referência mundial em dados abertos governamentais, as limitações de acesso à informação ainda são um entrave para o acompanhamento e fiscalização das ações de membros dos três poderes.

Através da libertação dos dados de 18.839 notas taquigráficas de eventos realizados por comissões da Câmara dos Deputados

Tabela 6: Exemplos de desvios temáticos identificados a partir dos resultados do *Latent Dirichlet Allocation*

Evento	Comissão	Tema	Assunto Discutido	Trechos
0800/14	CAPADR	2 (CDHM)	Identidade e consciência étnica indígena.	"*Na verdade, uma agenda de engenharia social, em que ONGs [...] estimulam o ressurgimento de uma consciência étnica.*" "*Eu vou mostrar um pouco mais como esse ato de terrorismo indígena – e isso aqui é uma violência étnica, que eu estou chamando de terrorismo neotupinambá.*"
0994/11	CCJC	2 (CDHM)	Abrigamento compulsório de usuários de *crack*.	"*O poder público tem [...] a obrigação de assegurar aquilo que é a maior garantia dada pelo ECA: o direito à vida e à integridade física. [...] No caso das crianças e adolescentes, fazer o que estamos chamando de abrigamento compulsório.*" "*Existe a concepção e visão dos agentes da saúde, existe a concepção daqueles que tratam do efeito das drogas serem criminalizadas em nosso país, [...] mas existe também um contingente de pessoas interessadas em debater esse assunto.*"
53486	CDHM	0 (CAPADR)	Novo traçado da BR-158 em Mato Grosso.	"*Do outro lado, estão produtores e comerciantes, que querem o projeto inicial sem desvios. É evidente que todos sabem que a área é de expansão agrícola e, portanto, [...] representa intranquilidade para o futuro.*" "*Ficou claro que a estrada é importante para promover o desenvolvimento regional, para promover o desenvolvimento da agricultura das comunidades e municípios daquela região e para integrá-la a um projeto nacional.*"

entre 1995 e 2020, este trabalho contribui significativamente para a melhoria desse cenário e viabiliza o surgimento de novos estudos retrospectivos sobre a atuação dos parlamentares brasileiros nesse período. Ainda, esse conjunto de dados pode ser associado a técnicas de Processamento de Linguagem Natural para a criação de novas abordagens e ferramentas de acompanhamento do Poder Legislativo pela sociedade civil brasileira, como demonstrado em nosso experimento. Neste, validamos a aplicabilidade do modelo *Latent Dirichlet Allocation* nos contextos de reconhecimento dos temas discutidos pelas comissões permanentes da Câmara dos Deputados e de identificação dos desvios temáticos ocorridos nessas discussões. Mais especificamente, encontramos três tópicos latentes no *corpus* adotado, todos eles alinhados aos temas das comissões presentes na amostra. Além disso, identificamos um pequeno percentual das notas taquigráficas (aproximadamente 7,6%) que descrevem discussões de temas fortemente associados a comissões diferentes de sua comissão de origem. Por fim, em análise mais detalhada, comprovamos a ocorrência de desvios temáticos nas transcrições desse subconjunto.

Todavia, compreendemos que o conjunto de dados adotado também constitui uma eventual ameaça à validade do presente trabalho. Em primeira instância, as características inerentes à amostra escolhida podem ter influenciado nossos resultados positiva ou negativamente, o que ressalta a necessidade de reproduzir o experimento com outras amostras. Ainda, considerando que as comissões da Câmara dos Deputados realizam o registro taquigráfico de seus eventos apenas sob determinação de seus respectivos presidentes, o conjunto de dados extraído também pode apresentar viés que o torne pouco capaz de representar as discussões dessa casa legislativa.

Para trabalhos futuros, sugerimos a reimplementação do ferramental desenvolvido na etapa de libertação do conjunto de dados, visando automatizar a extração de novas notas taquigráficas disponibilizadas ao longo do tempo, bem como permitir a filtragem prévia dos dados, buscando auxiliar os usuários mais inexperientes. Ademais, a metodologia adotada para o reconhecimento de temas e de desvios temáticos pode ser replicada em domínios semelhantes como, por exemplo, as comissões temporárias da Câmara dos Deputados e as comissões do Senado Federal. Esses resultados, por sua vez, podem viabilizar uma aplicação de acompanhamento contínuo dos temas discutidos pelo Poder Legislativo brasileiro.

AGRADECIMENTOS

Aos meus orientadores, Nazareno Andrade e Fábio Morais, por toda colaboração e inspiração que foram indispensáveis à construção deste trabalho. À minha família, pelo suporte e incentivo no decorrer de minha jornada. Em especial à minha mãe, Suely, que me ensinou sobre o potencial transformador da educação. Aos amigos e amigas que me acompanham e que, muitas vezes, me conduzem ao longo dessa vida. Aos membros da OpenDevUFCG, atuais e egressos,

por todas as experiências e desafios que compartilhamos. Por fim, àqueles e àquelas que seguem construindo e defendendo a ciência neste país.

REFERÊNCIAS

[1] Murugan Anandarajan, Chelsey Hill, and Thomas Nolan. 2018. Text Preprocessing. In *Practical Text Analytics*. Springer International Publishing, 45–59. https://doi.org/10.1007/978-3-319-95663-3_4

[2] Rachit Arora and Balaraman Ravindran. 2008. Latent Dirichlet Allocation Based Multi-Document Summarization. In *Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data - AND '08*. ACM Press, New York, NY, USA, 91–97. https://doi.org/10.1145/1390749.1390764

[3] David M. Blei and John D. Lafferty. 2009. Topic Models. In *Text Mining*, Ashok N. Srivastava and Mehran Sahami (Eds.). Chapman and Hall/CRC, 71–93. https://doi.org/10.1201/9781420059458

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. https://doi.org/10.5555/944919.944937

[5] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. 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, NY, USA, 288–296. https://doi.org/10.5555/2984093.2984126

[6] Gobinda G. Chowdhury. 2005. Natural Language Processing. *Annual Review of Information Science and Technology* 37, 1 (Jan. 2005), 51–89. https://doi.org/10.1002/aris.1440370103

[7] Alexandre de Moraes. 2017. *Direito Constitucional* (33ª ed.). Atlas, São Paulo, SP.

[8] Alex Graves, Abdel rahman Mohamed, and Geoffrey Hinton. 2013. Speech Recognition with Deep Recurrent Neural Networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 6645–6649. https://doi.org/10.1109/icassp.2013.6638947

[9] Derek Greene and James P. Cross. 2017. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis* 25, 1 (Jan. 2017), 77–94. https://doi.org/10.1017/pan.2016.7

[10] Gregory Grefenstette. 1999. Tokenization. In *Text, Speech and Language Technology*. Springer Netherlands, 117–133. https://doi.org/10.1007/978-94-015-9273-4_9

[11] Justin Grimmer. 2010. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis* 18, 1 (2010), 1–35. https://doi.org/10.1093/pan/mpp034

[12] C. Huyck and V. Orengo. 2001. Speech Recognition with Deep Recurrent Neural Networks. In *International Symposium on String Processing and Information Retrieval*. IEEE Computer Society, Los Alamitos, CA, USA. https://doi.org/10.1109/SPIRE.2001.10024

[13] Dan Jurafsky and James H. Martin. 2009. *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Pearson Prentice Hall, Upper Saddle River, N.J. https://doi.org/10.5555/1214993

[14] Elizabeth D. Liddy. 2017. Natural Language Processing for Information Retrieval. In *Encyclopedia of Library and Information Science, Fourth Edition*. CRC Press, 3346–3355. https://doi.org/10.1081/e-elis4-120008664

[15] Diana Maynard and Adam Funk. 2012. Automatic Detection of Political Opinions in Tweets. In *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 88–99. https://doi.org/10.1007/978-3-642-25953-1_8

[16] Davi Moreira. 2020. Com a Palavra os Nobres Deputados: Ênfase Temática dos Discursos dos Parlamentares Brasileiros. *Dados* 63, 1 (2020). https://doi.org/10.1590/001152582020204

[17] Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring Ideological Proportions in Political Speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 91–101. https://www.aclweb.org/anthology/D13-1010

[18] Jasmeet Singh and Vishal Gupta. 2016. Text Stemming: Approaches, Applications, and Challenges. *Comput. Surveys* 49, 3 (dec 2016), 1–46. https://doi.org/10.1145/2975608