Aplicação de Modelos de Tópicos em análises automatizadas de discursos de senadores brasileiros

Victor Landim Teixeirense Pinheiro
Thiago de Paulo Faleiros

[1]Universidade de Brasília

Abstract. *In this work, we intend to apply topics models technique and evaluate the results in order to obtain information relating to the senator's speeches and their thematic structure over time. It is understood that there is an overflow of information in our society. In this context, an automated approach to analysis can bring out patterns in large collections of data. One of those approaches is Topic Modeling. This tool typically outputs topics from collections of documents. Topics are a set of words that describe a clear semantic concept. In this regard, it is desired to extract patterns from the topics created from the large collection of Brazillian senator speeches, provided by the Federal Senate. The main hypothesis is that there is a correlation between the historical topic evolution and historical, political, and economic events. The results are matched with draft bills, relevant dates, and news articles. Thus, this work can contribute with transparency to Brazillian citizens regarding the patterns found in their politicians' speeches. Ultimately, this work can be extended with the evaluation of more modern implementations of Topic Models.*

Resumo. *Neste trabalho, pretende-se aplicar os modelos de tópicos e avaliar os resultados a fim de se obter informações relacionandos os assuntos dos discursos dos senadores ao longo do tempo. O contexto atual da sociedade é marcado por um excesso de informações. Dessa forma, uma abordagem de análise automatizada pode facilmente explicitar padrões em grandes coleções de dados. Uma dessas abordagens é a Modelagem de Tópicos. Tal ferramenta consome grandes coleções de documentos e evidencia padrões na forma de tópicos, que são conjuntos de palavras que descrevem um campo semântico. Neste contexto, almeja-se obter padrões ao analisar os tópicos obtidos a partir da extensa base textual de discursos de senadores, disponibilizada pelo Senado Federal. Acredita-se que é possível correlacionar a evolução temporal dos tópicos dos discursos com eventos históricos, políticos e econômicos. Os resultados encontrados são comparados com projetos de leis, datas relevantes e artigos jornalísticos. Com isso, este trabalho pode promover transparência aos cidadãos em relação às informações obtidas dos discursos de seus parlamentares. Por fim, o trabalho pode ser estendido com a avaliação de outras propostas de implementação de Modelos de Tópicos mais modernas.*

1. Introdução

Nos últimos tempos, deu-se início à uma iniciativa promovida pelas principais democracias da atualidade a disponibilizar dados abertos de forma livre e indiscriminada aos cidadãos de seus países. Sabe-se que tornar informações orçamentárias, governamentais

e legislativas acessíveis é essencial para promover transparência e aumentar a efetividade e a *accountability* [1] do setor público [Beghin and Zigoni 2014]. Essa iniciativa, no Brasil, ganhou força com a Lei de Acesso à Informação, de 2011 [2].

Diante deste contexto de dados abertos, o Senado Federal brasileiro disponibiliza para seus cidadãos o Portal da Transparência [3]. Em seu *website*, torna acessíveis tanto informações administrativas quanto legislativas de forma simples. No contexto deste trabalho, o conjunto de dados são as transcrições dos discursos proferidos por senadores. Especificamente para a casa legislativa do Senado Federal, é disponibilizada uma *API* (*Application Programming Interface*) que pode ser consultada pelos cidadão que desejem visualizar as agendas dos parlamentares, seus blocos políticos e também seus discursos proferidos. [4].

A priori, a base de discursos tem registrados centenas de milhares de pronunciamentos desde 1995. É visível como o processamento manual desse tipo de documento pode se tornar impraticável, tanto pelo fator custo, quanto pelo fator tempo. O processo de sumarização, por exemplo, é feito atualmente por profissionais do Senado, que devem realizar esse processo custoso de forma manual.

Neste sentido, uma poderosa ferramenta utilizada para auxiliar de analisar textos são os Modelos de Tópicos. Modelos de Tópicos têm destaque por que evidenciam características gerais de uma grande coleção de documentos, de forma simples e sem a necessidade da atuação humana [Boyd-Graber 2017]. A saída desse tipo de modelo é um conjunto de **tópicos**, ou temas, que indicam os assuntos que os documentos abordam. Um tópico nada mais é do que um conjunto de palavras semanticamente relacionadas que, quando unidas, fazem alusão a um tema específico. O modelo de tópicos mais popular é o LDA (*Latent Dirichlet Allocation*) [Blei et al. 2003]. Ele foi proposto por Blei em 2003 e marcou o início da área de estudos de Modelos de Tópicos Probabilísticos. Essa área tinha como principal objetivo extrair informações de forma eficiente de bases textuais de documentos, utilizando abordagens probabilísticas [Blei et al. 2003, Thiago de Paulo Faleiros 2016].

Para esse trabalho, serão utilizados conjunto de textos de transcrições dos discursos proferidos por senadores. Tem-se a hipótese de que, utilizando técnicas de mineração de texto, como a modelagem probabilística do *LDA*, pode-se, de forma automatizada, extrair diversas informações latentes dos discursos. Com isso, o objetivo principal do presente trabalho é mostrar que, com o auxílio de Modelos Probabilísticos de Tópicos, é possível identificar eventos históricos em discursos de senadores por meio de tópicos latentes.

Sabe-se que, em se tratando de bases de dados textuais, pesquisadores da área de Mineração de Texto oriundos de países lusófonos encontram dificuldades em obter material de qualidade na língua portuguesa, pois infelizmente estes ainda são escassos. Portanto, este trabalho possui como objetivo secundário contribuir com a criação de uma base de discursos normalizada, tratada e pré-processada que contém discursos de senadores

[1] Accountability pode ser entedida como prestação de contas, transparência e responsabilização.

[2] `http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm`

[3] `https://www12.senado.leg.br/transparencia`

[4] `https://legis.senado.leg.br/dadosabertos/docs/`

proferidos entre os anos 1995 e 2019.

Por fim, acredita-se que esse trabalho, ao avaliar tais dados, pode contribuir para a promoção da transparência do Senado Federal, reforçando a democracia no país.

2. Modelos Probabilísticos de Tópicos

Na área de modelos probabilísticos de tópicos, o algorítmo LDA (Latent Dirichlet Allocation) proposto em 2003 por Blei [Blei 2012] é tido como padrão e é a base das soluções seguintes desenvolvidas na área. LDA é um algoritmo generativo que assume que um documento é formado por uma mistura oculta de tópicos. Em geral, o objetivo do LDA é computar a distribuição condicional das variáveis ocultas dadas as variáveis observáveis. Neste contexto, as variáveis observáveis do modelo são o conjunto de palavras de um documento, ao passo que as variáveis ocultas são as efetivas distribuições dos tópicos. Ademais, o modelo é alimentado com hiper-parâmetros que influenciam na granularidade das distribuições. O hiper-parâmetro α ajusta a granularidade da distribuição de tópicos dos documentos, de forma a definir se um documento é formado por um conjunto maior ou menor de tópicos. De forma análoga, o hyper-parâmetro β é responsável pela granularidade da distribuição de palavras em cada tópico [Blei 2012, Thiago de Paulo Faleiros 2016].

2.1. Processo generativo do LDA

Para um melhor entendimento das variávies que compõe o processo generativo do LDA, faz-se necessárias as seguintes definições:

- **Palavra:** Também chamado de *token*. É uma unidade discreta de dado, um item do vocabulário.
- **Documento:** Sequência de n palavras denotadas por $d = (w_1, w_2, ..., w_n)$.
- **Corpus:** Coleção de m documentos de interesse, denotada por $D = \{d_1, d_2, ..., d_m\}$.
- **Vocabulário:** Também chamado de dicionário. É o conjunto das palavras únicas do corpus.

O processo generativo relacionado a geração de um documento d_j do corpus D é detalhado da seguinte forma:

1. Para cada um dos K tópicos (definido pelo modelador), crie as **distribuições de palavras**, definidas como distribuições de dirichlet $\phi_k \sim Dir(\phi_k, \beta)$, que seleciona as palavras que compõe cada um dos K tópicos.
2. Crie a **distribuição de tópicos**, definida como uma distribuição de dirichlet $\theta_j \sim Dir(\theta, \alpha)$ referente ao documento d_j. Essa distribuição define a composição de tópicos por documento.
3. Em seguida, para cada palavra w_i do documento d_j:
   (a) Amostre um tópico aleatório $z_{i,j}$ da distribuição de tópicos θ_j do documento d_j.
   (b) Selecione uma palavra aleatória $w_{j,i}$ dada a probabilidade $p(w_{j,i}|\phi_{z,j,i})$.

Como foi apresentado, o processo gerador é guiado por duas variáveis que descrevem as distribuições do modelo, θ e ϕ, com os respectivos hiper-parâmetros α e β . A distribuição θ é a distribuição de tópicos que define que tópicos fazem parte

da composição de um documento d_j. Dessa forma, possui dimensionalidade K, o número total de tópicos fixo e pré-definido. Analogamente, ϕ define a distribuição de palavras de um tópico k. Portanto, possui dimensionalidade n, o tamanho do vocabulário [Thiago de Paulo Faleiros 2016].

Ademais, uma das características diferenciais do LDA, em relação a uma simples técnica de agrupamento, é que, apesar de todos os documentos do *corpus* compartilharem os mesmos K tópicos, cada documento os apresenta com diferentes proporções, seguindo sua própria distribuição de tópicos θ_j [Blei 2012].

2.2. Inferência do LDA

O processo generativo do LDA descreve como os documentos são gerados dada as distribuições de tópicos, e não um processo descritivo para a extração de tópicos. Para extrair os tópicos, que é o principal problema computacional, deve-se realizar um processo de inferência do modelo descrito pelo LDA. No processo de inferência, apenas as palavras que formam os documentos são entidades conhecidas. As distribuições palavras que compõem os tópicos e a distribuição de tópicos de cada documento são elementos latentes. O desafio computacional dos modelos de tópicos probabilísticos é realizar a *operação inversa* da etapa generativa. Ou seja, dados os documentos, almeja-se *inferir* os elementos latentes supracitados. Essa relação pode ser observada na probabilidade conjunta entre as variáveis não observadas e as variáveis observadas. Essa distribuição tem a seguinte formulação:

$$p(z, w, \phi, \theta|\alpha, \beta) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{j=1}^{M} p(\theta_j|\alpha) \Big(\prod_{i=1}^{V} p(z_{j,i}|\theta_j) p(w_{i,j}|z_{i,j}, \phi_{z,j,i}) \Big) \quad (1)$$

A Equação 1 expressa diversas dependências complexas que, em si, definem o LDA. É desejado saber as distribuições de tópicos para os documentos e as distribuições de palavras para os tópicos. Porém, na prática, além dos hiper-parâmetros α e β, apenas as palavras w dos documentos são conhecidas. Ajustando a equação para isolar as variáveis, à priori, tem-se que:

$$p(z, \phi, \theta|w, \alpha, \beta) = \frac{p(z, w, \phi, \theta|\alpha, \beta)}{p(w)} \quad (2)$$

A Equação 2, que descreve a *posteriori* do LDA, computa a estrutura de tópicos dados os documentos observados. A princípio, esse problema computacional pode ser resolvido pela soma das distribuições conjuntas sobre todas as variáveis não observáveis [Blei 2012]. Contudo, o número de atribuições cresce exponencialmente, de forma a tornar o problema intratável computacionalmente. Então, a área de tópicos probabilísticos propõe uma série de métodos eficientes para aproximar distribuições à posteriori como esta. Em geral, considera-se duas categorias de algoritmos de aproximação: algoritmos baseados em amostragem e algoritmos variacionais [Blei 2012].

Algoritmos baseados em amostragem, como sugere o nome, obtém amostras da posteriori e tentam aproximá-la por meio de uma distribuição *empírica*. Já os métodos variacionais, em contraste aos métodos baseados em amostragem, propõem uma abordagem determinística, ao invés de uma abordagem empírica. Com isso, o problema computacional é visto como um desafio de otimização [Blei 2012].

2.3. Avaliação de Modelos de Tópicos

Uma métrica utilizada para avaliar a qualidade de um modelo probabilístico de tópicos é a verossimilhança *held-out* do modelo [Wallach et al. 2012]. Neste contexto, a questão principal é estimar o quão bem o modelo pode descrever corretamente documentos nunca vistos ou *held-out* (retidos). Neste contexto, criou-se o conceito de *perplexidade*. A métrica da perplexidade, matematicamente, indica o quanto um modelo se desvia do conjunto de treino em relação ao conjunto de teste [Thiago de Paulo Faleiros 2016, Blei et al. 2003]. Um valor de perplexidade inferior sugere um modelo capaz criar generalizações de forma mais eficiente.

A avaliação de modelos de tópicos por meio de métricas como perplexidade ignora a representação interna dos modelos [Chang et al. 2009]. Ou seja, estas métricas puramente estatísticas mostram-se pouco úteis para identificar a real coerência semântica nos tópicos gerados pelos modelos. Com isso, Chang et al. [Chang et al. 2009] rompem com a forma com que modelos de tópicos eram avaliados e propõe um método novo, que busca medir o quão interpretáveis são os tópicos, utilizando o julgamento humano. A relevância do trabalho de [Chang et al. 2009] foi mostrar que as métricas de avaliação tradicionais são pouco eficientes em determinar a real coerência de tópicos criados de modelos probabilísticos. Além disso, o trabalho de [Chang et al. 2009] expôs a conclusão contra-intuitiva de que os propostos métodos de avaliação de coerência semântica apresentam uma correlação **negativa** em relação à métricas probabilísticas como a perplexidade [Chang et al. 2009]. Isso significa que tentativas de otimização dos valores de perplexidade podem não gerar tópicos coerentes ou interpretáveis por humanos.

Newman et al. [Newman et al. 2011] apresentou em seu trabalho a métrica PMI (*Pointwise mutual information*), que calcula um escore baseado na presença de pares de palavras dos tópicos em grandes bases de dados externas como *Wikipedia*, *Google News* e *WordNet*. A metodologia proposta foi capaz de identificar tópicos pobres facilmente e, em geral, apresentou concordância com o julgamento humano.

O PMI define uma métrica de associação entre palavras. Para um dado tópico $w = (w_i, ..., w_{10})$, o escore é calculado como a mediana dos valores de PMI para cada par do tópico:

$$PMI\_score(w) = mediana\{PMI(w_i, w_j), ij \in 1...10\}, \tag{3}$$

onde o PMI de um par de palavras é calculado como:

$$PMI(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{4}$$

Note que um par de palavras semanticamente relacionadas apresenta um alto valor de PMI.

No trabalho de Lau et al. [Lau et al. 2014], dá continuidade aos esforços empregados na criação do PMI e avalia uma alternativa normalizada. A utilidade da normalização é que a escala de valores possui interpretação fixa ($NPMI \in [-1, 1]$), de forma que pode ser um bom candidato a substituir o PMI. Além disso, a normalização acaba com a tendência do PMI em favorecer palavras de baixa frequência.

O NPMI possui interpretação simples. Quando duas palavras aparecem juntas tem-se $NPMI(w_i, w_j) = 1$. Se as palavras são distribuídas de forma independente,

então $NPMI(w_i, w_j) = 0$. Por fim, quando as palavras são observadas separadamente, mas nunca em conjunto, tem-se que $NPMI(w_i, w_j) = -1$.

3. Descrição dos dados

O Senado não fornece os dados completos para *download* direto. É necessário, então, realizar uma coleta iterativa e programática dos dados. Para realizar esta tarefa, foi desenvolvido um *script* para a coleta dos dados. Utilizou-se o serviço *ListaSenadorService* e *PlenarioService* e, respectivamente, coleta dados da lista de senadores e coleta dados por tempo. Foi codificada uma rotina que *mescla* os resultados dos dois serviços. O que revelou-se uma co-ocorrência de **113,788** discursos.

Após criada a base de dados final, retorna-se a atenção para a natureza de cada discurso. Em destaque aos metadados de cada discurso, aquele identificado como —*TipoUsoDaPalavra*—, presente nas respostas da API, tem especial relevância para esse trabalho. Ele explicita o tipo de pronunciamento relacionado a um determinado discurso. No contexto do Senado Federal, é sabido que existem reuniões e encontros diversos, cada um com uma finalidade específica. O tipo de pronunciamento de interesse para esse trabalho é o tipo "Discurso", que compõe a maior parte dos pronunciamentos. Esta categoria de pronunciamento descreve discursos proferidos por qualquer senador, de forma livre. Tipicamente, estes pronunciamentos são direcionados à base apoiadora dos parlamentares e não dependem de nenhum direcionamento ou temática prévia, ao contrário dos outros tipos.

Em conclusão, realizou-se uma filtragem dos discursos usando como critério o tipo de pronunciamento. O conjunto resultante contém **81,858** discursos únicos.

3.1. Pré-processamento

A etapa de pré-processamento mostrou-se a que mais impactou na qualidade dos tópicos. Por isso, foi necessário adotar uma abordagem exploratória e iterativa. Seguiu-se as seguintes etapas: a) Remoção de números, acentos, caracteres especiais, letras maiúsculas, múltiplos espaços em branco e quebras de linha; b) Nomes compostos de estados brasileiros foram concatenados para que sua informação semântica fosse preservada nos tópicos; c) Removeu-se palavras com menos de 4 letras. Foi determinado experimentalmente que palavras com um número inferior de letras não carrega informação relevante; d) Excluiu-se *stopwords*, que é um conjunto de palavras rotuladas como "vazias" principalmente por não possuírem valor semântico. Tipicamente são artigos, preposições e outras palavras excessivamente frequentes. Inicialmente foram usadas as *stopwords* das bibliotecas Python Spacy[5] e NLTK[6]. Ademais, para o caso deste trabalho, criou-se modelos LDA e tomou-se nota, de forma manual e experimental, das palavras "irrelevantes" que compunham os tópicos gerados. Essas palavras foram, de forma iterativa, adicionadas à lista de *stopwords* total. Dessa forma, assegurou-se que o *corpus* continha o mínimo de ruído possível; e) Aplicou-se a técnica de *stemming*, utilizando a biblioteca *NLTK*. Para garantir melhor legibilidade adiante, palavras com um mesmo *stem* foram substituídas pela ocorrência mais frequente em todo o *corpus*; f) Finalmente, realizou-se o processo de filtragem de extremos no *corpus*. Adotou-se a metodologia observada na literatura

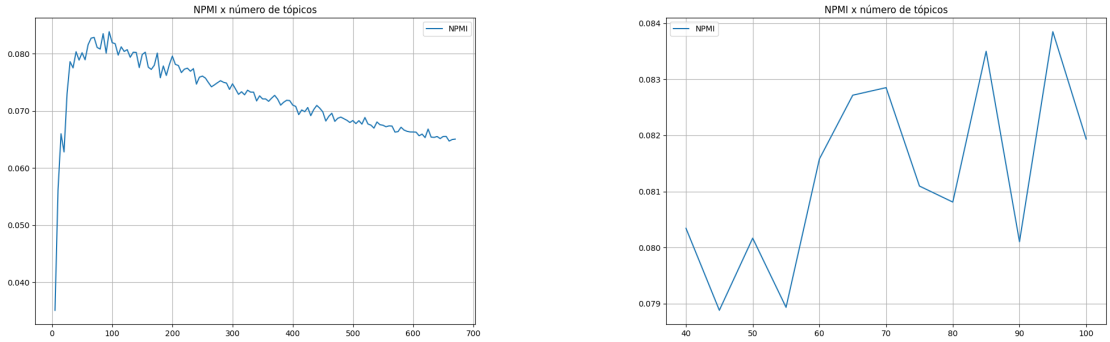
[5] https://spacy.io/
[6] https://www.nltk.org/

[Moreira 2020] - removeu-se palavras muito frequentes e pouco frequentes, isto é, que aparecem em mais de 90% e em menos de 0.5% de discursos, respectivamente. Apenas esta filtragem reduziu o tamanho do vocabulário de **262,601** palavras para **5,811** palavras distintas.

4. Extração dos tópicos

Para gerar as matrizes documento-tópico, utilizou-se a biblioteca Python *Gensim* com a implementação *MALLET* [7], por sua capacidade conhecida de gerar tópicos de qualidade, em comparação a outras implementações. Em termos de parâmetros, utilizou-se o recurso de α automático, que permite que o algoritmo encontre automaticamente um valor do hiper-parâmetro α ótimo para o corpus em questão. Além disso, foi utilizado o valor $10,000$ para o parâmetro do número de iterações.

Um dos desafios do problema de modelagem probabilística de tópicos é a seleção do número de tópicos K ideal, que depende tanto da coleção de documentos quanto da aplicação de interesse. Esta escolha tomou como direcionamento inicial a métrica $NPMI$. Computou-se então, o valor de $NPMI$ para diversos modelos LDA com diferentes número de tópicos. Os resultados estão descritos nas figuras 1(a) e na 1(b).



(a) NPMI para k entre 5 e 650

(b) NPMI para k entre 40 e 100

Figure 1. Qualidade dos tópicos pelo número de tópicos.

A Figura 1(a) expõe os valores de $NPMI$ para um grande intervalo entre $k = 5$ e $k = 650$. É visível um expressivo pico na vizinhança de $k = 100$ seguido de um rápido decrescimento de coerência para altos valores de k. Já a Figura 1(b), por sua vez, apresenta os mesmos valores, mas ampliados em torno do pico mencionado. O gráfico indica como candidatos do número de tópicos K os valores $\{40, 50, 60, 65, 85, 95\}$, que são os picos mais expressivos de coerência. Este trabalho optou por selecionar como valor final $k = 65$. Essa escolha se dá pelo fato do valor ser o primeiro grande pico expressivo, se diferenciando dos demais apenas por poucos milésimos. Além disso, o valor mediano oferece bom equilíbrio entre tópicos genéricos e tópicos específicos, que apresenta suficiente diversidade temática.

[7] `http://mallet.cs.umass.edu/`

5. Análise de tópicos por Tempo

O objetivo principal desse trabalho é comprovar a hipótese de que os tópicos dos discursos podem explicitar informações a respeito de contextos históricos, políticos e econômicos.

A metodologia proposta por esse trabalho sugere uma observação da evolução histórica dos tópicos ano a ano. Para cada tópico k e ano a, sugere-se a definição das seguintes métricas avaliativas:

- **Contribuição média**: Supõe-se que todos os discursos do ano a são combinados em 1 único discurso (discurso médio). Esta métrica observa a distribuição temática desse discurso geral e indica a contribuição percentual do tópico k.
- **Contagem de discursos dominantes**: É realizada uma contagem percentual do número de discursos do ano a em que o tópico k é o mais expressivo. Ou seja, é a contagem de quantos são os discursos que falam majoritariamente do tópico k no ano a.

Estas métricas, combinadas, se mostram úteis para que se possa verificar a evolução de um tema ao longo dos anos, de acordo com os discursos proferidos pelos parlamentares.

6. Análise dos Resultados

Após o processo de coleta massiva, tratamento e processamento da base de discursos do Senado Federal, caracteriza-se os resultados encontrados. Os tópicos mais frequentes (com mais discursos relacionados ao tópico) foram "corrupção", "legislação" e "programas de desenvolvimento". Por outro lado, revelou-se que os tópicos abordados com menos frequência foram "energia elétrica", "aviação" e "povos indígenas".

O tópico 5, cujo tema é "esporte", é retratado na Figura 2. Observa-se que houve picos expressivos nos anos de 2009 e 2014. Os respectivos picos são justificados pela vitória da seleção brasileira na Copa das Confederações de 2009 e na participação da Seleção na Copa do Mundo de 2014. Ainda, ao comparar os anos de 2008 e 2009, pode-ser constar que houve um aumento de aproximadamente 90% no número de discursos que abordam principalmente o tema em questão.

A Figura 3 apresenta um exemplo curioso. Ao invés de explicitar um evento histórico em particular, o gráfico torna visível um padrão relacionado à uma característica política. O tópico 27 diz respeito a questões eleitorais, votos e candidaturas. O gráfico expõe que há um claro padrão periódico relacionado aos picos observados. Os picos acontecem aproximadamente de 4 em 4 anos, período que coincide com as eleições para presidente, governadores, deputados e senadores. Nos principais discursos desse tópico, observa-se que os senadores usam seu espaço político para expor as principais conquistas de seu partido e governantes durante o mandato que se encerra.

Dentre todos os exemplos, o tópico 33 é provavelmente o mais icônico. O gráfico da Figura 4 tem um pico expressivo entre os anos de 2004 e 2006. Este foi precisamente o período em que foram apurados esquemas de corrupção na CPMI dos correios. O principal desses escândalos foi o esquema do mensalão, o mais conhecido esquema de corrupção e desvio de dinheiro da história do Brasil.

O tópico 41, representado pela Figura 5, tem como tema a aviação e, especificamente, acidentes aéreos. Os picos encontram-se nos anos de 2007 e 2014. Estes foram

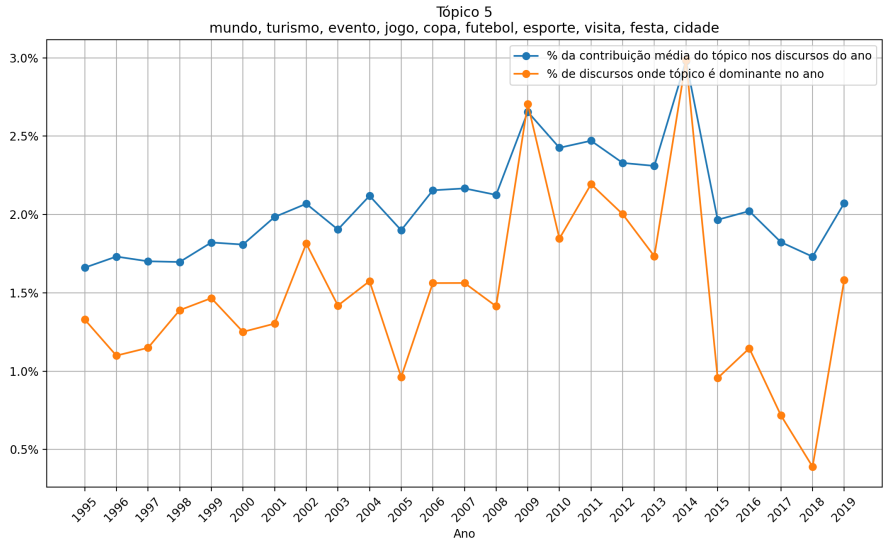


Figure 2. Evolução do tópico 5 entre 1995 e 2019

anos em que ocorreram conhecidos acidentes aéreos no Brasil e no exterior. O primeiro pico está relacionado ao acidente do voo TAM 3054, considerado a maior tragédia da aviação brasileira. Já o segundo, por outro lado, associa-se ao voo Malaysia Airlines 370, que causou comoção mundial com o desaparecimento de uma aeronave comercial cujos destroços nunca foram localizados. Nos discursos desse tópico, os senadores relembram as tragédias e reivindicam por mais segurança nas aeronaves e aeroportos brasileiros.

7. Resultados negativos

Apesar da preliminar coerência observada em grande parte dos tópicos encontrados, existiram também tópicos cujos gráficos não trouxeram boas conclusões. O exemplo inicial é o tópico 63, apresentado na Figura 6, que aborda de cultura em geral. No gráfico observado, não existem picos delimitados ou qualquer outro tipo de padrão. Neste caso, o entendimento que se extrai é que não existiram grandes acontecimentos em torno do universo semântico em questão.

Outra possibilidade é o fato de que o tema não atrai a atenção de um número grande de políticos, mais voltados para uma agenda fortemente concentrada em questões econômicas.

8. Conclusões e Trabalhos Futuros

O presente trabalho se empenhou em atingir dois objetivos principais. O primeiro é a criação e disponibilização de uma extensa base de discursos de senadores, que pode ser acessada por meio do seguinte endereço público: *https://github.com/VictorLandim/brazilian-senators-speeches*. O segundo objetivo é a verificação de se a decomposição temática dos discursos por meio do LDA é capaz de evidenciar fatos históricos ao longo dos anos.

Em se tratando dos resultados relacionados à análise temporal, apesar da imprecisão em observar correspondências históricas para cada um dos tópicos, temas populares

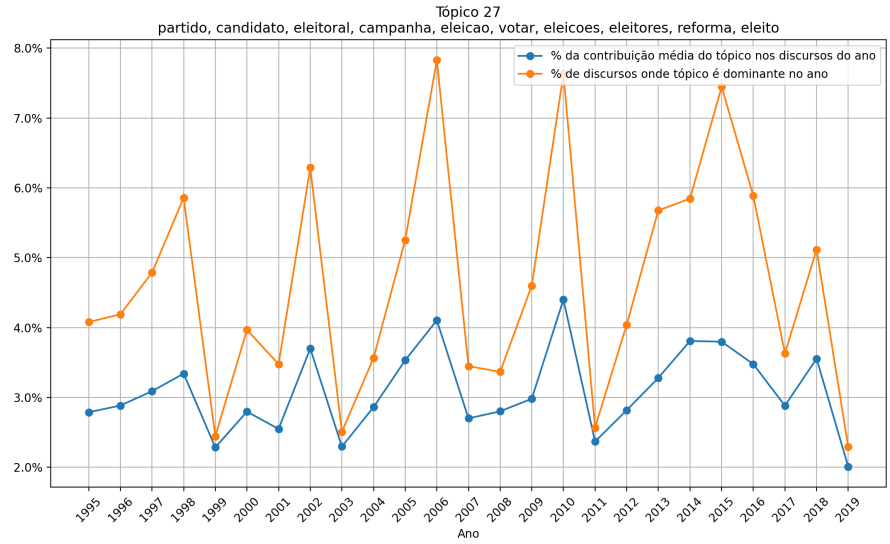


Figure 3. Evolução do tópico 27 entre 1995 e 2019

no contexto brasileiro como Corrupção, Futebol e Eleições puderam ser facilmente identificados temporalmente. Por fim, especula-se que a implementação proposta seja, ainda, incapaz de evidenciar as nuances contidas em temas com baixa relevância proporcional nos pronunciamentos. Há que se considerar que, a despeito de o parlamento dar uma resposta para os anseios da população e aos acontecimentos conjunturais, trata-se de uma instituição muito plural, que representa interesses e segmentos diversos. Diversos são, portanto, os tópicos abordados pelos senadores.

Julga-se possível que outros estudos possam propor avanços em diferentes frentes no contexto deste trabalho. Como trabalho futuro, pretende-se a experimentação com implementações de modelos de tópicos iterativos, como os *Interactive Topic Models* [Boyd-Graber et al. 2014]. Esse tipo de abordagem permite que, na prática, a atuação humana direcione os tópicos criados. Por fim, no contexto da análise temporal dos tópicos, sugere-se o uso de extensas bases de notícias como entrada para outro conjunto de modelos LDA. Dessa forma, é possível comparar as variações temáticas nos pronunciamentos com aquelas nas notícias, de forma automatizada.

References

[Beghin and Zigoni 2014] Beghin, N. and Zigoni, C. (2014). Avaliando os websites de transparência orçamentária nacionais e sub-nacionais e medindo impactos de dados abertos sobre direitos humanos no brasil. *Instituto de Estudos Socioeconômicos*.

[Blei 2012] Blei, D. M. (2012). Probabilistic topic models, surveying a suite of algorithms that offer a solution to managing large document archives. *Communications of the acm*.

[Blei et al. 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

[Boyd-Graber 2017] Boyd-Graber, J. (2017). Applications of topic models. *Department of Computer Science, umiacs, Language Science - University of Maryland*.

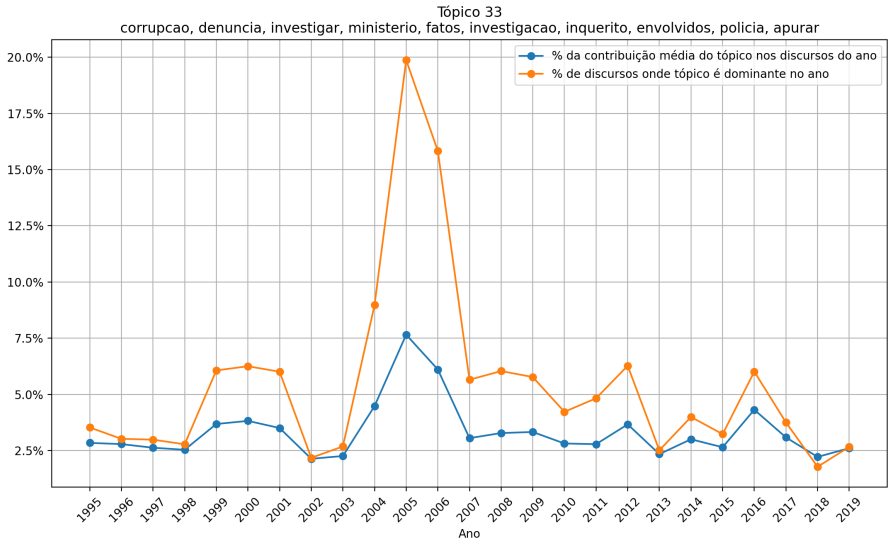


Figure 4. Tópico 33 dos discursos proferidos entre os anos de 1995 e 2019

[Boyd-Graber et al. 2014] Boyd-Graber, J., Mimno, D., and Newman, D. (2014). *Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements*. CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, Florida.

[Chang et al. 2009] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

[Lau et al. 2014] Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

[Moreira 2020] Moreira, D. (2020). Com a palavra os nobres deputados: ênfase temática dos discursos dos parlamentares brasileiros. *DADOS, Rio de Janeiro*.

[Newman et al. 2011] Newman, D., Karimi, S., and Cavedon, L. (2011). External evaluation of topic models. *ADCS 2009 - Proceedings of the Fourteenth Australasian Document Computing Symposium*.

[Thiago de Paulo Faleiros 2016] Thiago de Paulo Faleiros, A. d. A. L. (2016). Modelos probabilÍsticos de tÓpicos: Desvendando o latent dirichlet allocation. *Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP*.

[Wallach et al. 2012] Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2012). Evaluation methods for topic models. *Communications of the acm*.

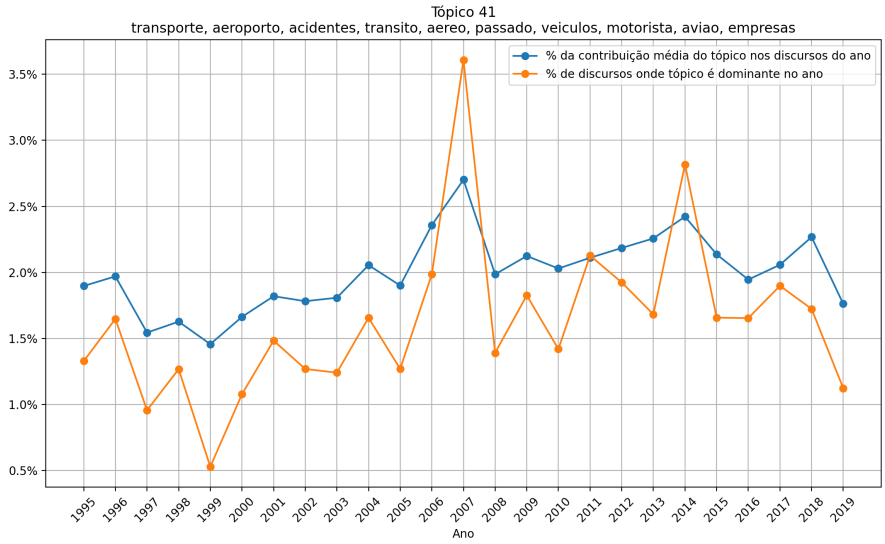


Figure 5. Evolução do tópico 41 entre 1995 e 2019

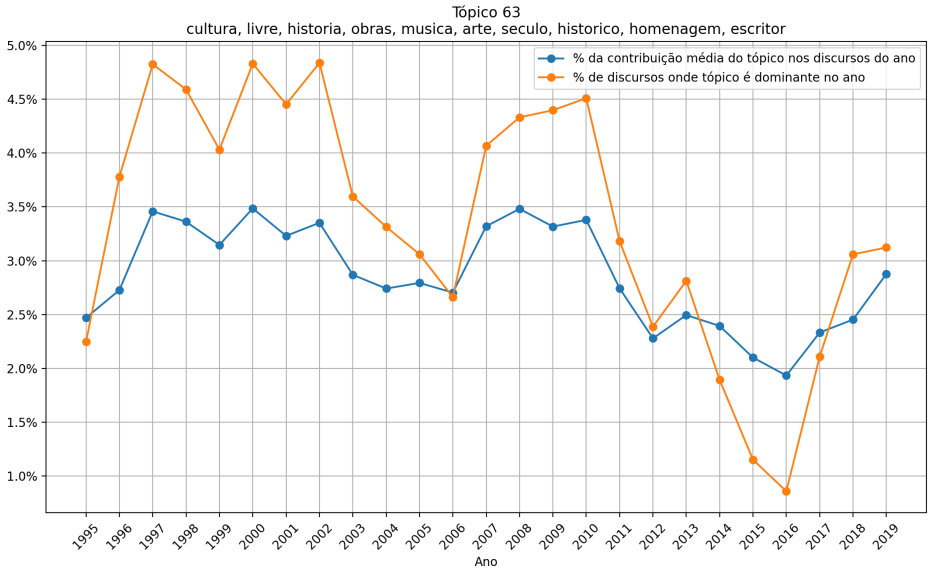


Figure 6. Evolução do tópico 63 entre 1995 e 2019