

Análise semântica latente aplicada a projetos de lei

1st Carlos Cardoso Dias

Programa de Eng. de Sistemas e Comp. (PESC/COPPE)

Universidade Federal do Rio de Janeiro (UFRJ)

Rio de Janeiro, Brasil

cdias@cos.ufrj.br

2nd Tales Mello Paiva

Programa de Eng. de Sistemas e Comp. (PESC/COPPE)

Universidade Federal do Rio de Janeiro (UFRJ)

Rio de Janeiro, Brasil

tmpaiva@cos.ufrj.br

Abstract—This work aims to use latent semantic analysis in bills from the House of Representatives in Brazil to extract topics and compare based on date range and political parties. The results are presented in wordcloud format. The analysis concluded that bills are following country's phenomena and that the party with most bills are indeed following its guidelines. The tool developed in this work is public available to help citizens and researchers to conduct its own researches.

Index Terms—Isa, pl, nlp, topics

I. INTRODUÇÃO

A análise por semântica latente (LSA) é uma técnica de processamento de linguagem natural (NLP) que busca uma relação entre os termos de um documento e os conceitos, ou ideias, por detrás dos mesmos. Essa técnica se baseia no uso da decomposição em valores singulares da matriz termo-documento para redução de dimensionalidade desta matriz, desta forma, aglutinando palavras em conceitos [2].

A Câmara dos Deputados, por meio do projeto Dados Abertos, oferece acesso a uma coleção de documentos de cunho legislativo tais como proposições, frentes parlamentares, deputados, discursos, despesas de parlamentares, dentre outros. Os dados disponíveis são alimentados por sistemas internos da câmara, sendo disponibilizados ao público de forma gratuita com o objetivo de que entidades da sociedade civil possam desenvolver aplicações alimentadas pelos dados disponibilizados [1].

Ao realizar a análise por semântica latente sobre as coleções de documentos disponibilizados pela Câmara sobre projetos de lei, a motivação deste trabalho é capturar as principais ideias que circulam na esfera política por período, ou partido, permitindo a identificação dos principais conceitos e tópicos em pauta, evidenciando tendências e anomalias e também simplificando o processo de acompanhamento legislativo pelo cidadão.

O trabalho será divido da seguinte forma: a seção II trará os objetivos a serem alcançados; a seção III irá discorrer sobre a metodologia de aquisição e tratamento dos dados; a seção IV apresenta os resultados e os discute; tendo a conclusão do trabalho na seção V.

II. OBJETIVOS

- Objetivo Geral: Identificar e apresentar, numa forma de fácil compreensão, os principais conceitos e ideias

associados aos projetos de lei propostos na Câmara dos Deputados do Brasil.

- Objetivos Específicos:

- Filtrar por período de tempo, de forma a identificar alterações nas temáticas mais evidentes ao longo dos anos.
- Filtrar por partido, de forma a identificar tendências e anomalias sob a legenda de um partido.
- Combinar os filtros de partido e período de tempo, de forma a identificar alterações e evoluções nos comportamentos dos partidos ao longo dos anos.
- Apresentar os filtros de uma maneira fácil de manipulá-los, de forma a possibilitar a experimentação, visando favorecer a descoberta de tendências e anomalias.

III. METODOLOGIA

Para compor a base de dados foram utilizados os dados de Proposições, Autores e Temas para todos os anos disponíveis (1934-2022) no site do projeto Dados Abertos da Câmara¹.

As proposições foram filtradas para conter apenas:

- Projetos de Lei (*siglaTipo* = “PL”)
- propostos por Deputados (*tipoAutor* = “Deputado”)

A base foi unificada a partir do conjunto de proposições utilizando *idProposicao* para a base de autores e *uriProposicao* para temas.

A base final apresenta 124.848 registros com *ano* e *ementa* devidamente preenchidos. Dentre esses, há siglas referentes a 64 partidos ao todo, apesar do Brasil contar hoje com apenas 32 partidos ativos registrados de acordo com o Tribunal Superior Eleitoral [5]. Há registros pertencentes a 32 partidos que foram extintos, incorporados a outro ou cuja sigla mudou devido a alguma mudança de nome ao longo da história de um determinado partido.

Ademais, um total de 11.933 registros, o equivalente a 9,56% da base, não conta com registro de sigla do partido do autor da proposição, estando o campo nulo ou com as especificações “PNI” (“Partido Não-Identificado”) e “SPART”/“S.PART.” (“Sem Partido”). Esses registros foram agrupados sob a sigla “NI” (“Não Identificado”), de forma a facilitar a análise. Algumas outras correções foram realizadas a respeito de nomenclatura das siglas, visando unificar grafias

¹<https://dadosabertos.camara.leg.br/swagger/api.html#staticfile>

distintas de um mesmo partido. Por fim, as proposições sob a sigla “PRP” foram separadas em 2 partidos: as proposições até 1965 são do “Partido de Representação Popular”, quando o mesmo tornou-se “ARENA”, e, portanto, foram renomeadas para “PRP*”; já referente às proposições de 1989 em diante, “PRP” é a sigla do “Partido Republicano Progressista”, atual “PATRIOTA”, e a sigla foi mantida como estava.

Quanto aos temas das proposições, a Câmara dos Deputados conta com 32 temas pré-definidos. A base conta com 50.560 registros cujo tema não se encontra declarado, havendo assim 74.288 registros com tema e ementa devidamente preenchidos, os quais apresentam possibilidades de análises interessantes de serem exploradas. Os registros cujo tema encontrava-se nulo não foram expurgados da base, uma vez que a ementa encontrava-se presente.

Para o processamento do texto, os campos *ementa*, *ementaDetalhada* e *keywords* foram concatenados, sendo removidos os registros vazios após a concatenação. Os documentos foram pré-processados para a remoção de acentos; transformação de letras para minúsculas; remoção de stopwords padrão (nltk) acrescido de stopwords relevantes encontradas na base (‘lei’, ‘federal’, ‘nacional’, ‘aplica’, ‘aplicando’, ‘aplicar’, ‘aplicado’, ‘aplicados’, ‘vigencia’, ‘programa’, ‘institui’, ‘publica’, ‘publico’, ‘sobre’); remoção de palavras não compostas inteiramente por caracteres alfabéticos; e remoção de sufixos utilizando o stemmer RSLP [4].

Após o pré-processamento montou-se a matriz termo-documento da coleção, sendo utilizada a métrica TD-IDF (do inglês, “Term Frequency-Inverse Document Frequency”) e em seguida foi feito o SVD (do inglês, “Singular Value Decomposition”) desta matriz para obter os termos semânticos ocultos, utilizando o valor de 32 componentes, ou tópicos, para a matriz diagonal de valores singulares, de forma a corresponder à mesma quantidade de temas pré-definidos pela Câmara.

A decomposição em valores singulares da matriz termo-documento, \mathbf{A} , resulta em três matrizes, onde

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \quad (1)$$

Sendo \mathbf{A} uma matriz com m termos e n documentos, e t o número de tópicos escolhido para a decomposição SVD de \mathbf{A} , temos que

- \mathbf{U} é uma matriz $m \times t$, que pondera os termos em relação aos tópicos
- Σ é uma matriz diagonal $t \times t$ que indica a relevância do i -ésimo tópico em relação à composição da base
- \mathbf{V}^T é uma matriz $t \times n$ que pondera os documentos em relação aos tópicos encontrados

Desta forma, um tópico, quando visto sob a ótica de álgebra linear, são os vetores que compõem a base de \mathbf{U} . Estes vetores atribuem um valor numérico para cada um dos m termos presentes no vocabulário da base. Para a caracterização de um tópico foi tomado um valor de *threshold* e um limite máximo de 5 termos de forma que para cada vetor da base de \mathbf{U} , o vetor foi ordenado pelo valor absoluto de suas

componentes e foram selecionados os k termos maiores em módulo que o *threshold*, com $k \leq 5$. A relevância, obtida da matriz Σ , para o conjunto de 32 tópicos, obtidos da matriz \mathbf{U} conforme descrito, considerando toda a base de dados pode ser vista na figura 1, onde os caracteres ‘+’ e ‘-’ no início de cada palavra que compõe o tópico está associado, respectivamente, ao valor positivo e negativo do termo na base de \mathbf{U} .

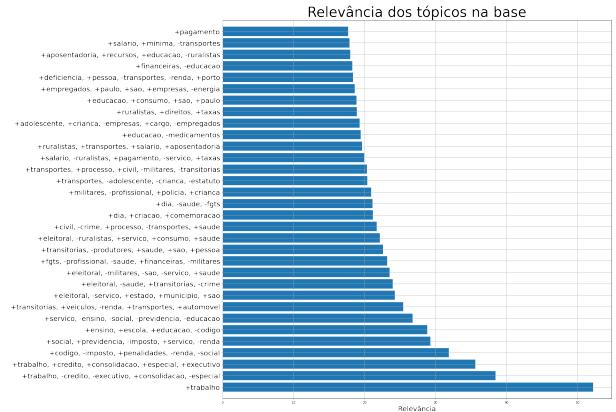


Fig. 1. Tópicos x Relevância.

Para facilitar a visualização dos termos que compõem um tópico e sua importância, tanto na pertinência ao tópico quanto à caracterização da base, foi utilizada a biblioteca *wordcloud*² para a exibição dos dados, tendo seu comportamento estendido para comportar a presença de palavras repetidas com frequências distintas (caso onde uma mesma palavra está presente em múltiplos tópicos). Desta forma, foi utilizado o padrão onde cada cor utilizada na *wordcloud* corresponde a um tópico e o tamanho do termo corresponde a sua influência com relação ao tópico ($\mathbf{U}[i]$) ou com relação à base ($\mathbf{U}[i]\Sigma[i]$). Para a análise considerando todos os documentos da base, a relevância de cada termo com relação ao seu tópico e de cada termo com relação à base podem ser vistas, respectivamente, nas figuras 2 e 3.



Fig. 2. Relevância termo-tópico.

IV. RESULTADOS

Como resultados primários, obtidos de forma a guiar as demais análises posteriores, temos que 6 temas correspondem

²<https://pypi.org/project/wordcloud/>



Fig. 3. Relevância termo-base.

a mais da metade das proposições com tema declarado, dentre os 32 temas possíveis:

- 1) Administração Pública - 12,35%
- 2) Trabalho e Emprego - 11,83%
- 3) Direitos Humanos e Minorias - 7,52%
- 4) Direito Penal e Processual - 6,32%
- 5) Educação - 6,12%
- 6) Finanças Públicas e Orçamento - 6,03%

Podemos observar também que apenas 8 partidos, dentre os 64 presentes na base, são responsáveis por mais da metade das proposições:

- 1) PT - 13,19%
- 2) PMDB - 9,01%
- 3) PSDB - 5,91%
- 4) PSD - 5,43%
- 5) PTB - 4,80%
- 6) PDT - 4,21%
- 7) MDB* - 3,40%
- 8) PFL - 3,01%

Tal concentração é ainda mais latente quando considerase que o partido cuja sigla é apresentada como "MDB*" é, na verdade, o próprio "PMDB" durante os anos entre 1966 e 1979, o qual em 2017 tornou a denominar-se MDB, constando na base como "MDB", de forma a diferenciar os 3 momentos do mesmo partido. Assim, temos que, efetivamente apenas 7 partidos respondem por mais da metade das proposições da base.

Para a verificação da relevância dos Projetos de Lei quando comparados ao cotidiano, foram analisados os períodos de 2000-2019 (figura 4) e 2000-2022 (figura 5), com especial foco nas mudanças entre as bases para este período.



Fig. 4. Relevância termo-base para o período de 2000-2019.

Na figura 5, pode-se notar o surgimento da palavra "pandemia" dentre os termos relevantes para a classificação em mais de um tópico, além do aumento da relevância de "saúde" com relação a base da figura 4.



Fig. 5. Relevância termo-base para o período de 2000-2022.

Além disso, quando comparados os tópicos gerados para os dois períodos, os termos "violência" e "mulher" apareciam nos tópicos "violência, mulher, doméstica, rural" e "idoso, violência, mulher" para o período de 2000-2019, o primeiro sugerindo um contexto específico regional de violência doméstica e o segundo sugerindo um contexto genérico aplicado ao grupo. Já para o período de 2000-2022, os mesmos termos aparecem nos tópicos "violência, profissional, mulher", "violência, mulher, doméstica", "fgts, violência, mulher", sugerindo a aparecimento de medidas específicas para o contexto de violência doméstica e também da profissional mulher para o segundo período. Quando pesquisado em sites de notícias³, verifica-se que há um aumento dos casos de violência contra a mulher durante o período da pandemia.

Quando analisado especificamente o período de 2020-2022, com a escolha de categorização em apenas 8 tópicos, podemos observar se o impacto da pandemia de coronavírus no Brasil teve relevância do ponto de vista legislativo na esfera nacional, como mostra a figura 6.

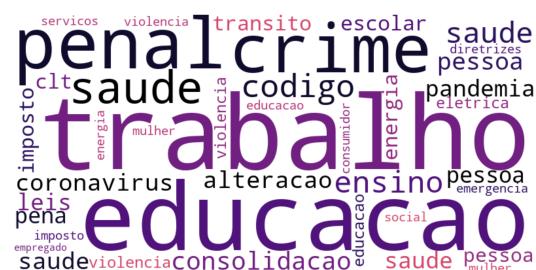


Fig. 6. Relevância termo-base para o período de 2020-2022.

Por fim, também foi selecionado o partido com maior número de proposições de Projetos de Lei da base ("PT") e feita a análise para a extração de 16 tópicos baseados nestes projetos. Foi analisado os tópicos e termos relevantes para formação do tópico, de acordo com a figura 7.

³<https://agenciabrasil.ebc.com.br/saude/noticia/2021-08/violencia-contra-mulheres-cresce-em-20-das-cidades-durante-pandemia#staticfile>



Fig. 7. Relevância termo-tópico para o PT.

Segundo o site oficial do partido, “o PT surgiu como agente promotor de mudanças na vida de trabalhadores da cidade e do campo, militantes de esquerda, intelectuais e artistas” [3]. Pode-se observar que “trabalhador”, palavra de destaque na descrição do partido, é também uma palavra com peso relevante e que figura em mais de um tópico. Além disso, para a caracterização de tópicos também figuram palavras como: benefício, social, criança, adolescente, rural, ensino, dentre outras. O gráfico na figura 8 mostra os tópicos definidos e sua relevância com relação à base composta por projetos de lei do PT.

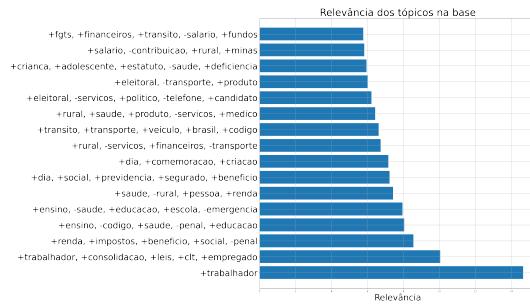


Fig. 8. Tópico x Relevância para o PT.

V. CONCLUSÃO

Como visto na seção IV, é possível estabelecer uma relação direta de acontecimentos relevantes do cotidiano com a alteração de temáticas definidas pela análise de semântica latente dos projetos de lei. Também é possível, através da delimitação da base de dados pelos filtros criados, avaliar o trabalho de partidos políticos de acordo com a premissa destes partidos através da criação de tópicos da base. Por fim, a representação dos tópicos através de uma *wordcloud* permite a rápida compreensão e visualização das temáticas do governo, facilitando o compartilhamento da informação.

Além disso, todo o código desenvolvido neste trabalho está disponibilizado em <https://colab.research.google.com/drive/1pKLW2xDtWffogS3lopReLua3XMLZuQcF?usp=sharing> como uma ferramenta que conta com campos de input disponíveis da plataforma que permitem a customização das análises por período, partido, número de tópicos e demais opções do algoritmo de acordo com a necessidade do usuário.

Como sugestão de trabalhos futuros, visualiza-se a criação de uma interface mais amigável ao uso, e a disponibilização dessa ferramenta em um ambiente público e de fácil acesso. Além disso, pode-se enriquecer a escolha das siglas a partir de um mapeamento da evolução dos partidos, a partir das suas fusões, incorporações e mudanças de nomenclatura. Na mesma linha, seria interessante uma forma de visualizar o efeito que essas fusões e incorporações de partidos tem nas novas legendas resultantes, de forma a visualizar a interferência ideológica antes e depois de tais eventos. Além disso, a análise aplicada neste trabalho pode ser estendida a outros documentos da mesma base, como a descrição detalhada dos projetos submetidos ou até mesmo o conteúdo dos discursos políticos, o que abriria espaço para novas análises de congruência.

REFERENCES

- [1] Câmara dos Deputados - Dados Abertos, <https://dadosabertos.camara.leg.br/index.html>, 22/06/2022.
- [2] G. Xexéo, “Tópicos em Busca e Aprendizado de Máquina com Textos com exemplos em Python e KNIME”.
- [3] Nossa história — Partido dos Trabalhadores, <https://pt.org.br/nossa-historia/>, 22/06/2022.
- [4] Orengo, V.M. and C.R. Huyck, “A Stemming Algorithm for the Portuguese Language”, in 8th International Symposium on String Processing and Information Retrieval (SPIRE). 2001: Laguna de San Raphael, Chile. p. 183-193.
- [5] Partidos políticos registrados no TSE — Tribunal Superior Eleitoral, <https://www.tse.jus.br/partidos/partidos-registrados-no-tse/registrados-no-tse#staticfile>, 25/06/2022.