MÉTODOS DE ANÁLISE AUTOMATIZADA DE CONTEÚDO APLICADOS AOS DISCURSOS PARLAMENTARES

AUTOMATED CONTENT ANALYSIS METHODS APPLIED TO PARLIAMENTARY SPEECHES

MÉTODOS AUTOMATIZADOS DE ANÁLISIS DE CONTENIDO APLICADO A DISCURSOS PARLAMENTARIOS

Ricardo Modesto Vieira[*]

Resumo: Ao reduzir substancialmente o tempo e os custos de se fazer pesquisa utilizando textos, os métodos de análise automatizada de conteúdo tornaram factível a investigação de grandes acervos, como os discursos parlamentares, quantitativamente. Esse artigo identifica o arcabouço de conceitos dessa nova fronteira do conhecimento, descreve um típico projeto de texto como dado, apresenta as principais metodologias e suas aplicações aos discursos parlamentares, mostra os principais desafios e as boas práticas dessa área de pesquisa e, finalmente, sugere como aplicar esses métodos em acervos de Casas Legislativas. A intenção não é esgotar a gama de métodos, técnicas e modelos; mas fornecer um guia para que esses métodos, cada vez mais populares entre cientistas políticos e sociais, sejam adequadamente aplicados aos discursos parlamentares e às Casas Legislativas.

Palavra-chave: análise automatizada de conteúdo; texto como dado; análise do discurso parlamentar.

Abstract: Through substantially reducing time and cost of doing text-as-data research, automated content analysis methods have made feasible to research large collections, such as parliamentary speeches, quantitatively. This article identifies the conceptual framework of this new frontier of knowledge, describes a typical text-as-data project, presents the main methodologies and their applications to parliamentary speeches, shows the main challenges and pitfalls and, finally, suggests how to apply these methods to Legislative Houses' corpora. The intention is not to exhaust the range of methods, techniques and models; but to provide a guide for applying appropriately these methods to parliamentary speeches and other Legislative Houses' corpora.

Key-words: automated content analysis; text-as-data; analysis of parliamentary speeches.

Resumen: al reducir sustancialmente el tiempo y el costo de la búsqueda utilizando textos, los métodos automatizados de análisis de contenido han permitido investigar cuantitativamente grandes colecciones, como discursos parlamentarios. Este artículo identifica el marco de conceptos de esta nueva frontera del conocimiento, describe un proyecto de texto típico como se presenta, presenta las principales metodologías y sus aplicaciones a los discursos parlamentarios, muestra los principales desafíos y las mejores prácticas en esta área de investigación, y finalmente sugiere cómo aplicar estos métodos a las colecciones de la Casa Legislativa. La intención no es agotar la gama de métodos, técnicas y modelos; pero brinde una guía para que estos métodos, cada vez más populares entre los científicos políticos y sociales, se apliquen adecuadamente a los discursos parlamentarios y las cámaras legislativas.

Palabra clave: análisis de contenido automatizado; texto como se da; Análisis del discurso parlamentario.

[*]Mestre em Ciência Política pela San Diego State University, Estados Unidos. E-mail: ricardomodestovieira@gmail.com

1 INTRODUÇÃO

A revolução do "*big data*" e da inteligência artificial forneceram grandes oportunidades para Casas Legislativas pouparem trabalho, como por meio da indexação automática; gerarem informações estratégicas, como o posicionamento ideológico dos partidos; e proverem dados para importantes descobertas, como as dos pesquisadores Moreira (2016), Izumi (2017) e Schwartz (2018), por meio da análise automatizada de conteúdo aplicada ao discurso parlamentar[1]. Esse artigo pretende demonstrar como os discursos podem ser analisados quantitativamente com o auxílio de ferramentas computacionais, a fim de facilitar a organização e a compreensão desse vasto acervo, bem como gerar indicadores e informações estratégicas para as Casas Legislativas, suas Mesas Diretoras e suas Lideranças.

Os parlamentares expressam suas opiniões, defendem suas posições e apresentam proposições por meio de palavras. Segundo a abordagem empírica adotada, o que interessa é a probabilidade de palavras e expressões 'comuns' e 'incomuns' aparecerem como eventos linguísticos (MANNING; SCHUTZE, 1999, p. 7). Assim, por meio da frequência com que essas palavras e expressões se manifestam, é possível descrever o uso da linguagem nas Casas Legislativas. Evidentemente, tal abordagem estatística estará sempre fadada a diversos erros e nunca substituirá o componente humano; entretanto, com o avanço das ferramentas computacionais, métodos automatizados de análise de conteúdo transformaram-se em ferramentas úteis para explorar um vasto acervo como os discursos parlamentares.

Nesse sentido, serão analisados dez métodos diferentes – quatro métodos de classificação, incluindo dois *topic models* (modelos de tópicos), três métodos de escalonamento e três métodos de similaridade de textos – e suas aplicações para Casas Legislativas. Todos esses métodos utilizam o *bag-of-words* (saco-de-palavras) como pressuposto: o que interessa é a frequência com que certas palavras são encontradas no corpus ou acervo. Como não há um método global para a análise automatizada de conteúdo (GRIMMER; STEWART, 2013), a escolha se deu pelo seu uso na literatura de Ciência Política e pela sua oportunidade de aplicação para as Casas Legislativas tendo em vista a variabilidade de possíveis propósitos.

A fim de compreender melhor essas dez diferentes técnicas e suas possíveis aplicações, é necessário primeiro explicar os fundamentos do processamento de linguagem natural, bem como as técnicas de pré-processamento de dados que possibilitam a análise automatizada de conteúdo.

[1] Não se trata aqui das linhas francesas de Análise de Discurso – que tem como referências Michel Pêcheux e Dominique Maingueneau – ou de Análise de Conteúdo – que tem como referência Laurence Bardin, mas da linha estadunidense de Processamento de Linguagem Natural, cujas referências são George Kingsley Zipf e Alan Turing. Assim, a visão empiricista desses dois autores se diferenciava da linguística racionalista de Chomsky. Nesse artigo, o enfoque não é o sentido do discurso nem seu conteúdo, mas a utilização de técnicas quantitativas automatizadas aplicadas a um conteúdo em específico: os discursos e debates parlamentares. O termo "discursos e debates" foi oficialmente definido pela Câmara dos Deputados como o apropriado para se referir tanto a fala do orador sozinho na tribuna – como no Pequeno e no Grande Expediente – quanto a discussão das matérias na Ordem do Dia.

Após a explicação do pré-processamento, bem como da descrição das técnicas e de suas aplicações, serão analisados os desafios mais comuns relacionados a esses métodos de análise automatizada aplicados ao discurso parlamentar, bem como as melhores práticas preconizadas na literatura para enfrentar esses desafios. Finalmente, serão sugeridas algumas possíveis aplicações desses métodos para as Casas Legislativas.

2 FUNDAMENTOS E PRÉ-PROCESSAMENTO DE DADOS

Em uma abordagem empírica, o cérebro humano parte de associações, reconhecimento de padrões e generalizações para aprender a estrutura detalhada da linguagem natural. Não há, assim, uma faculdade inata da linguagem, como uma herança genética. Não há diferença entre a competência linguística — o conhecimento da estrutura da linguagem — e o desempenho linguístico de uma pessoa no mundo. Como a linguagem é inseparável de seu contexto social, o que importa são os padrões comuns que ocorrem no uso da linguagem (MANNING; SCHUTZE, 1999, p. 5-6).

Nesse sentido, palavras e frases são vistas como "usuais" e "incomuns", e a frequência dos tipos "usuais" demonstra as preferências que ocorrem no uso da linguagem (MANNING; SCHUTZE, 1999, p. 9). Por meio de uma abordagem estatística, é possível aprender automaticamente essas preferências lexicais e as palavras e expressões que tendem a se agrupar e a formar um campo lexical próprio. Como o significado de palavras/expressões está relacionado ao contexto em que elas são usadas, esse conhecimento sobre as preferências lexicais pode ser explorado para compreender relações semânticas mais profundas (MANNING; SCHUTZE, 1999, p. 18-19).

Seguindo essa abordagem estatística, Zipf (1949) descobriu que há uma relação entre a frequência de uma palavra e sua posição (*rank*) em uma lista de palavras de um mesmo corpus[2]. Essa constante é útil como uma descrição aproximada da distribuição de frequência de palavras em idiomas humanos: há algumas palavras muito comuns, um número intermediário de palavras de média frequência e muitas palavras de baixa frequência[3]. Zipf também obteve evidências empíricas para a tendência de as palavras pertencentes a um mesmo conteúdo agruparem-se (MANNING; SCHUTZE, 1999, p. 24-25).

Assim, a fim de reduzir a complexidade e o tamanho do vocabulário, bem como focar no que é usual e significativo no texto, analisam-se apenas as palavras de média frequência. Por isso é preciso retirar as palavras de conteúdo desnecessário e pouco frequentes — as que aparecem em

[2] Um conjunto de textos é chamado de corpus. Várias dessas coleções de textos são chamadas corpora (MANNING; SCHUTZE, 1999, p. 6).

[3] Como a linha é muito baixa para a maioria das classificações baixas e muito alta para classificações superiores a 10.000, Mandelbrot obteve uma relação mais precisa entre classificação e frequência, utilizando três outros parâmetros de texto como variáveis (MANNING; SCHUTZE, 1999, p. 25).

99% ou em menos de 1% dos documentos (GRIMMER; STEWART, 2013; HOPKINS; KING, 2010, *apud* IZUMI; MOREIRA, 2018; QUINN *et al.*, 2010, *apud* CASAS). As palavras mais frequentes normalmente não geram conteúdo significativo e correspondem a conjunções, preposições, artigos, pronomes e verbos como os de ligação.

Quando a ordem em que as palavras ocorrem no texto não é levada em consideração, assume-se o princípio do saco de palavras (*bag of words*) (BLEI, 2012; GRIMMER; STEWART, 2013). O princípio do saco de palavras apresenta cada documento como um único vetor com o comprimento igual ao número de palavras únicas no texto (ROBERTS *et al*., 2015; IZUMI; MOREIRA, 2018). A fim de reestabelecer alguma importância para a ordem das palavras, pode-se utilizar bigramas, trigramas (GRIMMER; STEWART, 2013) ou *collocations*[4].

É também importante frisar a necessidade de retirar o que na literatura é chamado de *stopwords* (GRIMMER; STEWART, 2013; ROBERTS *et al*., 2015; CASAS; WILKERSON, 2017). Cada contexto gera uma lista diferente de *stopwords*, ou seja, palavras, frases e expressões que são utilizadas muito comumente naquele específico contexto, mas que não geram conteúdo significativo. Lauderdale e Herzog (2016), por exemplo, retiraram todas as falas procedurais do processo legislativo, como as falas do Presidente, a leitura da ata e da agenda, eleições da Mesa, orações e homenagens.

Após a remoção das *stopwords*, é necessário reduzir a variabilidade das palavras por meio de *stemming* ou *lemmatization*. *Stemming* é a redução da palavra a seu radical por meio da remoção de seu final, como em plurais ou conjugações verbais. Na verdade, *stemming* é uma aproximação de um conceito linguístico chamado de *lemmatization*, o qual busca reduzir as palavras à sua forma básica e agrupá-las, utilizando para isso um algoritmo mais complexo que identifica a origem da palavra e retorna apenas sua *lemma* ou raiz (GRIMMER; STEWART, 2013; ROBERTS *et al*., 2015; CASAS; WILKERSON, 2017). No português, é utilizada uma adaptação do algoritmo de Porter (1980) para *stemming* (IZUMI; MOREIRA, 2018).

Após o *stemming*, as ocorrências individuais de cada palavra são chamadas de *tokens* (MANNING; SCHUTZE, 1999, p. 22) e o conteúdo do documento está finalmente pronto para ser convertido em dados quantitativos. Cabe lembrar que antes do pré-processamento é necessário

[4] Uma colocação é uma expressão de uso frequente em que o todo é maior que a soma das partes. Qualquer expressão que as pessoas repetem é candidata a uma colocação. Os dois padrões mais frequentes para colocações de palavras são "substantivo adjetivo" e "substantivo substantivo" (MANNING; SCHUTZE, 1999, p. 29-31).

obter[5], codificar[6] e tratar[7] os documentos. O saco de palavras é o método mais simples e mais usado para transformar cada documento em um vetor único cujo valor é determinado pela ausência ou presença de um *token* no documento, pela frequência desses *tokens* ou pela frequência normalizada pelo tamanho do documento (DIERMEIER *et al*., 2012). O objetivo é criar uma matriz de documentos e termos (*Document Term Matrix* – DTM) na qual cada linha representa um documento e cada coluna representa um *token* único. Como cada célula da matriz denota o número de vezes que o *token* indicado na coluna aparece no documento indicado na linha, cada documento é representado por um vetor único (ROBERTS *et al*., 2015; CASAS; WILKERSON, 2017; IZUMI; MOREIRA, 2018).

3 MÉTODOS E APLICAÇÕES

Após a obtenção, a codificação, o tratamento e o pré-processamento dos discursos parlamentares, é possível utilizar diversos métodos de análise automatizada de conteúdo para diferentes propósitos — como poupar trabalho, gerar informações estratégicas e realizar descobertas científicas. Esses métodos podem ser divididos em três conjuntos, de acordo com os objetivos e tarefas que se propõem realizar: (1) classificação; (2) escalonamento e (3) similaridade de textos. O primeiro se propõe a classificar os textos ou documentos em categorias conhecidas ou não conhecidas; o segundo se propõe a estimar a localização de atores utilizando uma escala; e o terceiro se propõe a mensurar a similaridade e/ou homogeneidade de textos ou de partes desses textos.

Há também duas diferentes abordagens: métodos supervisionados e não supervisionados. A principal diferença entre essas abordagens é que, na supervisionada, é necessário especificar a estrutura conceitual dos textos de antemão, enquanto que na não supervisionada utiliza-se um modelo para encontrar um resumo de baixa dimensionalidade que melhor explique os documentos observados, sendo necessário informar de antemão apenas o número de categorias (ROBERTS *et al*., 2015). O significado dessa dimensão que o método recupera, no caso dos discursos parlamentares, depende do contexto político, pois as preferências e as motivações para discursar variam de acordo com esse contexto.

[5] A obtenção dos documentos pode ocorrer de diversas formas: (1) dados abertos em formato JSON (JavaScript Object Notation), XML (Extensible Markup Language), CVS (Comma-Separated Values), etc.; (2) documentos de várias espécies convertidos em dados editáveis por meio de reconhecimento óptico de caracteres (OCR); (3) métodos de raspagem de dados (*web scraping data*) nos quais o computador acessa páginas na Internet, copiando e organizando seu conteúdo; (4) plataformas on-line como a Mechanical Turk, da Amazon, e aplicativos (APIs) que permitem a solicitação de conteúdo selecionado de um banco de dados (GRIMMER; STEWART, 2013; CASAS; WILKERSON, 2017; IZUMI; MOREIRA, 2018).

[6] A codificação (*encoding*) é a maneira pela qual o computador traduz caracteres individuais e únicos em *bytes* para que o texto possa ser lido pela máquina (ROBERTS *et al*., 2015; IZUMI; MOREIRA, 2018). Para caracteres em latim, é utilizado o UTF-8 (GRIMMER; STEWART, 2013).

[7] É necessário obter documentos formatados de forma semelhante. A formatação idêntica torna possível escrever um único script para extrair conteúdo mais específico de vários documentos de uma só vez (CASAS; WILKERSON, 2017).

Entretanto, os resultados da baixa dimensionalidade dos indicadores elaborados por meio da análise automatizada de conteúdo foram reconhecidos como uma característica importante da tomada de decisões no Congresso, principalmente quando se leva em consideração que o uso das palavras no debate político também varia de acordo com o tópico debatido[8] e o tipo de escala política ou ideológica estimada (LAUDERDALE; HERZOG, 2016). Assim, a baixa dimensionalidade no Congresso implica, por exemplo, que o registro de voto de um representante em um assunto será um bom preditor da sua escolha de voto em relação a outro assunto não relacionado (DIERMEIER *et al*., 2012).

3.1 Métodos de Classificação

Os métodos de classificação podem ser supervisionados e não supervisionados. Os métodos de aprendizagem supervisionada usam a frequência em que as palavras aparecem em um texto para classificar os documentos em categorias predeterminadas ou para medir a extensão em que os documentos pertencem a categorias específicas. O algoritmo então "aprende" como classificar os documentos nessas categorias usando um conjunto de treinamento. Ou seja, o algoritmo usa características dos documentos para classificá-los nas categorias. Como existem estatísticas claras que resumem o desempenho do modelo, métodos supervisionados são mais fáceis de validar (GRIMMER; STEWART, 2013).

Dentro dos métodos supervisionados, aqueles que exigem previamente a identificação das palavras que separam as classes são chamados de métodos de dicionário. Os métodos de dicionário usam a frequência relativa de palavras-chave para medir a presença de cada categoria em textos. Utilizando uma lista de palavras associadas à tonalidade do discurso (dicionário anotado), bem como a frequência relativa em que essas palavras-chave ocorrem, é possível medir o tom de um documento (GRIMMER; STEWART, 2013). Entretanto, a análise de sentimento — outra área importante da pesquisa de classificação, em que o objetivo é classificar o texto ordinalmente (de negativo para positivo, por exemplo) em vez de categoricamente — pode ser empregada utilizando métodos supervisionados e não supervisionados (CASAS; WILKERSON, 2017).

Modelos não supervisionados de associação mista ou modelos de tópicos são um conjunto de modelos generativos bayesianos que codificam a estrutura específica do problema em uma estimativa de categorias (GRIMMER; STEWART, 2013). Em outras palavras, são um conjunto de algoritmos que visam descobrir e identificar documentos com informações temáticas (BLEI, 2012). Os algoritmos analisam as palavras dos textos para descobrir os tópicos, como esses tópicos estão conectados entre si e como eles mudam com o tempo. Os algoritmos não têm

[8] A associação política para determinadas palavras depende do debate em que essas palavras foram usadas — uma palavra que implica uma posição de esquerda em um debate pode implicar uma posição de direita em outro debate —, mas também algumas palavras são usadas de forma semelhante para denotar a posição em muitos debates.

informações sobre esses tópicos, e os documentos não são rotulados com tópicos ou palavras-chave (BLEI, 2012). Como estatisticamente um tópico é uma função probabilística sobre as palavras, para estimar um tópico, os modelos usam a concorrência de palavras entre documentos (GRIMMER; STEWART, 2013).

Dessa forma, esses métodos não supervisionados utilizam características do texto sem impor categorias predeterminadas, utilizando apenas premissas de modelagem e propriedades dos textos para estimar um conjunto de categorias e simultaneamente atribuir documentos (ou partes de documentos) a essas categorias (GRIMMER; STEWART, 2013). As distribuições de tópicos interpretáveis surgem computando a estrutura oculta que provavelmente gerou a coleção de documentos observada. Assim, o processo que gera os tópicos define uma probabilidade de distribuição em relação a variáveis randômicas observáveis e não observáveis (BLEI, 2012).

Os métodos de classificação e modelos de tópicos podem ser utilizados para diversos propósitos. Esses métodos podem ser utilizados no estudo de questões, tópicos e agendas legislativas, como, por exemplo, entender como as agendas políticas mudam ao longo do tempo (DIERMEIER *et al*., 2012). Métodos supervisionados podem servir para categorização e indexação automática de documentos, poupando trabalho, bem como para análise de sentimento e posicionamento de atores políticos. Métodos não supervisionados de classificação podem responder perguntas sobre o funcionamento interno do governo, a influência de diferentes grupos e atores políticos na formulação de políticas públicas, posicionamento político de atores, etc. (ROBERTS *et al*., 2015).

Como exemplos práticos recentes, citam-se Moreira (2016), que pesquisou a relação governo-oposição com base nos discursos parlamentares por meio do modelo de tópico Expressed Agenda Model e constatou que a ênfase temática dos deputados não segue a relação governo-oposição verificada no âmbito das votações nominais; Izumi (2017), que estimou posições políticas no Senado por meio do classificador Naive Bayes (análise de sentimento) aplicado aos discursos dos Senadores; e Diermeier *et al*. (2012), que utilizou o Support Vector Machines (SVM) em conjunto com a base de dados de votações nominais para prever a posição ideológica dos parlamentares, bem como medir o grau de coesão intrapartidária no Senado. Comparado com o Naive Bayes, os resultados de Diermeier demonstraram que o SVM é superior para classificar posicionamento ideológico.

3.1.1 Naive Bayes

O Naive Bayes é um dos métodos supervisionados de classificação mais utilizados. Embora parta de um pressuposto ingênuo — o modelo assume que as palavras são geradas de forma independente para uma dada categoria (*the naive assumption*), quando na verdade o uso de palavras é altamente correlacionado em qualquer conjunto de dados —, o modelo fornece um método alternativo útil para atribuir documentos a categorias predeterminadas (GRIMMER;

STEWART, 2013; IZUMI; MOREIRA, 2018). No caso de grandes acervos, o classificador Decision Tree pode ser utilizado em conjunto com o Naive Bayes para aumentar a precisão (KOHAVI, 1996).

Para que o método supervisionado funcione, é preciso primeiro (1) construir um conjunto de textos de treinamento (*training set*) — no qual haja (1.1) a criação de um esquema de codificação feito manualmente e (1.2) a escolha aleatória de documentos de amostragem (como regra geral entre 100 e 500 documentos); depois (2) aplicar o método de aprendizado supervisionado para que o algoritmo "aprenda" — aprendendo a relação entre as características e as categorias no conjunto de treinamento e, em seguida, usando isso para inferir os rótulos no conjunto de testes — como classificar os documentos nas categorias utilizando o conjunto de treinamento; e (3) validar a saída do modelo, comparando a saída da codificação automatizada com a saída da codificação manual. Os métodos de aprendizado supervisionado são muito mais fáceis de validar, com estatísticas claras que resumem o desempenho do modelo. Após esses três procedimentos é possível classificar com sucesso os demais documentos (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018).

3.1.2 Support Vector Machines (SVM)

O Support Vector Machines (SVM) é um método supervisionado que utiliza um algoritmo de classificação de texto para extrair os termos mais indicativos das posições conservadoras e liberais dos discursos legislativos e para prever as posições ideológicas dos parlamentares. O SVM é baseado no princípio da minimização do risco estrutural da teoria de aprendizagem estatística. Primeiro, os discursos feitos pelos 25 senadores mais liberais e mais conservadores dos Estados Unidos foram analisados — utilizando as pontuações do DW-Nominate (voteview.com), que mede o posicionamento ideológico com base nas votações nominais — para treinar o algoritmo de classificação (DIERMEIER *et al*., 2012).

Na fase de treinamento, rotulou-se arbitrariamente uma categoria como "negativa" e a outra como "positiva". Como os pontos de dados em cada categoria são separáveis por um hiperplano, há dois hiperplanos paralelos em que a distância entre os pontos de cada categoria é a maior possível. Esses pontos de dados são chamados vetores de suporte, e a distância entre os dois hiperplanos paralelos é chamada de margem. A tarefa do SVM na fase de treinamento é encontrar os dois hiperplanos de separação de forma que a margem seja máxima. Em seguida, a mesma metodologia foi utilizada para investigar os senadores moderados, treinando o classificador em senadores moderados (DIERMEIER *et al*., 2012).

3.1.3 Latent Dirichlet Allocation (LDA)

O Latent Dirichlet Allocation (LDA) é o modelo mais simples de tópico ou de associação mista de dados agrupados. A ideia por trás do LDA é que cada documento é uma mistura de vários

tópicos em diferentes proporções. Além da suposição do saco de palavras, há a suposição de que o número de tópicos é conhecido e fixo. Tecnicamente, o modelo assume que os tópicos são gerados primeiro, antes dos documentos (BLEI, 2012; GRIMMER; STEWART, 2013).

Para cada documento do corpus, as palavras são geradas em duas etapas: (1) escolhe-se aleatoriamente uma distribuição sobre os tópicos; (2) para cada palavra no documento (2.1) escolhe-se aleatoriamente um tópico da distribuição da primeira etapa e (2.2) escolhe-se aleatoriamente uma palavra da distribuição correspondente sobre o vocabulário. A distribuição que é usada para desenhar a distribuição do tópico do pré-documento é chamada de distribuição Dirichlet, e o resultado do Dirichlet é usado para alocar as palavras dos documentos para diferentes tópicos. Todos os documentos compartilham o mesmo conjunto de tópicos, mas cada documento exibe esses tópicos em uma proporção diferente (BLEI, 2012).

3.1.4 Structural Topic Model (STM)

O STM também é um modelo de tópico de associação mista, mas fornece uma maneira flexível de incorporar metadados associados ao texto — como quando o texto foi escrito, onde foi escrito, quem o escreveu, as características do autor, etc. — como covariáveis na análise de documentos (ROBERTS *et al*., 2015). A inclusão de metadados de documentos segue e amplia o Dynamic Topic Model — modelo no qual a probabilidade de observar um tópico se modifica ao longo do tempo — e o Expressed Agenda Model — modelo que inclui informações sobre os autores dos documentos, supondo que cada autor divide sua atenção a um conjunto de tópicos. As covariáveis também permitem o compartilhamento de metadados em função da frequência de tópicos, como a probabilidade de mulheres falarem um tópico em específico ou de usarem certas palavras em relação aos homens. O STM também se distingue do LDA por substituir a distribuição de Dirichlet por uma distribuição logística normal, como no Modelo de Tópicos Correlacionados, para estimar a correlação entre os tópicos. Com isso, é possível desenhar uma rede de tópicos correlacionados para um modelo de tópico estruturado, utilizando um tema como preditor de tópicos em um determinado corpus (GRIMMER; STEWART, 2013; ROBERTS *et al*., 2015).

3.2 Métodos de Escalonamento

Os métodos de escalonamento são utilizados para localizar políticos e partidos em espaços ideológicos contínuos (KLUVER, 2009; DIERMEIER *et al*., 2012; GRIMMER; STEWART, 2013; LAUDERDALE; HERZOG, 2016; CASAS; WILKERSON, 2017; IZUMI; MOREIRA, 2018). Esses métodos baseiam-se na suposição de que as tendências ideológicas dos atores políticos determinam o que é discutido nos textos — suposição da dominância ideológica na fala (GRIMMER; STEWART, 2013). Mas isso pode não ser verdade, pois, como Mayhew (1974) demonstrou, os políticos se envolvem regularmente em reivindicações de crédito não ideológicas.

Assim, os métodos de escalonamento terão melhor desempenho se forem acompanhados de métodos que separem enunciados ideológicos e não ideológicos (CASAS; WILKERSON, 2017).

A utilização mais conhecida desses métodos é o projeto Vote View (voteview.com). Como a ideologia não é diretamente observável, Poole e Rosenthal (1991) desenvolveram um modelo espacial bidimensional (*D-Nominate scores*) para classificar parlamentares por meio de seus votos nominais. A primeira dimensão representa a tradicional visão de esquerda-direita associada ao papel do governo na economia e na redistribuição de renda. A segunda dimensão representa questões de interferência do Estado na vida privada, escravidão e, posteriormente, questões raciais e de direitos civis. O modelo consegue classificar corretamente 85% das decisões individuais de voto de cada membro do Congresso (DIERMEIER *et al*., 2012).

Outra aplicação conhecida é o Manifesto Project, que utiliza decodificação manual para classificar mais de 1.000 partidos em 50 países de acordo com seus manifestos políticos (manifesto-project.wzb.eu). Nesse caso, a estimativa pontual ideal é usada para classificar as partes como pontos de referência em uma escala ideológica esquerda-direita (DIERMEIER *et al*., 2012).

Além disso, Lauderdale e Herzog (2016) estimaram as posições políticas individuais de cada parlamentar estadunidense utilizando o método de escalonamento Wordshoal, que combina o Wordfish e a análise fatorial Bayesiana. Os resultados dessa pesquisa também sugerem que as posições políticas individuais de parlamentares são preditoras precisas da coesão intrapartidária e do comportamento dissidente em países onde os sistemas eleitorais fornecem fortes incentivos aos votos pessoais (LAUDERDALE; HERZOG, 2016). Finalmente, Pritoni (2014) analisou as dificuldades em mensurar a influência de grupos de interesse extraindo o posicionamento político desses grupos e o comparando com produtos legislativos por meio de métodos de escalonamento.

3.2.1 Wordscore

O Wordscore é um algoritmo supervisionado para dimensionamento e um caso especial de método de dicionário. O primeiro passo é a seleção de textos de referência que definem as posições políticas no espaço, como liberal e conservadora. Os textos de referência (treinamento) são usados para gerar uma pontuação para cada palavra. A pontuação mede a taxa relativa com que cada palavra é usada nos textos de referência. Isso cria uma medida de quão bem a palavra separa parlamentares liberais e conservadores. Depois a pontuação das palavras é usada para dimensionar os textos restantes (GRIMMER; STEWART, 2013). Ou seja, o segundo passo é gerar scores para as palavras dos textos de referência baseados na posição política atribuída a priori e ponderada pela probabilidade de observá-la em um documento. O mesmo procedimento pode ser utilizado para a escala ideológica esquerda-direita (IZUMI; MOREIRA, 2018).

O Wordscore é baseado em uma série de suposições: (1) as posições políticas são refletidas na frequência relativa das palavras usadas dentro e entre os textos; (2) o significado das palavras

permanece estável ao longo do tempo; (3) todas as palavras têm o mesmo peso no processo de estimativa; e (4) todas as palavras de interesse estão contidas nos textos de referência (KLUVER, 2013). Ainda é necessário lidar com a definição da dimensão política a ser investigada e escolher um conjunto de textos de referência com estimativas de posição política conhecidas, de preferência de uma fonte independente (KLUVER, 2013). Também é importante que os textos de referência utilizem o mesmo léxico que os textos a serem testados, que cubram todo o espectro ideológico e que possuam um conjunto diversificado de palavras (IZUMI; MOREIRA, 2018).

3.2.2 Wordfish

O Wordfish é um algoritmo não supervisionado que estima a importância das palavras para discriminar as posições políticas baseado na teoria de resposta a itens (IRT). O modelo assume uma distribuição de Poisson para a contagem de palavras e parte do pressuposto de que a probabilidade de observarmos uma palavra em um documento é independente da posição das outras palavras no mesmo documento. Assim, ele pode ser usado para descobrir palavras que distinguem posições em um espectro político (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018). O Wordfish não requer textos de referência em contraste com o Wordscore. Nesse caso, o pesquisador deve definir a dimensão política a ser analisada, selecionar os documentos que tratam dessa dimensão política e remover todas as passagens de texto que não se referem à dimensão investigada (PRITONI, 2014). Portanto, uma validação cuidadosa é necessária para confirmar que o espaço ideológico pretendido foi identificado (GRIMMER; STEWART, 2013).

3.2.3 Wordshoal

O Wordshoal é um modelo de fator hierárquico para uso da palavra em debates legislativos que combina duas abordagens em uma única estratégia de estimativa. A primeira é limitar a análise a discursos sobre um único tópico legislativo, mantendo constante a variação tópica. A segunda abordagem é combinar muitos discursos sobre muitos tópicos legislativos em um único documento para cada legislador ou partido. Ou seja, no primeiro estágio o modelo usa a escala de texto existente Wordfish para mensurar a variação do uso de palavras em cada tópico separadamente. No segundo estágio, utiliza-se a análise fatorial bayesiana para construir uma escala comum a partir das posições específicas do debate estimadas no primeiro estágio. Assim, o modelo apresenta os resultados das estimativas com base no Wordfish para cada debate e, em seguida, usa essas estimativas como dados para o modelo de agregação do segundo estágio para avaliar se um parlamentar está geralmente à direita ou à esquerda de outro parlamentar em um conjunto de debates sobre temas heterogêneos (LAUDERDALE; HERZOG, 2016).

Mais especificamente, o modelo utiliza uma escala unidimensional Wordfish aplicada a um conjunto de textos dentro de um único debate político, mantendo assim uma variação constante

de uso de palavras orientada por tópicos. Esse modelo de escala de Poisson aplicado a cada debate resulta em uma estimativa específica do debate da posição relativa de cada parlamentar. Tendo estimado a posição expressa para todos os parlamentares em um determinado tópico, o modelo agrega dimensões específicas do debate que envolvam subconjuntos variáveis de legisladores em um número menor de dimensões que incluam todos os legisladores. Como essa abordagem não depende da variação do uso da palavra em qualquer debate para estimar posições em uma dimensão latente de discordância, é possível gerar uma (ou mais) posição geral latente para cada legislador com grande poder preditivo (LAUDERDALE; HERZOG, 2016).

A suposição empírica do Wordshoal é que a discordância política é mais clara e consistentemente refletida na variação dentro do debate no uso da palavra do que na variação do uso da palavra em vários debates. Uma de suas principais inovações é que o Wordshoal permite que o significado e o poder discriminatório de uma determinada palavra variem de debate para debate. A variação no uso de palavras entre discursos é tanto uma função do tópico de um debate quanto uma função da posição que um legislador toma. Além disso, o método fornece estimativas de incerteza significativas das posições agregadas dos legisladores, levando em conta a frequência com que os parlamentares falaram e a consistência com que expressaram suas posições nos debates (LAUDERDALE; HERZOG, 2016).

3.3 Métodos de Similaridade de Texto

A reutilização de texto consiste em descobrir instâncias de similaridade no uso da linguagem. A característica distintiva dos algoritmos de reutilização de texto é que eles valorizam explicitamente o sequenciamento de palavras ao julgar a similaridade do documento (CASAS; WILKERSON, 2017). Entretanto, como a semelhança de corpus é inerentemente multidimensional (eles serão semelhantes de alguma forma e diferentes em outras), uma medida de similaridade só faz sentido quando comparados dois corpus homogêneos. Sendo assim, a similaridade só pode ser interpretada à luz da homogeneidade do corpus. Nesse sentido, a mesma medida pode ser usada para similaridade e homogeneidade, comparando a distância entre dois corpus (*within-corpus distance*) (KILGARRIFF; ROSE, 1998).

Os métodos de similaridade podem ser utilizados para comparar documentos como um todo ou encontrar pequenos pedaços de texto correspondentes entre dois documentos, como aqueles encontrados em legislação elaborada a partir de múltiplas fontes. Para estudar se projetos de lei do Congresso reutilizam textos de outros projetos do Congresso, Wilkerson *et al*. (2015) usaram o algoritmo de Smith-Waterman para comparar cadeias de texto de projetos de lei do Congresso Estadunidense introduzidos desde 1990. Além de revelar padrões sobre os parlamentares que introduzem projetos similares em uma legislatura ou entre legislaturas, o algoritmo pode ser usado para determinar em que medida a linguagem que um legislador introduz combina com a de outros legisladores (BURGESS *et al*., 2016).

Já Hertel-Fernandez e Kashin (2015) utilizaram métodos de similaridade para rastrear as origens de proposições e desvendar a influência de grupos de interesse no processo legislativo (CASAS; WILKERSON, 2017). Nesse mesmo sentido, Burgess *et al*. (2016) desvendaram a reutilização de textos em proposições, utilizando o ElasticSearch para limitar o número de comparações, a fim de detectar a influência de grupos de interesse na proliferação de legislações estaduais nos Estados Unidos (Legislative Influence Detector Project: dssg.uchicago.edu/lid/).

Finalmente, os métodos de similaridade também podem ser utilizados para posicionar parlamentares e partidos em um espectro ideológico. Nesse sentido, Schwartz (2018) utilizou a técnica do X^2 para testar a similaridade de discursos de Grande e Pequeno Expediente — por meio da frequência de palavras e bigramas (*collocations*) nos conjuntos de discursos dos partidos políticos (PT, PSDB, PMDB, PSOL, PCdoB e PTB) comparados dois a dois.

3.3.1 Smith-Waterman

O algoritmo Smith-Waterman foi desenvolvido para encontrar subsequências similares dentro de longas cadeias de DNA, mas também pode ser utilizado para encontrar trechos de um documento que são semelhantes a trechos de outros documentos por meio de uma pontuação de alinhamento baseada em três parâmetros: correspondência de palavras (*matching*), incompatibilidade de palavras (*mismatching*) e lacunas (*gap*). Esse algoritmo é uma boa escolha para comparar um número relativamente pequeno de documentos, mas pode levar muito tempo para ser executado em um grande corpus (BURGESS *et al*, 2016).

Para resolver esse problema, Burgess *et al* (2016) utilizaram o mecanismo de pesquisa ElasticSearch, configurado com a função de pontuação Lucene padrão, para classificar os documentos para uma determinada consulta e, assim, filtrar o conjunto de documentos executados, identificando um subconjunto de documentos no corpus com maior probabilidade de conter texto semelhante ao do documento de consulta. A filtragem mostrou aumentar a eficiência, pois o algoritmo de alinhamento local compara apenas os documentos retornados pelo módulo de pesquisa, sem sacrificar a precisão nas tarefas de similaridade de documentos (BURGESS *et al*, 2016).

3.3.2 Cosseno

Como já foi demonstrado, cada documento pode ser representado por um vetor, com o comprimento igual ao número de palavras únicas no texto. Assim, assume-se que quanto maior a similaridade na frequência relativa das palavras utilizadas, maior será a similaridade do conteúdo entre os textos. Sendo dois vetores "u" e "v", é possível calcular a similaridade por meio do produto interno entre eles, pois quanto maior o produto interno entre eles, maior a frequência para as mesmas palavras. Como essa medida ainda é problemática, a solução é dividir o produto interno pelo produto dos comprimentos dos vetores, o que é representado matematicamente pelo cosseno

do ângulo formado entre os vetores "u" e "v" (IZUMI; MOREIRA, 2018).

3.3.3 X^2

Kilgarriff e Rose (1998) apresentaram um método para avaliar a similaridade de corpus – denominado Similaridade Conhecida de Corpora – e testaram abordagens comumente discutidas na literatura: medidas de entropia cruzada[9], *Spearmen* e X^2. Para o tamanho do corpus utilizado – um subconjunto do *British National Corpus* contendo 300 mil palavras de jornais e periódicos divididos em pares de 10 mil palavras – as abordagens X^2 e *Spearmen* tiveram melhor desempenho[10] que qualquer uma das medidas de entropia cruzada; entre as duas, X^2 superou *Spearmen*. Para cada uma das palavras mais comuns, os autores calcularam o número de ocorrências esperadas em cada corpus, se ambos os corpus fossem amostras aleatórias do mesmo corpora. Como um corpus jamais é uma amostra aleatória de palavras, a diferença na frequência de cada palavra entre dois corpora tende a aumentar, mas não aumenta na ordem de grandeza, como acontece com frequências brutas (KILGARRIFF; ROSE, 1998).

Quadro 1- Resumo dos Métodos de Análise Automatizada de Conteúdo

Método	Técnica	Descrição e Aplicação
Classificação	Naive Bayes	Método supervisionado que classifica documentos em categorias conhecidas a partir de um conjunto de treinamento. O classificador Decision Tree é utilizado para aumentar a precisão em caso de grandes acervos. Pode ser utilizado para indexação automática, para a classificação de tópicos mais debatidos, para análise de sentimento e posicionamento ideológico de parlamentares.
	SVM	Método supervisionado, o Support Vector Machines (SVM) é utilizado em conjunto com a base de dados de votações nominais para prever a posição ideológica dos parlamentares, correlacionar o que eles falaram e como eles votaram, bem como medir a consistência ideológica dos partidos e o grau de coesão intrapartidária no Congresso.
	LDA	Método não supervisionado, o Lattent Dirichlet Allocation (LDA) classifica documentos sem a necessidade de especificar as categorias previamente, mas é necessário informar o número de categorias.
	STM	Método não supervisionado, o Structural Topic Model (STM) permite a inserção de metadados como covariáveis, combinando o Expressed Agenda Model e o Dynamic Topic Model. Pode ser usado para classificação de tópicos mais debatidos, bem como compreender as posições políticas, os padrões de liderança e as influências no processo legislativo.
Escalonamento	Wordscore	Método supervisionado que posiciona politicamente documentos em dimensões conhecidas a partir de um conjunto de treinamento. Pode ser usado para posicionar partidos políticos e parlamentares em uma escala ideológica – como esquerda-direita ou liberal-conservador –

[9] Perplexidade é, grosseiramente falando, uma medida do tamanho do conjunto de palavras a partir do qual a próxima palavra é escolhida, dado o histórico das palavras. Perplexidade é usada para avaliar quão boa é uma estratégia de modelagem de linguagem, considerando o mesmo corpus. Assim, a perplexidade pode ser usada para medir uma propriedade semelhante à homogeneidade se a estratégia de modelagem de linguagem for mantida constante e os corpora forem variados. Com a estratégia de modelagem de linguagem mantida constante, a entropia cruzada torna-se uma medida de similaridade (KILGARRIFF; ROSE, 1998).

[10] O teste de confiabilidade (chamado de gold standard) dos métodos foi realizado tendo com parâmetro a comparação de dois conjuntos de pares pelos codificadores (KILGARRIFF; ROSE, 1998).

		por meio de seus discursos.
	Wordfish	Método não supervisionado que posiciona politicamente documentos sem a necessidade de especificar as dimensões previamente, mas é necessário informar o número de dimensões. Pode ser usado para prever posições políticas e posicionar os partidos políticos e os parlamentares em uma escala ideológica.
	Wordshoal	O Wordshoal combina o Wordfish e a análise fatorial Bayesiana para estimar a posição de cada parlamentar em relação aos demais dentro do partido e entre partidos tendo como parâmetro os seus discursos. Também pode ser usado para estimar a coerência do posicionamento do parlamentar ao longo do tempo, bem como a coesão intrapartidária e do comportamento dissidente.
Similaridade de texto	Smith-Waterman	O Smith-Waterman é um algoritmo que mede o reuso de partes de um texto, sendo usado para medir a similaridade de trechos entre dois textos. Pode ser utilizado para identificar a origem de partes de proposições, relatórios, emendas e legislação aprovada. Também pode ser usado para mensurar a influência de grupos de interesse nos produtos legislativos.
	Cosseno	A técnica do cosseno pressupõe que quanto maior a similaridade na frequência relativa das palavras utilizadas, maior a similaridade entre dois textos como um todo. Pode ser utilizado para identificar quão similares são duas proposições ou emendas.
	X^2	A estatística X^2 é utilizada para mensurar a similaridade e/ou homogeneidade entre corpus. O X^2 calcula o número de ocorrências das palavras mais comuns em cada corpus e mede a diferença entre a frequência dessas palavras nos dois corpus. Pode ser utilizado para medir a identidade ideológica entre partidos e blocos partidários baseados nos discursos dos parlamentares.

Fonte: Elaboração própria, 2018.

4 DESAFIOS METODOLÓGICOS E BOAS PRÁTICAS

Como os métodos de análise automatizada de conteúdo não substituem a leitura humana, é necessária uma cuidadosa validação dos resultados (GRIMMER; STEWART, 2013), tendo como princípios a replicabilidade dos resultados e a rigorosa codificação manual, preferencialmente por mais de um codificador. Sendo assim, a validação é um componente crítico de todo projeto de texto-como-dado (CASAS; WILKERSON, 2017). Para métodos de classificação e escalonamento supervisionados, é importante demonstrar que a classificação computadorizada replica a codificação manual. Um classificador pode ser refinado por meio de um livro de código e iterações de codificação manual (GRIMMER *et al*, 2018). Já nos métodos não supervisionados não há um *gold standard* desse tipo, a validação acontece à medida que se ajusta os parâmetros para examinar novos resultados (GRIMMER *et al*, 2018). Uma maneira de fazer isso é examinar o grau de coesão e a distinção das palavras de cada tópico (ROBERTS *et al*, 2014; CASAS; WILKERSON, 2017). Métodos não supervisionados também requerem a validação de que as medidas produzidas correspondam aos conceitos reivindicados (GRIMMER; STEWART, 2013).

Cabe lembrar que, além de multidimensionais, os textos são mais flexíveis do que outros tipos de variáveis, criando uma gama mais ampla de potenciais propriedades do texto a serem analisadas e validadas (GRIMMER *et al*, 2018). A questão da multidimensionalidade de textos

também é mencionada por Lauderdale e Herzog (2016), Roberts *et al* (2015), Diermeier *et al* (2012) e Kilgarriff e Rose (1998). Como todo texto é multidimensional, é necessário escolher umas ou poucas dimensões (*low-dimensional representation*) para compreender o corpus e fazer inferências. Essa limitação é inerente aos três conjuntos de métodos[11]. Evidentemente, o problema da multidimensionalidade no caso da posição ideológica também esbarra na dificuldade de operacionalização de complexos conceitos, como ideologia, esquerda-direita, liberal-conservador etc.

Portanto, é importante validar os resultados com eventos do mundo real e eventos esperados (GRIMMER; STEWART, 2013; CASAS; WILKERSON, 2017), conectando-os, por exemplo, com fatos, dados quantitativos e resultados da atividade legislativa (SCHWARTZ, 2018). Outra boa prática de validação é utilizar diferentes algoritmos para o mesmo propósito e comparar se os resultados são similares (QUINN *et al*, 2010, *apud* CASAS; WILKERSON, 2017, GRIMMER; KING, 2011, ROBERTS *et al*, 2014). Nesse sentido, métodos de aprendizado supervisionado podem ser usados para validar ou generalizar as descobertas fornecidas por métodos não supervisionados (GRIMMER; STEWART, 2013). Há também uma troca (*trade-off*) entre o grau de generalização do conceito e a validade dos indicadores empíricos (PRITONI, 2014). É tentador generalizar, mesmo quando a propriedade do texto a ser analisada é mais específica. Isso aumenta a relevância teórica, mas diminui a fidelidade do indicador (GRIMMER *et al*, 2018).

Um problema comum relacionado à validação é o chamado *overfitting* (excesso de ajuste). Ao utilizar os mesmos documentos para descobrir propriedades do texto, é comum interpretar de forma equivocada os resultados. Para resolver esse problema, uma solução é separar os conjuntos de treinamento e de teste (*split sample design*). Ao dividir as amostras e separar um conjunto de treinamento para uso na descoberta (fixando resultados potenciais) e um conjunto de teste na estimativa (análise), quebra-se a dependência entre a descoberta das propriedades do texto a serem analisadas e seu efeito causal (GRIMMER *et al*, 2018). Para Grimmer e Stewart (2013), o procedimento de validação ideal seria dividir os dados em três subconjuntos. O primeiro seria o subconjunto de teste, no qual o modelo seria ajustado. O segundo seria o subconjunto de validação, codificado à mão e usado para avaliar a performance do modelo. Finalmente, o modelo seria aplicado para classificar o terceiro subconjunto.

Casas e Wilkerson (2017) também consideram salutar treinar o algoritmo em um conjunto de textos já rotulados antes de testar sua precisão em um conjunto desconhecido. Entretanto, eles acreditam que repetir esse processo várias vezes, usando diferentes conjuntos para treinamento e para testes e, em seguida, agregar a validação dos resultados (validação cruzada em N-*fold*) é uma

[11] Na classificação é necessário definir os tópicos ou ao menos o número de tópicos. No escalonamento é necessário definir as dimensões a serem mensuradas. Na similaridade de corpus, dois textos serão semelhantes de alguma forma e diferentes em outras, pois a similaridade só pode ser interpretada à luz da homogeneidade do corpus, ou seja, não é adequado, por exemplo, comparar discursos parlamentares com manuais técnicos (KILGARRIFF; ROSE, 1998).

abordagem ainda melhor. Além do *split sample design* (desenho de amostras separadas), outra maneira de evitar o *overfitting* (excesso de ajuste) é utilizar algoritmos diferentes para demonstrar se existem *clusters* (agrupamentos) semelhantes (QUINN *et al*, 2010, *apud* CASAS; WILKERSON, 2017; GRIMMER; KING, 2011 e ROBERTS *et al*, 2014).

Além do *overfitting* (excesso de ajuste), há também o problema da instabilidade de tópicos nos métodos não supervisionados. Para evitar essa instabilidade é salutar utilizar resultados de diferentes modelos do mesmo algoritmo. Casas e Wilkerson (2017), por exemplo, utilizaram 17 modelos de alocação latente de Dirichlet (LDA) – variando o número de tópicos entre 10 e 90, em incrementos de cinco – para produzir 850 tópicos (10 + 15 + 20… + 90). A fim de determinar quais tópicos eram consistentes, eles primeiro calcularam a similaridade de cosseno para todos os pares de tópicos (resultando em 722.500 pontuações de similaridade) e, em seguida, utilizaram o algoritmo Spectral Clustering para agrupar os 850 tópicos com base na similaridade de cosseno e verificar quais tópicos se mantinham constantes (CASAS; WILKERSON, 2017).

Outras duas características típicas dos discursos parlamentares são, como demonstram Lauderdale e Herzog (2016), a esparsidade e a seleção de oradores. Levando em consideração que na vida real apenas poucos parlamentares discursam em determinados debates, como resultado, a matriz de documentos e termos – DTM – é esparsa. Outro ponto é o controle partidário e o controle de agenda. A fim de evitar esse enviesamento, Moreira (2016) utilizou apenas discursos do Pequeno Expediente, e Schwartz (2018) utilizou discursos do Pequeno e do Grande Expediente, pois nesses dois momentos os parlamentares são livres para se expressar sobre qualquer assunto. Lauderdale e Herzog (2016) também descobriram que em sistemas com fortes incentivos ao voto pessoal e não partidário, os parlamentares falam mais livremente porque os partidos reconhecem a necessidade de reconhecimento dos nomes dos parlamentares.

5 POSSÍVEIS APLICAÇÕES E RECOMENDAÇÕES

Há diversas possíveis aplicações de métodos automatizados de conteúdo para Casas Legislativas. Métodos supervisionados, por exemplo, podem ser mais indicados para poupar trabalho – como por meio da indexação ou classificação automáticas – e métodos não supervisionados para realizar descobertas e fornecer insights sobre essas classificações (GRIMMER; STEWART, 2013; CASAS; WILKERSON, 2017). Entretanto, eles também são métodos complementares, uma vez que a aprendizagem supervisionada pode ser utilizada para validar ou generalizar as descobertas fornecidas por métodos não supervisionados (GRIMMER; STEWART, 2013).

No caso dos métodos supervisionados, é evidente a possibilidade de poupar trabalho com a utilização da indexação automática, ou ao menos semiautomática – no caso de o servidor confirmar o termo sugerido pelo algoritmo. Dentre os diversos processos de indexação executados

pela Câmara dos Deputados, por exemplo, citam-se a indexação de discursos parlamentares (realizada pelo Departamento de Taquigrafia, Revisão e Redação - DETAQ), de proposições (realizada pelo Centro de Documentação e Informação - CEDI) e de matérias jornalísticas (realizada pela Secretaria de Comunicação Social - SECOM). Além da indexação, outros processos – como o apensamento e a distribuição de proposições às Comissões (realizados pela Secretaria Geral da Mesa – SGM) – também poderiam ganhar muito com o uso de métodos supervisionados. Além de poupar trabalho, a automatização também traria maior uniformização à indexação da Casa e, consequentemente, facilitaria a busca e recuperação de informações.

Assim, é recomendável utilizar um método supervisionado – como o Naive Bayes – para implementar a automatização dos processos de indexação; bem como refinar o classificador por meio de um livro código, como a base Tesauro da Câmara, e validar os resultados dos testes por meio de iterações de codificação manual. Ainda assim, é salutar manter o componente humano no processo de trabalho para revisar os resultados apresentados pela máquina. A fim de evitar o *overfitting*, é recomendável utilizar o *split sample design*[12].

Outra aplicação da análise automatizada para Casas Legislativas é o fornecimento de diversas informações estratégicas. Quão importante é saber quais foram os temas mais debatidos pela Casa durante a legislatura? Qual é o posicionamento dos deputados e dos partidos em cada um desses temas? Essa é uma legislatura majoritariamente de direita ou de esquerda? Liberal ou conservadora? Há uma identificação ideológica nos partidos? Os parlamentares votam segundo essa identificação? Há consistência no que falam e como votam os parlamentares? Há coesão partidária nos discursos parlamentares? Há alinhamento entre os discursos da base e da oposição? Quem influência mais o processo legislativo e os discursos dos parlamentares? O Poder Executivo? Os partidos? A base eleitoral? Os grupos de interesse?

A quantidade de informações que podem ser extraídas dos discursos parlamentares – principalmente se levada em conta a combinação desses discursos com os produtos e resultados da atividade legislativa – parece infindável. Para fins didáticos, é possível dividir essas informações em três conjuntos: (1) temas ou tópicos de debate, (2) posicionamento em uma escala política ou ideológica e (3) influência no processo legislativo. Como o uso de palavras no debate político pode variar de acordo com o tópico debatido, a classificação de tópicos serve também como dimensão para analisar o posicionamento político e a influência.

Assim, a classificação em tópicos emerge como o primeiro desafio de a Casa fornecer informações estratégicas de qualidade por meio da análise automatizada de conteúdo. Dentre as diversas abordagens possíveis, sugere-se a utilização de métodos supervisionados e não

[12] Também é possível utilizar o método não supervisionado STM para descobrir palavras e expressões novas que estão sendo utilizadas frequentemente pelos deputados, mas não estão na indexação. Esse pode ser o caso de neologismos e expressões como "ideologia de gênero", "escola sem partido", "pixuleco", etc.

supervisionados como complementares, utilizando um algoritmo para validar os resultados do outro. Nesse sentido, há dois caminhos possíveis: (1) utilizar o método não supervisionado STM para classificação e validar os resultados com o método supervisionado Naive Bayes/Decision Tree; ou (2) utilizar o Naive Bayes/Decision Tree para classificação e validar os resultados com o STM[13]. Nos dois caminhos é possível utilizar o resultado de binômios/trinômios para rejeitar a hipótese nula em relação à utilização dos dois algoritmos, bem como validar os resultados com fatos, eventos e resultados da atividade legislativa.

Além disso, é recomendável utilizar o *split sample design* para o Naive Bayes, bem como refinar o classificador e validar o conjunto de treinamento por meio de iterações de codificação manual. No caso do STM, é recomendável utilizar – seguindo os ensinamentos de Casas e Wilkerson (2017) e tendo em vista os 32 temas utilizados pela Casa em sua tabela de classificação[14] – ao menos seis diferentes modelos com incrementos de oito tópicos cada (8, 16, 24, 32, 40, 48). Assim, é possível dirimir o problema da instabilidade de tópicos agrupando todos os tópicos resultantes dos seis modelos e verificando quais tópicos permanecem consistentes.

Cabe lembrar que os métodos automatizados contam a frequência das palavras no discurso, não considerando o tempo desse discurso. Assim, um parlamentar pode falar a mesma palavra dezenas de vezes em um discurso de cinco minutos e outro pode falar apenas algumas vezes em um discurso de grande expediente. Apesar de a utilização de campos lexicais para definir os temas amenizar esse problema, é recomendável mensurar o tempo de debate por tema – considerado o período de início e fim de um debate como um metadado do discurso – e utilizá-lo também como um indicador de tema mais debatido na Casa em determinado período. Acredita-se que é possível conseguir esses resultados utilizando uma abordagem em que o tempo (período de início e fim) dos discursos é utilizado como covariável no método STM. Nesse mesmo sentido, é possível mensurar o tempo de debate investido em um subtema, em um conjunto de proposições, ou em uma política pública. Assim, o tempo de fala pode ser visto como um indicador de poder, ao mensurar a dominância de partidos e parlamentares nos trabalhos da Casa, bem como um indicador de desempenho, ao mensurar o tempo total investido para debater políticas públicas.

Entretanto, os temas debatidos também refletem as ideologias presentes no Parlamento. Como diria Bakhtin (1995), o discurso parlamentar é também um conteúdo ideológico, remete a algo situado fora de si mesmo ao firmar interesses e estabelecer níveis de dominação, transformando o Parlamento em uma arena onde são travadas as batalhas dos diversos interesses nacionais. Nesse sentido, as ideologias se expressam não necessariamente falando diferentemente sobre os mesmos

[13] O STM é o método não supervisionado mais recomendado por permitir a utilização de metadados como covariáveis, como, por exemplo, o tempo do discurso. Outra importante característica desse método é a possibilidade de se criar um campo lexical, como uma rede de palavras, em torno do tema. Já o Naive Bayes/Decision Tree é um algoritmo bastante testado cujo uso já é dominado por alguns servidores da Câmara.

[14] Como os 32 temas são muito amplos, a definição de subtemas é imprescindível para compreender o debate parlamentar e facilitar a classificação de conjuntos de proposições.

problemas, mas também falando sobre questões diferentes (DIERMEIER *et al*, 2012). Assim, busca-se compreender por meio de modelos estatísticos como esses conteúdos ideológicos se manifestam nas formas linguísticas e como essas refletem e/ou refratam tais conteúdos.

Empiricamente, a ideologia permite predizer o posicionamento político de um parlamentar em um assunto por meio de seu posicionamento em outro assunto não correlato (DIERMEIER *et al*, 2012). Tradicionalmente, cientistas políticos utilizam as votações nominais para estimar o posicionamento individual dos parlamentares. Entretanto, muitas vezes as votações são apenas simbólicas e não são registradas nominalmente. Também há um forte controle partidário nas votações para evitar votos dissidentes. Nesse sentido, é mais preciso estimar a diversidade de posicionamentos dos parlamentares por meio do conjunto de seus discursos (LAUDERDALE; HERZOG, 2016). Os discursos parlamentares seriam, assim, um "produto" que demonstra a conexão eleitoral.

Dado que ideologias podem se manifestar por meio da escolha de tópicos, expressões ou palavras em determinados contextos; é necessário primeiro definir o que se entende por ideologia. Pragmaticamente, segundo Converse (1964) ideologia é um conjunto de crenças que orienta o posicionamento do indivíduo em diversos assuntos (DIERMEIER *et al*, 2012). Essa é uma visão parecida com o que preconiza Cancian (2007) para pesquisas empíricas: ideologia é descrita como o conjunto de ideias, valores ou crenças que orientam percepção e o comportamento dos indivíduos sobre diversos assuntos (SCHWARTZ, 2018).

Ao definir ideologia como um conjunto de crenças que orienta o posicionamento parlamentar, é possível examinar se as posições ideológicas dos parlamentares expressas em seus discursos determinam seus votos, tendo em vista restrições institucionais, como o controle da agenda e dos partidos Diermeier *et al* (2012). Se o voto é explicado mais por uma ideologia política preexistente (como a expressa no discurso) do que por fatores institucionais[15], conhecer o posicionamento dos parlamentares para um conjunto de questões é preditivo para saber sua visão para outras questões não correlacionadas[16] (DIERMEIER *et al*, 2012).

Entretanto, dada à necessidade de reeleição do parlamentar – principalmente em sistemas com fortes incentivos ao voto pessoal e não partidário – os parlamentares precisam, além de assumir posicionamentos sobre diversos temas, fazer propaganda para sua base eleitoral e ganhar crédito. Nesse ponto, Casas e Wilkerson (2017) concordam com Mayhew (1974) e concluem: nem todo discurso parlamentar é ideológico. Não é por acaso que os resultados da pesquisa de Diermeier *et al* (2012) demonstraram que tanto Republicanos quanto Democratas passam a maior parte de seus

[15] Diermeier *et al* (2012) compara o que disseram e como votaram os Senadores estadunidenses e conclui que votar e debater são expressões diferentes, mas correlacionadas, de um mesmo sistema de crenças ideológicas.

[16] Como foi alcançada uma precisão de 94% na classificação de Senadores extremos e 52% nos senadores moderados – além do reconhecimento, por meio de experimentos adicionais, de que os senadores moderados são versões atenuadas de senadores extremos, em vez de uma categoria distinta – acredita-se que há, de fato, um campo lexical distinto para legisladores conservadores e liberais. Apesar das diversas questões discutidas no Congresso, os conservadores e os liberais sempre falam sobre qualquer um desses tópicos de maneiras distintas e estáveis (DIERMEIER *et al*, 2012).

discursos felicitando eleitores[17]. Assim, fica evidente a necessidade de separar o máximo possível o conteúdo não ideológico do corpus, retirando, por exemplo, falas procedurais, homenagens e agradecimentos.

Uma boa prática é observar as variações no uso das palavras por tópicos, antes de analisar as variações por posicionamento. Também é recomendável usar dimensões específicas em cada debate, depois sumarizar a variação nessas dimensões específicas de cada debate utilizando duas dimensões gerais – como esquerda/direita, conservador/liberal ou governo/oposição. Assim, é possível recuperar a multidimensionalidade nas estimativas de preferências com etiquetas de tópicos. Outra boa prática é medir a variação por tópico e por parlamentar, depois juntar todos os tópicos de um mesmo parlamentares para recuperar a multidimensionalidade (LAUDERDALE; HERZOG, 2016).

Como essas práticas dependem de uma minuciosa definição das duas (ou mais) dimensões gerais utilizadas, recomenda-se a criação de um grupo de estudo, com a participação de consultores da Casa, com o objetivo de criar um dicionário anotado por tópico e por política pública com as definições e o campo lexical do que seria esquerda e direita, bem como conservador e liberal, em cada um desses tópicos – tendo em vista a indexação da Casa e os trabalhos prévios sobre o assunto[18]. Em termos de métodos, todos os três conjuntos analisados podem medir posicionamento com sucesso. Há experiências com o SVM – que pode ser utilizado em conjunto com as votações nominais para esse propósito – com o Wordshoal – que combina o Wordfish com a análise fatorial Bayesiana para estimar posições – e com o X^2 – abordagem de similaridade de texto combinada com a análise fatorial Bayesiana para o mesmo propósito. É recomendável utilizar diferentes métodos para validar uns aos outros, como preconiza a literatura.

Além de medir o posicionamento, estimar posições políticas e prever votações, também é possível analisar a coerência de um parlamentar – observando a variação de seu posicionamento ao longo do tempo e a diferença entre seus discursos e suas votações nominais –, bem como a coesão de partidos, blocos e frentes parlamentares. A coesão partidária pode ser mensurada pela aglomeração dos deputados na escala ideológica ou pela similaridade de discurso. Nesse sentido, Diermeier *et al* (2012) acharam evidências de que a filiação partidária e a classificação ideológica são altamente correlacionadas nos EUA, calculando a concordância "*kappa"* como uma medida de consistência entre os rótulos ideológicos dos senadores e sua filiação partidária (houve uma concordância quase perfeita: *kappa* = 0,932). Um tipo especial de coesão é o alinhamento da base do governo, como demonstrou Schwartz (2018). Nesse sentido, é recomendável focar também na utilização de métodos e abordagens diferentes para medir o grau de coesão partidária, pois esse

[17] Os resultados indicaram que os Republicanos são mais propensos a dar discursos de felicitações (17%), seguidos de discursos sobre defesa (8%), educação (7%) e família (6%). Democratas também dão mais discursos de felicitações (9%), seguidos de discursos sobre educação (9%), saúde (7%) e família (5%).
[18] Como o livro Direita e Esquerda de Noberto Bobbio (1994) e o projeto voteview.com.

tipo de informação é de grande importância para as Lideranças da Casa.

Além dos tópicos de debate e do posicionamento no espectro ideológico, há uma terceira utilização para os métodos automatizados: a identificação de influências nos documentos e resultados do processo legislativo. Nos estudos mais recentes influência é conceituada como controle sobre resultados, mais especificamente a diferença entre o que expressamente desejam os atores envolvidos e o resultado[19] (PRITONI, 2014). Pode ser difícil desvendar o que realmente querem os atores envolvidos, mas possível identificar os atores que influenciaram uma política pública conhecendo a origem das ideias que estão expressas na legislação aprovada (BURGESS *et al*, 2016). Assim, torna-se possível e relevante identificar a influência do Poder Executivo, dos partidos, blocos e frentes parlamentares, bem como dos grupos de interesse, no processo legislativo, por meio do reuso de texto.

Nesse sentido, Burgess *et al* (2016) utilizou o algoritmo Smith-Waterman e o mecanismo de pesquisa ElasticSearch para traçar a origem de proposições. Essa mesma abordagem pode ser utilizada para verificar o reuso de texto em pareceres e emendas parlamentares. Já Hertel-Fernandez e Kashin (2015) utilizaram a combinação de três métodos – envolvendo similaridade de bigramas e trigramas, LDA e SVM – para identificar a influência de grupos de interesse na legislação estadual dos EUA. Entretanto, essas abordagens focam em produtos legislativos – como proposições, pareceres, emendas e estudos – e na própria legislação[20]. Recomenda-se a utilização da abordagem de Burgess *et al* (2016) em documentos como proposições e emendas. Além de fornecer valiosas informações sobre o reuso de texto, os resultados dessa aplicação podem também colaborar no processo de apensamento de proposições e de apreciação de emendas.

REFERÊNCIAS

BAKHTIN, Mikhail. **Marxismo e filosofia da linguagem**. São Paulo: Hucitec, 1995.

BLEI, D. Probabilistic Topic Models. **Communications of the ACM**, New York, v. 55, n. 4, p. 77-84, 2012.

BURGESS, M. *et al*. The Legislative Influence Detector: Finding Text Reuse in State Legislation. *In*: CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD), 16, San Francisco. **Anais** […], New York: ACM, 2016. p. 57-66.

CASAS, A.; WILKERSON, J. Large-scale computerized text analysis in political science: Opportunities and challenges. **Annual Review of Political Science**, Palo Alto, v. 20, p. 529-544, 2017.

[19] Entretanto, Pritoni (2014) adianta que os atores envolvidos podem exagerar suas pretensões iniciais a fim de facilitar a barganha e atingir o melhor resultado possível para o momento, independentemente se a intenção é mudar ou manter uma política pública. Assim, apenas os grupos de interesse que querem a manutenção de políticas públicas de fato mostrariam sua real intenção, não mudar nada.

[20] Em relação aos discursos parlamentares, é possível utilizar métodos parecidos para traçar a origem de palavras, expressões e argumentos nos debates parlamentares. Tal abordagem, por exemplo, pode servir para analisar a similaridade entre o discurso dos deputados e os discursos dos representantes de grupos de interesse.

DIERMEIER, D. *et al*. Language and ideology in Congress. **British Journal of Political Science**, Cambridge, v. 42, n. 1, p. 31-55, 2012.

GRIMMER, J. *et al*. How to Make Causal Inferences using Texts. **arXiv preprint arXiv:1802.02163**, New York, 2018.

GRIMMER, J. We are all social scientists now: how big data, machine learning, and causal inference work together. **PS: Political Science & Politics**, Cambridge, UK, v. 48, n. 1, p. 80-83, 2015.

GRIMMER, J.; STEWART, B. M. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. **Political Analysis**, Oxford, v. 21, n. 3, p. 267-297, jan. 2013.

HERTEL-FERNANDEZ, A.; KASHIN, K. 2015. Capturing business power across the states with text reuse. *In*: CONFERENCE OF THE MIDWEST POLITICAL SCIENCE ASSOCIATION, Chicago, **Anais** […], Chicago, Apr., 2015. p. 16-19.

IZUMI, M. **Velhas questões, novos métodos**: posições, agenda, ideologia e dinheiro na política brasileira. 2017. 113 f. Tese (Doutorado em Ciência Política) – Departamento de Ciência Política, Universidade de São Paulo, São Paulo, 2017.

IZUMI, M; MOREIRA, D. O Texto como Dado: Desafios e Oportunidades para as Ciências Sociais. **Revista Brasileira de Informação Bibliográfica em Ciências Sociais – BIB**. São Paulo, n. 86, p. 138-174, 2018.

KILGARRIFF, A.; ROSE, T. Measures for corpus similarity and homogeneity. **Information Technology Research Institute Technical Report Series**, Brighton, p. 46-52, July 1998.

KLUVER, H. **Lobbying in the European Union**: Interest groups, lobbying coalitions and policy change. Oxford: Oxford University Press, 2013.

KLUVER, H. Measuring interest group influence using quantitative text analysis. **Eur. Union Polit.**, Konstanz, v. 10, n. 4, p.535– 49, 2009.

KOHAVI, R. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. **.** *In*: THE SECOND INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING (KDD), Portland. **Anais** […], Menlo Park: AAAI, Aug. 1996. p. 202-207.

LAUDERDALE, B. HERZOG, A. Measuring political positions from legislative speech. **Political Analysis**, Cambridge, UK, v. 24, n. 3, p. 374-394, 2016.

MANNING, C.; SCHÜTZE, H. **Foundations of statistical natural language processing**. Cambridge, EUA: The MIT Press, 1999.

MAYHEW, D. Congress: **The Electoral Connection**. New Haven: Yale University Press, 1974.
MOREIRA, D. **Com a palavra os nobres deputados**: frequência e ênfase temática dos discursos dos parlamentares brasileiros. 2016. 204 f. Tese (Doutorado em Ciência Política) - Departamento de Ciência Política, Universidade de São Paulo, 2016.

PRITONI, A. How To Measure Interest Group Influence: Evidence From Italy. *In*: ECPR JOINT SESSIONS OF WORKSHOPS,42, Salamanca, **Anais** […], Salamanca: Universidad de Salamanca, 2014. p. 11-18.

ROBERTS, M. *et al*. Computer-Assisted Text Analysis for Comparative Politics. **Political Analysis**, Cambridge, UK, v. 23, p. 254-277, 2015.

ROBERTS, M. *et al*. The structural topic model and applied social science. *In*: NIPS 2013 WORKSHOP ON TOPIC MODELS: COMPUTATION, APPLICATION, AND EVALUATION, Cambridge, EUA. **Anais** […], Cambridge, EUA: Harvard University, 2014.

SCHWARTZ, F. **Análise do Discurso Parlamentar por Meio da Técnica do Processamento de Linguagem Natural**: Abordagem Estatística e Aprendizagem de Máquina. 2018. 76 f. Relatório de Pesquisa (Pós-Doutorado em Tecnologia) – Faculdade Tecnologia, Universidade de Brasília, 2018.

WILKERSON, J. *et al*. Tracing the flow of policy ideas in legislatures: a text reuse approach. **American Journal of Political Science**, Charlottesville, v. 59, n. 4, p. 943–56, 2015.

Artigo recebido em: 2018-12-27

Artigo reapresentado em: 2019-02-20

Artigo aceito para publicação em: 2019-03-12