DOI: 10.17666/bib8607/2018

O texto como dado: desafios e oportunidades para as ciências sociais[1]

Maurício Izumi[2]
Davi Moreira[3]

Introdução

A despeito da importância que a análise de conteúdo possui para as ciências sociais, a escassez de tempo, de recursos e a vulnerabilidade à qual o rigor do pesquisador está sujeito ao analisar de forma manual o conteúdo de grandes bases de documentos (*corpus*)[4] fez que boa parte das pesquisas com esse propósito se limitasse à análise de pequenos acervos. No caso da ciência política, por exemplo, são raros os trabalhos como os do *Manifesto Research Group* que, desde os anos 1970, analisa a ênfase temática de manifestos partidários[5] ou o *Comparative Agendas Project*, que coleta e analisa dados sobre agendas de políticas públicas em diferentes países[6].

Mais recentemente, porém, o avanço tecnológico e científico permitiu que técnicas automatizadas de análise do conteúdo fossem desenvolvidas e aplicadas de forma simples a grandes acervos. Longe de se distanciar desse processo, as ciências sociais como um todo têm se mantido presente nessa fronteira do conhecimento, e o campo de pesquisa *text as data* ganha cada vez mais força (MONROE; SCHRODT, 2008; ROBERTS, 2016).

Em linha com esta fronteira do conhecimento, a ciência social brasileira já apresenta de forma satisfatória resultados científicos obtidos pelo uso intensivo da análise automatizada de conteúdo. A tese de doutorado de Davi Moreira (2016), por exemplo, analisou mais de 127 mil discursos proferidos por deputados federais no Pequeno Expediente ao longo dos anos de 1999 a 2016, concluindo que a atividade parlamentar no âmbito da Câmara dos Deputados não é integralmente governada pela relação governo-oposição, como acontece no caso das votações nominais. Por sua vez, Izumi (2017) desenvolveu um modelo bayesiano da teoria da resposta ao item (TRI) para estimar posições políticas utilizando textos como dados. O autor estimou as posições políticas dos partidos brasileiros entre 1995 e 2014 utilizando os discursos dos senadores.

1 Ambos os autores contribuíram igualmente para este manuscrito, sendo sua ordem aleatoriamente definida.

2 Maurício Izumi é doutor em Ciência Política pela Universidade de São Paulo (DCP/USP) e pesquisador do Centro de Política e Economia do Setor Público da Fundação Getúlio Vargas (Cepesp/FGV). O autor contou com apoio da FAPESP, processo número 2018/08118-4. E-mail: mauricioizum@hotmail.com

3 Davi Moreira é doutor em Ciência Política pela USP e pós-doutorando pela UFPE. Vencedor do Prêmio Capes de Tese 2017 na área de Ciência Política e Relações Internacionais. Especialista em análise automatizada de conteúdo, discursos políticos e métodos quantitativos para ciências sociais. Idealizador do projeto Retórica Parlamentar, implementado pelo Laboratório Hacker da Câmara dos Deputados do Brasil. E-mail: davi.moreira@gmail.com

4 Neste artigo, os termos coleção de documentos, base de documentos, acervo e *corpus* são usados como sinônimos.

5 Para mais detalhes, ver: <https://bit.ly/1NthwlD>. Acesso em: 27 jun. 2018.

6 Para mais detalhes, ver: <https://bit.ly/2o0MEpP>. Acesso em: 27 jun. 2018.

O resultado apresenta que, em vez de uma clivagem ideológica, os discursos dos senadores organizam os partidos em uma dimensão que representa o conflito entre governo e oposição. Esse resultado indica que presidentes exercem uma influência não apenas em como os senadores votam, mas também em como eles falam.

Nesse escopo, o principal objetivo do artigo é manter as ciências sociais brasileiras na fronteira desse processo, conforme se verifica em trabalhos como o de Campos, Feres e Guarnieri (2017), e apresentar ao leitor um leque atualizado das principais metodologias de análise automatizada de conteúdo[7]. Convém ressaltar que o espaço dedicado a cada um dos métodos apresentados não reflete sua importância diante dos demais, o campo de estudos que trata o texto como dado (*text as data*) é rico e diversificado, sendo a escolha da metodologia ou de um conjunto de métodos totalmente dependente da questão de pesquisa que se deseja responder. Sem esgotar o leque de métodos, técnicas e modelos, este artigo é um guia para essa intensa e instigante área de pesquisa, apresenta conceitos essenciais para a área e provê um arcabouço relevante para o uso do texto como dado (*text as data*) nas pesquisas em ciências sociais.

Este artigo está dividido em sete seções, além desta introdução. Na primeira, "O texto como dado", apresentamos os principais conceitos, princípios e desafios da área. Na segunda, "Obtenção e pré-processamento dos dados", apresentamos procedimentos básicos que são comumente necessários antes do uso de modelos automatizados de análise de conteúdo. Na terceira, "Semelhança entre textos", apontamos técnicas para obter medidas que indiquem quão parecidos são os conteúdos de dois documentos. Na quarta, "Métodos de classificação", mostram-se técnicas para a classificação automatizada de documentos em categorias, sejam elas conhecidas ou não. Na quinta, "Métodos de escalonamento", apresentamos os dois principais algoritmos para a extração de posições políticas a partir de textos. Na sexta, "Desenvolvimentos e aplicações mais recentes", destacamos as novas tendências da área. Por fim, a última seção traz breves considerações.

Com foco sobre a língua portuguesa, ao menos um exemplo de aplicação ao caso brasileiro será apresentado nas partes de 2 a 5[8]. O acervo completo que será utilizado para aplicação das diferentes metodologias, com exceção da parte 3, refere-se aos 35.398 discursos proferidos pelos deputados federais no Pequeno Expediente durante a 54ª legislatura – 2011-2014 (MOREIRA, 2016, p. 54).

O texto como dado

Dada a complexidade da linguagem, o processo de geração, produção e seleção de dados que resultam na comunicação humana é ainda um mistério para a ciência. Tal complexidade faz com que os modelos estatísticos desenvolvidos falhem na tarefa de prover um relato preciso do processo de geração de dados utilizados na produção de conteúdo e, principalmente, em seu significado.

Os modelos de análise de conteúdo, portanto, não devem ser avaliados pelo quanto explicam do processo de geração dos dados, mas sim por sua performance ao auxiliar o pesquisador em suas atividades acadêmicas. Transformar palavras em números não substitui a leitura cuidadosa e atenta de documentos,

7 Outros trabalhos publicados já enfrentaram esse desafio. Recomendamos especialmente que o leitor tenha contato com os seguintes artigos: Grimmer e Stewart (2013), Lucas et al. (2015), Wilkerson e Casas (2017) e Welbers, Van Attenveldt e Benoit (2017).

8 Para conhecer técnicas e métodos de análise automatizada de conteúdo para múltiplas línguas disponíveis para uso nas pesquisas em perspectiva comparada, recomendamos a leitura do artigo de Lucas et al. (2015).

mas permite a análise sistemática de grandes bases de texto sem a necessidade de mão de obra em larga escala e de um enorme montante de recursos financeiros para financiamento de pesquisas, amplificando o potencial científico dos trabalhos acadêmicos.

Logo, o uso do texto como dado para análise automatizada de conteúdo permite: (1) a utilização de diferentes técnicas independentemente do idioma sob análise; (2) o cálculo de medidas de incerteza, sendo possível julgar se as diferenças entre os textos são substantivas ou apenas fruto de erros de mensuração e variação amostral; (3) reduzir a necessidade de intervenção humana, facilitando a replicabilidade dos resultados; (4) a análise de um volume de informações manualmente inviável.

Contudo, reconhecendo que "métodos de análise automatizada de conteúdo são modelos incorretos de linguagem" (GRIMMER; STEWART, 2013, p. 2), a performance de qualquer método automatizado não é garantida sem a consideração de ao menos quatro princípios (Ibidem):

1. Todos os modelos quantitativos de análise de conteúdo estão errados, mas alguns são úteis;
2. Métodos quantitativos de análise de conteúdo amplificam a capacidade humana, mas não a substitui;
3. Não há um método global para a análise automatizada de conteúdo;
4. Alidar, validar, validar.

Justamente por conhecer as vulnerabilidades dos modelos estatísticos de análise de conteúdo que esses jamais substituirão a leitura e análise humana. Como será apresentado, é por essa razão que a grande maioria dos modelos disponíveis atrelam sua capacidade inferencial a medidas e informações fornecidas pelo pesquisador antes e depois de produzidos os resultados. Para ser profícua, essa interação deve resultar da leitura cautelosa de amostras do acervo a ser analisado e da análise cuidadosa dos resultados obtidos.

A escolha do modelo, da família de modelos ou de eventuais combinações a serem utilizadas é resultado dos objetivos de pesquisa almejados. Como será visto, há uma variedade de modelos disponíveis e nenhum deles se sobrepõe aos demais. Pelo contrário, são indicados para questões e aplicações diante do auxílio analítico que o pesquisador deseja receber e das perguntas que deseja responder. Mais do que sobrepor uma abordagem a outra, convém ao pesquisador identificar a abordagem que melhor lhe atende. Neste artigo, por exemplo, o maior ou menor espaço dedicado a cada um deles não deve ser visto como algo que reduz ou aumenta sua importância para propósitos científicos.

Como sugerem Grimmer e Stewart (2013), o que precisa ser evitado é o uso cego de qualquer método. Por essa razão, em acordo com esses pesquisadores, desencorajamos o uso de *softwares* comerciais para análise quantitativa de textos. A despeito de resultados positivos, por vezes é impossível conhecer o método aplicado e o pesquisador se torna refém dos resultados apresentados sem a possibilidade de interagir e interferir no procedimento de análise do modelo.

Na próxima seção apresentamos o processo de obtenção/coleta de dados e o pré-processamento dos documentos brutos.

Obtenção e pré-processamento dos dados

Obtenção dos dados

A disponibilidade de uma infinidade de informações publicadas na internet ampliou de forma exponencial a possibilidade de acesso a um volume expressivo de conteúdo. Atualmente, projetos de lei, documentos históricos, discursos, material de campanha eleitoral, entre outras

fontes, podem ser acessados de qualquer parte do planeta sobre qualquer parte do planeta.

Se de um lado o acesso foi ampliado, de outro multiplicaram-se as formas de obtenção de conteúdo para análise automatizada nas pesquisas em ciências sociais. A um custo baixíssimo, por exemplo, o conteúdo virtualmente publicado pode ser obtido com o uso de métodos de raspagem de dados (*web scraping*), pelo qual o computador é programado para acessar páginas na internet, copiar seu conteúdo e organizar os dados para o pesquisador. De forma adicional, como destacam Berinsky, Huber e Lenz (2012), mesmo quando dados não estão facilmente disponíveis na internet, plataformas on-line, como a *Mechanical Turk* da Amazon, podem ser alternativas eficientes para sua obtenção e compilação.

A maior dificuldade na obtenção de acervo para análise automatizada de conteúdo reside no interesse de análise de documentos hitóricos. Para tanto, é necessário o escaneamento em alta qualidade de documentos e o uso de um bom *software* para reconhecimento óptico de caracteres (OCR). Para teste fim, já existem pacotes[9] na linguagem R[10] que utilizam uma Interface de Programação de Aplicativos (API) do Google[11] para execução dessa tarefa. Além disso, existem diversos pacotes que facilitam a extração de dados da internet, como o "xml2" (WICKHAN; HESTER; OOMS, 2018), o "httr" (WICKHAM, 2014) e o "rvest" (Idem, 2016).

Para os exemplos de aplicação que serão apresentados ao longo deste artigo, os mais de 35 mil discursos proferidos pelos deputados federais no Pequeno Expediente ao longo dos anos de 2011 e 2014 foram obtidos através do webservice da Câmara dos Deputados com o uso de um programa desenvolvido em linguagem R para raspagem dos dados publicados (MOREIRA, 2016). Com o fantástico trabalho de registro e divulgação dos pronunciamentos parlamentares, praticamente tudo que é dito nos momentos institucionais previstos pelo Regimento Interno da Câmara dos Deputados (RICD) é gravado, arquivado e disponibilizado ao cidadão pela internet em linguagem de máquina sob escopo do projeto Dados Abertos da Câmara dos Deputados[12].

Pré-processamento dos dados[13]

Como já apontamos, dada a complexidade da linguagem, a análise automatizada de conteúdo pressupõe a perda de informação para ganho de eficiência dos modelos. Nesse sentido, o pré-processamento dos dados exige a adoção de procedimentos que reduzam a dimensionalidade com a qual os modelos de análise de conteúdo vão lidar. A título de exemplificação, apresentamos um roteiro básico comum para o pré-processamento dos dados.

O roteiro básico de pré-processamento consiste em cinco passos: (1) codificação de caracteres (*encoding*); (2) remoção de palavras e conteúdo desnecessário; (3) construção de uma sacola de palavras (*bag of words*); (4) *stemming*; (5) construção da matriz de documentos e termos (*Document Term Matrix* – DTM).

1. **Codificação de caracteres:** Como apontam Lucas et al. (2015), apesar dos métodos estatísticos para análise de conteúdo serem

9 Atualmente, o principal pacote para esta tarefa é o "tesseract" (OOMS, 2018).

10 O R é uma linguagem e um ambiente de programação estatística. Para mais detalhes, acesse: <https://bit.ly/1gm1uk2> Acesso em: 8 jul. 2018.

11 Para mais detalhes, acesse: <https://cloud.google.com/vision/>. Acesso em: 29 jun. 2018.

12 Para mais detalhes, acesse: <https://bit.ly/2w7kY6Z>. Acesso em: 29 jun. 2018.

13 Para referências sobre o pré-processamento de acervos com múltiplas línguas, consulte Lucas et al. (2015).

agnósticos em relação à liguagem, as ferramentas para pré-processar os textos não são. Isso significa que o pré-processamento precisa ter em conta o modo como o computador interpreta a informação a ele repassada. A codificação de texto é a maneira pela qual o computador traduz caracteres individuais e únicos em *bytes* para seu armazenamento na memória. Desse modo, dados de várias fontes diferentes e em línguas diferentes têm grande chance de estar em codificações distintas, sendo necessário ao analista converter todos os documentos do acervo para a mesma codificação. Em seguida, o analista deve garantir que o computador interprete corretamente a codificação dos caracteres[14].

2. **Remoção de palavras e conteúdo desnecessário:** Harmonizada a codificação dos caracteres para a correta interpretação por parte do computador, o primeiro procedimento de redução de dimensionalidade a ser considerado é a remoção de palavras e elementos do texto que não representem substantivamente o alvo de interesse do pesquisador. Pontuação, números, artigos, pronomes e preposições (*stop words*) costumam ser retirados dos documentos[15]. Em seguida, é conveniente a retirada de uma lista predeterminada de palavras irrelevantes para os propósitos da pesquisa, bem como palavras pouco ou muito frequentes que apareçam em menos de 1% ou mais de 99% dos documentos (HOPKINS; KING, 2010; QUINN et al. 2010)[16]. A título de exemplo, se considerarmos a seguinte sentença: "O Partido dos Trabalhadores é contra a reforma trabalhista", nesta etapa do pré-processamento, teríamos: "partido trabalhadores contra reforma trabalhista".
3. **Sacola de palavras (*bag of words*):** De forma geral, os métodos automatizados de análise de texto tratam documentos como um vetor composto por palavras, desconsiderando a ordem em que aparecem. Com essa abordagem, cada documento é representado como um único vetor, sendo seu comprimento igual ao número de palavras únicas que possui.
4. ***Stemming***[17]**:** Até aqui, mesmo que os procedimentos adotados tenham diminuído enormemente a dimensionalidade do acervo de documentos ao transformá-lo em uma sacola de palavras, ainda é necessária a adoção de procedimentos que possam reduzir a complexidade do conteúdo a ser analisado. Nesse sentido, podemos pensar que determinado documento tenha em sua composição as seguintes palavras únicas: trabalho, trabalhador, trabalhista. Apesar de seus diferentes signficados, cada uma dessas palavras pode ser reduzida ao seu radical, *trabalh*, dando ao pequisador informações suficientes para sua análise e, assim reduzindo, o *n* de três palavras únicas para uma palavra que tem a soma das frequências anteriores. Para garantir que

14 No caso da linguagem R de programação, a função *encoding* é de grande importância. Para mais detalhes: <https://bit.ly/2w81Cyw>. Acesso em: 30 jun. 2018.

15 O pacote "tm" (FEINERER; HORNIK, 2018) da linguagem R pode ser utilizado para a remoção de *stop words*. Ele utiliza o acervo disponível no projeto *Snowball* para realizar as operações. Para mais informações acessar: <https://bit.ly/2OYva8V>. Acesso em: 30 jun. 2018.

16 Como apontam Lucas et al. (2015), para cada idioma, a escolha de quais palavras devem ser removidas é uma decisão substantiva que, em alguns casos, pode ter efeitos importantes nos resultados da análise (CAMPBELL; PENNEBAKER, 2003; FOKKENS et al., 2013).

17 O processo de *stemming* é uma aproximação do processo de lematização, que reduz palavras às suas formas básicas (JURAFSKY; MARTIN, 2009; MANNING; RAGHAVAN; SCHÜTZE, 2008).

palavras que variam apenas na flexão, número ou conjugação sejam consideradas iguais, reduzindo o número de dimensões contido no acervo, por meio da adaptação do algoritmo de Porter (1980) para o português já desenvolvida por diferentes projetos (*Snowball* e NILC-USP), podem ser obtidos os *stems* das palavras restantes[18]. Retomando a sentença de exemplo, nesta etapa do pré-processamento teríamos: "part trabalh contr reform trabalh".

5. **Matriz de Documentos e Termos – DTM:** Com os procedimentos até aqui adotados, cada documento i ($i=1, \ldots, N$) é representado por um vetor que contabiliza o número de vezes que cada uma das M palavras únicas ocorrem, $P_i = (P_{i1}, P_{i2}, \ldots, P_{iM})$. Cada P_{im} representa o número de vezes a *m-ésima* palavra ocorre no *i-ésimo* documento. Estruturando lado a lado os vetores de contagem forma-se a Matriz de Documentos e Termos (*Document Term Matrix* – DTM). Essa matriz será esparsa (com grande número de zeros) e conterá a frequência de cada termo (palavra ou *stem*) para cada documento do acervo. É essa matriz a matéria-prima para os modelos de análise automatizada de conteúdo.

Aplicação para o caso brasileiro – pré-processamento dos dados

Como destacado, a aplicação dos procedimentos apresentados deve levar em consideração o modelo que será utilizado para a análise automatizada do conteúdo do acervo e os propósitos da análise. Desse modo, em posse do conteúdo dos 35.398 discursos proferidos pelos deputados federais no Pequeno Expediente durante a 54ª legislatura (2011-2014), os procedimentos de tratamento dos dados foram aplicados para o uso do *expressed agenda model* (GRIMMER, 2010) com o objetivo de se identificar os temas proferidos pelos parlamentares em seus discursos[19].

Aplicados para todos os discursos cujos oradores proferiram no mínimo mais de uma fala[20], obteve-se uma *Document Term Matrix* (DTM), $d \times P$, com 33.941 linhas e 3.906 colunas, cujas linhas representam cada discurso e as colunas representam cada *stem* presente no *corpus*. É essa matriz que será utilizada na estimação dos tópicos[21] das falas dos deputados na seção 4.

Como enfatizado ao longo de toda a seção, é possível ver, pela última coluna da Tabela 1, que a aplicação do pré-processamento dos dados reduz de forma drástica o volume informacional do conteúdo presente nos documentos. Para o caso em que foi aplicado, de mais de 150 mil palavras únicas, o modelo terá como matéria-prima de análise pouco menos de quatro mil *stems*[22]. Apesar da sensação de que não será possível obter resultados robustos após tamanha perda de informação, além de como será visto mais adiante neste artigo, pesquisadores têm mostrado que a representação do acervo por meio

18 Ver: <http://snowball.tartarus.org/> e <https://bit.ly/2OXkVS6>. Acesso em: 30 jun. 2018.

19 Mais detalhes sobre o *expressed agenda model* estão apresentados na quarta parte deste artigo.

20 Grimmer (2010) retira do *corpus* todas as palavras que estão presentes em 0,5% dos documentos e também retira dos documentos de cada autor as palavras que aparecem em mais de 90% dos documentos de cada um. Além desse procedimento, neste trabalho, foram retiradas palavras selecionadas após análise de uma amostra aleatória de discursos.

21 Neste artigo as palavras *tópico* e *tema* são usadas como sinônimos.

22 Além da DTM, o *expressed agenda model* recebe como argumento uma matriz de autores, $n \times 2$, cujas linhas representam cada autor, a primeira coluna representa a linha do primeiro discurso do autor na DTM e a segunda coluna representa a linha do último discurso deste autor na DTM.

da DTM é suficiente para inferências substativas a respeito de seu conteúdo (HOPKINS; KING, 2010).

Na próxima seção apresentaremos um conjunto de técnicas utilizadas para mensurar quão similiares são dois documentos.

Tabela 1
Resultado do tratamento aplicado à coleção de discursos

	Oradores	Discursos (*D*)	Palavras únicas (*P*)
# Inicial	591	35.398	153.111*
# Final	552	33.941	3.906**

* Número de palavras únicas antes do processo de *stemming*.
** Número de *stems* únicos.

Semelhança entre textos

Medidas de similaridade são ferramentas muito interessantes para aqueles que analisam o texto como dado e buscam identificar semelhanças entre textos. Essas medidas podem ter diversas aplicações nas ciências sociais. A primeira é o uso como uma técnica exploratória. Textos são entidades complexas e em grandes quantidades se torna difícil saber se existe algum padrão entre os documentos. A partir de medidas de similaridade podemos ver como os dados se estruturam e se relacionam. Além disso, elas podem ser utilizadas em tarefas mais complexas, como traçar a origem de projetos de lei (LI; LAROCHELLE; LO, 2014; WILKERSON; SMITH; STRAMP, 2015) e verificar como grupos de interesse e da sociedade civil influenciam o conteúdo das leis (GARRETT; JANSA, 2015; KROEGER, 2015). A seguir veremos dois tipos de métodos para análise de similaridade.

Similaridade de cosseno

Esse primeiro método busca avaliar quanto um documento é similar a outro como um todo. Assume-se que quanto maior a similaridade na frequência relativa das palavras utilizadas, maior será a similaridade do conteúdo entre os textos. Ou seja, busca-se obter uma medida que indique quão parecido é o uso das palavras em dois documentos.

Partimos da ideia de que um texto pode ser representado como um vetor (*bag of words*). Esse vetor é representado em um espaço de dimensão igual ao comprimento do universo de palavras utilizadas (vocabulário). Vamos supor que queremos comparar dois projetos de lei, PL1 e PL2. Vamos supor também que os deputados que redigiram esses projetos possuem um vocabulário reduzido. Eles conhecem apenas cinco palavras: "política", "orçamento", "presidente", "futebol" e "copa". Assim, o PL1 e o PL2 podem ser definidos como vetores (u e v, respectivamente) de dimensão igual a 5, R^5. Na Tabela 2 apresentamos a frequência de palavras em cada projeto.

Tabela 2
Frequência de palavras no PL1 e no PL2

PL	política	orçamento	presidente	futebol	copa
PL1	1	10	15	30	0
PL2	0	15	20	0	3

A partir dessa tabela podemos construir os vetores u e v como $u = (1, 10, 15, 30, 0)$ e $v = (0, 15, 20, 0, 3)$. Uma forma de avaliarmos quão similares são esses vetores é calculando o produto interno entre eles, isto é, $u \cdot v = \sum_{i=1}^{5} u_i v_i$. Quando os dois vetores possuem frequências altas para as mesmas palavras, maior o produto interno entre eles. Em outras palavras, se observamos um produto interno alto, temos evidências de que a distribuição do conjunto de palavras nos dois documentos é similar. No exemplo dado, o produto interno é igual $u \cdot v = (1 \times 0) + (10 \times 15) + (15 \times 20) + (30 \times 0) + (0 \times 3) = 450$.

Como é possível observar, essa medida ainda é problemática. Quanto maior o comprimento do vetor $\left(|v| = \sqrt{\sum_{i=1}^{5} v_i^2}\right)$, maior será o produto interno. Assim, documentos que possuem um vocabulário mais rico, bem como aqueles que utilizam palavras muito frequentes, provavelmente terão um produto interno maior.

A solução para esse problema é dividir o produto interno pelo produto dos comprimentos dos vetores, $\frac{u \cdot v}{|u| \cdot |v|}$. Mas, se lembrarmos da geometria analítica, em que $u \cdot v = |u| \cdot |v| \cdot \cos \theta$, temos que $\cos \theta = \frac{u \cdot v}{|u| \cdot |v|}$. Logo, $\cos \theta$, que representa o cosseno do ângulo formado entre os vetores u e v, indica quão similares são os vetores. Essa medida é conhecida como similaridade de cosseno (JURAFSKY; MARTIN, 2009) e varia de zero a um. Quanto mais próxima de um, maior a similaridade dos vetores.

Quadro 1
Comparação entre o PL1375/2011 e o PL7480/2014

PL1375/2011	PL7480/2014
Altera a redação do art. 11 da Lei no 11.180, de 23 de setembro de 2005, no que se refere ao valor da bolsa-permanência do Programa Universidade para Todos – Prouni.	Altera a redação do art. 11 da Lei no 11.180, de 23 de setembro de 2005, com relação à concessão de bolsa-permanência para estudantes beneficiários do Programa Universidade para Todos (Prouni).
O Congresso Nacional decreta:	O Congresso Nacional decreta:
Art. 1º O art. 11 da Lei no 11.180, de 23 de setembro de 2005, passa a vigorar com a seguinte redação:	Art. 1º O art. 11 da Lei no 11.180, de 23 de setembro de 2005, passa a vigorar com a seguinte redação:
Art. 11. Fica autorizada a concessão de bolsa-permanência, no valor de até um salário mínimo mensal, exclusivamente para custeio das despesas educacionais, a estudante beneficiário de bolsa integral do Programa Universidade para Todos – Prouni, instituído pela Lei nº 11.096, de 13 de janeiro de 2005, matriculado em curso de turno integral, conforme critérios de concessão, distribuição, manutenção e cancelamento de bolsas a serem estabelecidos em regulamento, inclusive quanto ao aproveitamento e à freqüência mínima a ser exigida do estudante. (NR)	Art. 11. Fica autorizada a concessão de bolsa-permanência, até o valor equivalente ao praticado na política federal de concessão de bolsas de iniciação científica, exclusivamente para custeio das despesas educacionais, a estudantes beneficiários de bolsa integral do Programa Universidade para Todos (Prouni), instituído pela Lei nº 11.096, de 13 de janeiro de 2005, conforme critérios de concessão, distribuição, manutenção e cancelamento de bolsas a serem estabelecidos em regulamento, inclusive quanto ao aproveitamento e à frequência mínima a ser exigida do estudante.
Art. 2º Esta lei entra em vigor na data de sua publicação.	Parágrafo único. Os critérios de concessão referidos no caput considerarão especialmente a situação de impossibilidade de compatibilidade entre a frequência ao curso, em turno parcial ou integral, e o exercício de atividade remunerada, no caso de o estudante não contar com renda própria ou familiar suficiente para prover sua subsistência. (NR)

Fonte: Dados abertos da Câmara dos Deputados.

No Quadro 1 apresentamos um exemplo concreto no qual comparamos dois projetos de lei. O primeiro é o PL1375/2011, de autoria da deputada professora Dorinha Seabra Rezende, do partido Democratas do Tocantins (DEM-TO), e o segundo é o PL7480/2014, de autoria do deputado Gustavo Petta, do Partido Comunista do Brasil de São Paulo (PCdoB-SP). A semelhança entre os projetos é clara. O cosseno do ângulo formado pelos dois vetores é igual a 0,85[23].

Uma limitação desse método é o fato de ele desconsiderar completamente o ordenamento das palavras. Vejamos um exemplo que ressalta essa limitação. Se observamos as seguintes frases: (1) "faça amor, não faça guerra"; e (2) "faça guerra, não faça amor", a similaridade de cosseno será máxima pois ambas as frases utilizam o mesmo conjunto de palavras. No entanto, elas possuem sentidos completamente opostos. Ao desconsiderar o ordenamento das palavras, essa medida não é capaz de diferenciar esse tipo de detalhe. Obviamente, esse é um caso extremo. No geral, documentos com sentidos opostos tendem a utilizar palavras diferentes. A seguir apresentamos um método que enfrenta esse desafio.

Algoritmo de Smith-Waterman

Na similaridade de cosseno, estamos interessados em quanto um documento é similar a outro como um todo. Mas outra tarefa que um cientista social pode estar interessado é aquela em que buscamos os trechos de um documento que são semelhantes a trechos de outros documentos. Nessa tarefa estamos interessados na similaridade de apenas um pedaço do texto. A família de algoritmos que cumpre esse tipo função é a de alinhamento local e sua origem está na biologia molecular.

Nesse campo de estudos é comum querer saber quão similares são as sequências genéticas de organismos diferentes. O algoritmo mais utilizado para essa tarefa é o de alinhamento local de Smith-Waterman (SMITH; WATERMAN, 1981). Ele compara segmentos das sequências de todos os comprimentos possíveis de modo a maximizar a similaridade entre eles. A similaridade é baseada em uma função que atribui um valor positivo quando há um *matching* nas duas sequências (*matching* = 2) e penaliza com valores negativos quando há um *mismatching* (*mismatching* = –1) ou um *gap* (*gap* = –1) nas sequências. Por exemplo, se observarmos as sequências "AAACGTCA" e "CGTA", podemos tentar alinhar os trechos "AAAC" e "CGTA" (índice: –1–1–1–1 = –4). Outra possibilidade é alinhar "CGTCA" e "C#GTA" (índice: 2 – 1 – 1 – 1 + 2 = –1). Aqui incluímos um *gap* (#) na segunda posição da segunda sequência. O melhor alinhamento possível das duas sequências (isto é, aquele que gera o maior valor no índice) é dado pelos trechos "CGTCA" e "CGT#A" (índice: 2 + 2 + 2 – 1 + 2 = 7). Esse seria o alinhamento encontrado pelo algoritmo.

A diferença é que, em vez de comparar sequências de nucleotídeos em cadeias de DNA, vamos comparar sequências de palavras em textos. Esse algoritmo parece ser apropriado para diversos problemas, como a comparação de projetos de lei, por dois motivos. O primeiro é o fato de ele fazer comparações locais e não globais. Isso é relevante porque dois documentos semelhantes podem compartilhar apenas alguns trechos e serem bem diferentes em todos os outros. O segundo ponto está relacionado ao fato de ele permitir pequenas diferenças nos trechos ao adicionar os *gaps*. Isso é relevante, já que mudanças na linguagem são esperadas quando ideias migram de um contexto a outro[24].

23 Para mais detalhes, ver Izumi (2017).

24 Para uma aplicação ao caso norte-americano, ver Wilkerson, Smith e Stramp (2015).

Quadro 2
Resultado do algoritmo de Smith-Waterman para a comparação entre o PL1375/2011 e o PL7480/2014

PL1375/2011	PL7480/2014
Altera a redação do art 11 da Lei nº 11.180 de 23 de setembro de 2005 no ### que ####### se # refere ######### ao ## valor da bolsa permanência #### ########## ############# do Programa Universidade para Todos PROUNI	Altera a redação do art 11 da Lei nº 11.180 de 23 de setembro de 2005 ## com ### relação ## à ###### concessão ## de ##### ## bolsa permanência para estudantes beneficiários do Programa Universidade para Todos Prouni
O Congresso Nacional decreta	O Congresso Nacional decreta
Art 1º O art 11 da Lei no 11.180 de 23 de setembro de 2005 passa a vigorar com a seguinte redação	Art 1º O art 11 da Lei no 11.180 de 23 de setembro de 2005 passa a vigorar com a seguinte redação
Art 11 Fica autorizada a concessão de bolsa permanência no ### # valor ########## ## ######### ## ######## ####### de até ######### um ## salário ###### mínimo ##### mensal ######### ########## exclusivamente para custeio das despesas educacionais a estudante ########## beneficiário ############# de bolsa integral do Programa Universidade para Todos Prouni instituído pela Lei nº # # 11.096 de 13 de janeiro de 2005 matriculado em curso de turno integral conforme critérios de concessão distribuição manutenção e cancelamento de bolsas a serem estabelecidos em regulamento inclusive quanto ao aproveitamento e à freqüência ########## mínima a ser exigida do estudante	Art 11 Fica autorizada a concessão de bolsa permanência ## até o valor equivalente ao praticado na política federal de ### concessão ## de ####### bolsas ###### de ###### iniciação científica exclusivamente para custeio das despesas educacionais a ######### estudantes ########### beneficiários de bolsa integral do Programa Universidade para Todos Prouni instituído pela Lei ## nº 11.096 de 13 de janeiro de 2005 ########## ## ##### ## ##### ######## conforme critérios de concessão distribuição manutenção e cancelamento de bolsas a serem estabelecidos em regulamento inclusive quanto ao aproveitamento e à ########## frequência mínima a ser exigida do estudante

Fonte: Dados Abertos da Câmara dos Deputados.

No Quadro 2 apresentamos o resultado da aplicação do algoritmo de Smith-Waterman para o PL1375/2011 e o PL7480/2014. Como podemos observar, esse algoritmo conseguiu captar de forma precisa a semelhança entre os projetos. A ementa foi mantida com a inclusão de alguns *gaps*, assim como o artigo 11. O 1º artigo se manteve inalterado. Embora o artigo 2º de ambos projetos sejam idênticos, ele não foi alinhado, pois sua inclusão levaria a uma penalidade muito grande, já que o parágrafo único aparece em apenas um dos projetos.

Métodos de classificação

A classificação automatizada organiza o acervo de documentos em categorias, sejam elas conhecidas ou não. Utilizaremos esse critério básico, conhecimento ou desconhecimento das categorias nas quais deve ser classificado o acervo, para apresentar métodos de cassificação automatizada.

Categorias conhecidas

Ancorados na teoria, pesquisadores em ciências sociais almejam classificar documentos em categorias já conhecidas. Inúmeras são as aplicações, mas entre elas descatam-se duas: (1) o uso de métodos de dicionário; e (2) métodos de aprendizagem supervisionada (*supervised learning methods*).

Métodos de dicionário

É comum que pesquisadores queiram saber se um documento tem conotação positiva ou negativa sobre um tópico qualquer. Podemos

usar o método de análise de sentimentos para esse tipo de tarefa. A análise de sentimentos está associada ao uso da análise quantitativa de textos para extração de estados afetivos contidos em documentos. Esses estados afetivos são conhecidos como sentimentos ou opiniões. Em geral, na análise de sentimentos, o foco está sobre opiniões que expressam sentimentos positivos ou negativos. Queremos saber se um consumidor possui opinião positiva ou negativa sobre o produto ou serviço que ele adquiriu; se as propagandas de um candidato em campanha política possuem conotação positiva ou negativa; se a cobertura jornalística sobre determinado candidato é positiva ou negativa (LIU, 2012; PANG; LEE, 2008).

De modo mais preciso, um sentimento é definido como uma quádrupla (*g, s, h, t*), em que *g* é o alvo do sentimento, *s* é o sentimento a respeito do alvo, *h* é o detentor da opinião e *t* é o momento em que a opinião foi expressada. Um bom exemplo está na frase "eu tenho ódio e nojo à ditadura", proferida por Ulysses Guimarães no momento da promulgação da Constituição de 1988. Nessa frase, Ulysses Guimarães é o detentor da opinião (*h*). O alvo de seu sentimento negativo (*s*) é a ditadura (*g*) e seu sentimento foi expressado no dia 5 de outubro de 1988 (*t*).

Uma das abordagens mais simples na análise de sentimentos, ou de modo mais geral, para a classificação de documentos em categorias preestabelecidas, é a abordagem por meio de dicionários anotados (TABOADA et al., 2011). Como o nome sugere, nessa abordagem, o sentimento de um documento é determinado com o auxílio de um dicionário no qual as palavras são anotadas com sua orientação semântica (positiva ou negativa)[25]. Assim, palavras como "ódio" e "nojo" possuem conotação negativa, já palavras como "amor" e "gostar" têm conotação positiva. Desse modo, documentos que utilizam mais palavras positivas do que negativas são classificados como documentos que expressam um sentimento positivo.

Formalmente, sejam W_{dp} o número de vezes que a palavra *p* aparece no documento *d* e s_p o sentimento associado à palavra *p*. Para simplificar, $s_p = -1$, quando o sentimento associado for negativo e $s_p = 1$, quando o sentimento associado for positivo. O sentimento s_d de um documento *d* é:

$$s_d = \sum_{p=1}^{M} \frac{s_p W_{dp}}{\sum_{p=1}^{M} W_{dp}}$$

Classificamos um documento como positivo, se $s_d > 0$ e o classificamos como negativo, caso contrário.

Métodos de aprendizagem supervisionada (*Supervised learning methods*)

Métodos de aprendizado supervisionado replicam a familiar tarefa de codificação manual, porém com enorme redução de custos e grande ganho de escala. Sua implementação pressupõe a classificação manual de uma amostra do acervo em um conjunto predeterminado de categorias. Essa amostra classificada, conhecida como conjunto de treinamento ou *training set*, é usada para treinar modelos estatísticos, cuja principal aplicação é a classificação do restante do acervo, conjunto de teste ou *test set*, nas categorias predeterminadas. Ao final da classificação, procedimentos de

25 O Grupo de Processamento da Linguagem Natural da Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS) possui o *OpLexicon*, um léxico de sentimento para a língua portuguesa (SOUZA; VIEIRA, 2012; SOUZA et al., 2011). Na atual versão, esse dicionário apresenta 31.719 termos dos quais 14.254 carregam sentimentos negativos, 8.469 sentimentos positivos e 8.996 sentimentos neutros.Mais informações podem ser encontradas em: <https://bit.ly/2OZGxNP> Acesso em: 7 jul. 2018.

validação devem ser adotados para se averiguar a performance do modelo utilizado.

Vejam que o método de classificação a partir de dicionários anotados não é supervisionado. Essa abordagem tem como vantagem não depender de regras ou procedimentos criados fora do domínio específico do conteúdo que está sob análise. Tal característica evita problemas de uso de referências externas, como pode ocorrer no caso de modelos de dicionário. Métodos de aprendizado supervisionado requerem que os pesquisadores desenvolvam regras de codificação manual para as categorias de interesse. Tal necessidade força os analistas a desenvolverem definições coerentes de conceitos para aplicações particulares, o que leva à clareza sobre a classificação pretendida. Outra vantagem dos métodos de aprendizado supervisionado é a facilidade de validação e verificação da performance do modelo utilizado.

Como apontam Grimmer e Stewart (2013), todos os métodos de aprendizagem supervisionada pressupõem três etapas básicas após os procedimentos de pré-processamento já apresentados: (1) construir um conjunto de treinamento (*training set*); (2) aplicar o método de aprendizado supervisionado; e (3) validar a saída do modelo.

1. **Construção de um conjunto de treinamento – *training set***
   Esquema de codificação: Para a construção de um conjunto de treinamento, deve ser criado um esquema de codificação manual que supere dificuldades relacionadas a ambiguidades na linguagem, a atenção limitada dos codificadores e o entendimento sobre conceitos presentes no acervo. Com um livro de códigos elaborado, devem-se realizar exercícios de testes para que sejam identificadas ambiguidades no esquema de codificação ou nas categorias negligenciadas. Esse procedimento leva, subsequentemente, a uma revisão do livro de códigos, que então precisa ser aplicado a um novo conjunto de documentos para assegurar que as ambiguidades tenham sido suficientemente resolvidas. Logo, somente após os codificadores aplicarem o esquema de codificação aos documentos sem perceber ambiguidades, o esquema estará pronto para ser aplicado ao restante do conjunto de dados[26].
   Seleção do conjunto de treinamento: Idealmente, os documentos presentes no conjunto de treinamento devem ser representativos do acervo. Logo, para um bom desempenho do modelo, é aconselhavel que o conjunto de treinamento seja construído a partir de uma amostra aleatória da coletânea à qual pertencem (HAND, 2006). Isto posto, resta saber qual a quantidade ideal de documentos para o conjunto de treinamento. Hopkins e King (2010) indicam quinhentos como regra geral, sendo cem documentos já suficientes para alguns casos. No entanto, o número ideal dependerá da aplicação específica de interesse, pois, conforme o número de categorias aumenta, mais documentos são necessários em cada categoria do conjunto de treinamento para uma boa performance do modelo. Como será exposto adiante, uma vantagem do uso da aprendizagem supervisionada para

26 Como apontam Grimmer e Stewart (2013), a criação de esquemas de codificação é uma literatura rica em ciências sociais. Para mais detalhes, ver Krippendorff (2004), Neuendorf (2002), Weber (1990) e a documentação disponível no pacote do R "ReadMe" (HOPKINS; KING, 2017).

classificação é a possibilidade de usar o processo de validação para verificar se não é necessário aumentar o N de documentos do conjunto de treinamento.

2. **O uso do método de aprendizagem supervisionada**

Os métodos de aprendizagem supervisionada são diversos, mas compartilham de uma estrutura comum (HASTIE; TIBSHIRANI; FRIEDMAN, 2001). Como apresentam Grimmer e Stewart (2013), suponha que existam N_{train} documentos ($i=1, \ldots, N_{train}$), sendo esse o conjunto de treinamento manualmente classificado em uma das K categorias, ($k=1, \ldots, K$). A categoria de cada documento i é representada por $Y_i \hat{I} \{C_1, C_2, \ldots, C_k\}$ e todo o conjunto de treinamento é representado como $Y_{train} = (Y_1, \ldots, Y_{Ntrain})$. Cada modelo de aprendizagem supervisionada assume que existe uma função, f, não observada que associa a configuração de palavras dos documentos às categorias preestabelecidas, $Y_{train} = f(P_{train})$. Assim, com base no conjunto de treinamento, o algoritmo busca compreender essa associação e replicá-la aos demais documentos, $\hat{Y}_{train} = \hat{f}(P_{train})$. Com base nessa estrutura comum, apresentamos a seguir um método de inferência da relação entre palavras e categorias para classificação individual de documentos.

Naive Bayes (MARON; KUHNS, 1960)[27]**:** Esse é um dos mais simples e poderosos métodos de classificação individual. Nele, o conjunto de treinamento é usado para aprender sobre a distribuição de palavras para documentos de cada categoria k. Essa distribuição é usada para classificar cada um dos documentos no conjunto restante do acervo. Com base na regra de Bayes, o modelo opera essa classificação ao inferir a probabilidade de que o documento i pertença à categoria k, dado o perfil de palavras P_i. Aplicando a regra de Bayes, $p(C_k|P_i) \mu p(C_k)\, p(P_i|C_k)$, sabemos que é necessário estimar $p(C_k)$ e $p(P_i|C_k)$. Sendo o conjunto de treinamento uma boa representação do acervo, temos que o estimador de máxima verossimilhança de $p(C_k)$ é dado pela proporção de documentos do conjunto de treinamento em cada categoria k. Por sua vez, a estimação de $p(P_i|C_k)$ é mais complexa e necessita do pressuposto (*naive assumption*) de que, dada a categoria de um documento, as palavras são geradas de forma independente, $p(P_i \mid C_k) = \prod_{i=1}^{M} p(P_{im} \mid C_k)$. Obviamente, tal pressuposto é equivocado, visto que o uso de palavras é altamente correlacionado. Contudo, mesmo com esse pressuposto, o modelo ainda é capaz de capturar informações úteis para classificação dos documentos. Usando essa suposição, a estimativa de $p(P_{im}|C_k)$ é dada por:

$$\hat{p}\,(P_{im} = j|C_k) = \frac{\text{\#documentos do conjunto de treinamento na categoria } k \text{ e com a palavra } m \text{ usada } j \text{ vezes}}{\text{\#documentos na categoria } k}$$

27 Em vez de classificar cada documento do acervo, pode ser de interesse do pesquisador somente ter informação sobre a proporção de documentos presentes em cada categoria. Entre os modelos disponíveis para mensuração de proporções está o ReadME, desenvolvido por Hopkins e King (2010).

Dado o número de zeros presentes na DTM, algumas contagens específicas de palavras nunca ocorrem no conjunto de dados. Grimmer e Stewart (2013) apontam que a solução comum é adicionar uma pequena quantidade a cada probabilidade. Desse modo, o classificador estimado para Naive Bayes fica:

$$\hat{f}(P_i) = \arg\max_k \left[\hat{p}(C_k) = \prod_{i=1}^{M} \hat{p}(P_{im} \mid C_k) \right]$$

Não sendo as palavras condicionalmente independentes, o classificador de Naive Bayes se ajusta perfeitamente ao primeiro princípio apresentado por Grimmer e Stewart (2013), de que todos os modelos quantitativos de análise de conteúdo estão errados, mas alguns são úteis[28].

3. **Validação**
Se o método de aprendizagem supervisionada aplicado tiver bom desempenho, ele será capaz de se assemelhar à classificação manual nas tarefas de codificação de documentos em categorias ou medir a proporção de documentos em categorias. Esse objetivo claro implica um padrão preciso para sua avaliação, ou seja: a comparação da saída da codificação automatizada com a saída da codificação manual. Logo, o procedimento de validação ideal se divide em três: (1) que o ajuste inicial do modelo seja realizado no conjunto de treinamento; (2) depois que um modelo final é escolhido, que um segundo conjunto de documentos codificados manualmente – o conjunto de validação – seja usado para avaliar o desempenho do modelo; (3) que o modelo final seja então aplicado ao restante do acervo para completar a classificação. Para validação, pode ser aplicado o procedimento de validação cruzada – *cross-validation* (EFRON; GONG, 1983; HASTIE; TIBSHIRANI; FRIEDMAN, 2001). Nele, o conjunto de treinamento é particionado aleatoriamente em V ($v = 1, \ldots, V$) grupos, procedimento conhecido como *V-fold cross-validation*. Para cada grupo v, o modelo é treinado nos outros $V-1$ grupos, depois aplicado ao grupo V-*th* para avaliação de seu desempenho.

Categorias desconhecidas[29]

Vimos que com um conjunto definido de categorias os métodos de aprendizagem supervisionada auxiliam o pesquisador na tarefa de classificação do acervo. No entanto, não é difícil encontrar situações nas quais o conjunto de categorias não seja conhecido pelo pesquisador. Com base nos dados utilizados para aplicação deste artigo e analisados por Moreira (2016), pode, por exemplo, ser do interesse do pesquisador identificar quais tópicos são enfatizados pelos deputados federais nos discursos proferidos ao longo da 54ª legislatura (2011-2014)[30]. Uma vez que a atividade do representante político se debruça sobre inúmeras esferas da sociedade e da vida, predeterminar categorias temáticas de fala dos deputados federais pode limitar o conhecimento a ser obtido sobre o acervo. Para

28 Naive Bayes é apenas um exemplo de uma rica literatura que inclui outros modelos como: Random Forests (BREIMAN, 2001), Support Vector Machines (VENABLES; RIPLEY, 2002) e redes neurais (BISHOP, 1995).

29 Dedicamos esta seção aos algoritmos de *fully automated clustering* (FAC). Para detalhes sobre algoritmos de *computer assisted clustering* (CAC), ver Grimmer e King (2011).

30 Esse exemplo prático será apresentado mais adiante nesta seção.

enfrentar esse desafio, a seguir apresentamos métodos de aprendizado não supervisionado (*unsupervided learning methods*).

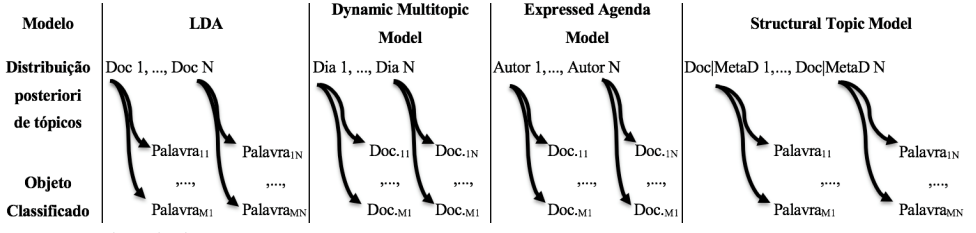
Métodos de aprendizado não supervisionado *(unsupervided learning methods)*

A classe de métodos de aprendizado não supervisionado revela características subjacentes ao texto sem a necessidade de imposição direta de categorias de interesse. Logo, ao invés de exigir que pesquisadores determinem previamente as categorias conhecidas, os métodos de aprendizado não supervisionado usam premissas e propriedades de modelagem dos textos para estimar um conjunto de categorias e, simultaneamente, atribuir a elas os documentos (ou partes de seu conteúdo). Ao informar para o algoritmo o número *k* de categorias nas quais os documentos devem ser alocados, há a oportunidade de se descobrir qual a composição de categorias que tenha melhor aderência ao conteúdo que está sendo analisado. Dada a incerteza do pesquisador sobre a performance do modelo em relação ao acervo que está sendo analisado, procedimentos de validação são essenciais.

Modelos de tópico (***topic models***)[31]. Para fins deste artigo, apresentaremos métodos de aprendizagem não supervisionada também conhecidos como modelos de tópico (*topic models*). Os modelos de tópicos possuem duas principais características. Em primeiro lugar, definem estatisticamente um tópico como função densidade de probabilidade sobre palavras. Para um tópico *k* (*k=1, …, K*), essa função de probabilidade é representada com um vetor *M´1,* q_k, em que q_{mk} descreve a probabilidade de o *k-ésimo* tópico usar a *m-ésima* palavra. Logo, para estimar um tópico, os modelos usam a ocorrência de palavras entre documentos e pressupõem, em sua grande maioria, o uso da DTM obtida através do pré-processamento apresentado anteriormente.

Em segundo lugar, os modelos de tópicos compartilham uma estrutura hierárquica básica. Como apresenta a Figura 1, para cada modelo temos qual o elemento sobre o qual os tópicos estão distribuídos. Em outras palavras, o elemento que terá uma distribuição de tópicos que somada resulta em um. Em seguida, na parte inferior da hierarquia, palavras ou documentos que são atribuídos a um único tópico. O *Expressed Agenda Model*, por exemplo, pressupõe que cada documento seja classificado em apenas um tópico.

Figura 1
Estrutura comum entre modelos de tópico



Fonte: Adaptado de Grimmer e Stewart (2013, p. 18)

31 Estes modelos diferem-se de modelos de clusterização como o *K-means*. O objetivo do algoritmo K-means é identificar uma partição de documentos que minimiza o quadrado da distância euclidiana entre *clusters*. Para mais detalhes, ver MacQueen (1967). Para conhecer outros modelos, recomendamos consultar Grimmer e Stewart (2013).

Latent Dirichlet Allocation – LDA (BLEI; NG; JORDAN, 2003)[32]. Como primeiro e mais difundido modelo de tópico, o LDA (BLEI; NG; JORDAN, 2003) assume que cada documento é constituído por diferentes (uma mistura de) tópicos. Para cada documento *i*, p_{ik} representa a proporção do documento dedicada ao tópico *k*, sendo $p_i = (p_{i1}, p_{i2}, \ldots, p_{iK})$ as proporções pelos tópicos. *A priori*, o LDA assume que a proporção de cada documento advém de uma distribuição Dirichlet, $p_i \sim$ Dirichlet (a), em que a representa o parâmetro da distribuição.

Em cada documento, as palavras são extraídas de acordo com a distribuição dos tópicos. Suponha que um documento contenha um total de palavras N_i, ($j=1$, …, N_i). Como Grimmer e Stewart (2013) apresentam, o LDA assume que um processo de duas etapas gera cada palavra. Para obter a *j-ésima* palavra no *i-ésimo* documento, o primeiro passo é extrair seu tópico, $t_{ij} \sim$ Multinomial (1, p_i). Condicional ao tópico atribuído, a palavra é extraída se a *j-ésima* palavra no *i-ésimo* documento está atribuída ao *k-ésimo* tópico, então extrai-se do tópico correspondente, $P_{ij} \sim$ Multinomial (1, q_k).

Foi com base na invovação promovida pelo LDA que a ciência política se apresentou mais uma vez na fronteira do conhecimento metodológico ao desenvolver outros modelos de tópico. Entre eles, o *Dynamic Multitopic Model* (QUINN et al., 2010), o *Expressed Agenda Model* (GRIMMER, 2010) e o *Structural Topic Model* (ROBERTS et al., 2013, 2014). Resguardadas as inovações promovidas por cada um dos modelos, conforme a Figura 1, todos apresentam uma estrutura semelhante ao LDA.

Dynamic Multitopic Model **(QUINN et al., 2010)**. Com base no caso dos discursos em plenário do Senado americano, o modelo supõe que a cada dia exista uma distribuição diferente de atenção a uma diversidade de tópicos. De forma análoga à atribuição de palavras a tópicos de determinado documento no LDA, cada discurso proferido no Senado americano é atribuído a um tópico. Por fim, uma *priori* dinâmica é usada para fazer inferências sobre a proporção de discursos de cada dia alocada a cada tópico.

Expressed Agenda Model **(GRIMMER, 2010)**[33]. Projetado para medir como os autores dividem sua atenção sobre temas, o modelo apresenta outra maneira de explorar a mesma estrutura inaugurada pelo LDA. Sua principal suposição é que cada autor divide sua atenção a um conjunto de tópicos. Assim, condicionado a tal distribuição de atenção dos atores, o tópico de cada documento é extraído. Mais adiante nesta seção será apresentado um exemplo de aplicação do *expressed agenda model* para o caso brasileiro.

Structural Topic Model **– STM (MONROE et al., 2015; ROBERTS et al., 2013, 2014)**[34]. Com base na mesma estrutura do LDA, o STM inova com duas características. Em primeiro lugar, permite que, ao nível dos documentos, seus metadados[35], entendidos

32 Para aplicação do LDA, recomendamos o pacote "topicmodels" da linguagem R (GRÜN; HORNIK, 2011). Para mais detalhes, ver: <https://bit.ly/2o0VjZb>. Acesso em: 3 jul. 2018.

33 Até o momento da elaboração deste artigo, o autor do modelo não disponibilizou o pacote em linguagem R prometido em Grimmer (2010).

34 Para estimação de tópicos com o STM, é possível utilizar gratuitamente o pacote "stm" na linguagem R (ROBERTS; STEWART; TINGLEY, 2018).

35 Metadados podem ser entendidos como dados sobre outros dados. Ou seja, informação que ajude a compreender características de um dado específico. No caso desta seção, nos referimos a dados que podem ajudar a compreender características dos documentos analisados como: data de publicação, autoria, georreferenciamento de onde foi criado, número de pessoas diferentes responsáveis pela produção do documento etc.

enquanto covariáveis, sejam incorporados ao modelo. Logo, informações como autoria e data de publicação podem ser incluídas para contribuir com a estimação dos tópicos. Em segundo lugar, ele permite estimar a correlação entre os tópicos de modo a identificar, por exemplo, quando dois tópicos podem ocorrer simultaneamente num documento. Tal informação pode auxiliar o pesquisador na identificação de temas que transcendam os tópicos.

Assim, com o objetivo de estimar a relação entre metadados e tópicos, no STM, estes são definidos como uma mistura sobre palavras em que cada palavra tem uma probabilidade de pertencer a um tópico. Logo, um documento é uma mistura sobre tópicos, o que significa que um único documento pode ser composto de vários tópicos do mesmo modo, como no caso do LDA.

O uso de modelos de aprendizagem não supervisionada

Como apontam Grimmer e Stewart (2013), todos os métodos de aprendizagem não supervisionada pressupõem duas etapas básicas após os procedimentos de pré-processamento apresentados: (1) a definição do número de categorias; e (2) validação.

1. **Definindo o número de categorias**: Determinar o número de categorias é uma das questões mais complexas no aprendizado não supervisionado[36]. Ao mensurarem quão bem modelos se ajustam aos dados, medidas de ajuste estatístico tornam-se inúteis diante da brusca redução de informação que o uso de métodos não supervisionados para análise de conteúdo pressupõem após o pré-processamento dos dados. Os textos pré-processados representam uma simplificação substancial dos documentos, sendo o objetivo do uso de métodos não supervisionados a revelação de informações substantivamente relevantes. Logo, em vez de ajuste estatístico, a seleção de modelos deve ser tratada como um problema de mensuração substantiva. Em linha com Grimmer e Stewart (Ibidem), acreditamos que a abordagem *mixed-method* fornecida por Quinn et al. (2010) é adequada. Nela, os modelos candidatos são ajustados variando-se o número de categorias para, em seguida, ser realizada uma avaliação manual e qualitativa para seleção do modelo final com base na qualidade das categorias obtidas.

2. **Validação**: Se, de um lado, o uso da aprendizagem não supervisionada reduz os custos de análise manual do acervo por parte do pesquisador antes do uso do modelo, de outro, a carga de trabalho para a validação de seus resultados é imensa. É a validação extensiva das categorias estimadas e dos documentos classificados que permite a realização de inferências concretas sobre o acervo. Veremos, no exemplo prático a seguir, abordagens para validação dos resultados do *expressed agenda model* aplicado ao caso brasileiro[37].

Aplicação para o caso brasileiro – *expressed agenda model*

O *expressed agenda model* foi utilizado por Moreira (2016) com o objetivo de identificar as ênfases temáticas proferidas pelos deputados

36 Alguns métodos tentam eliminar essa decisão e estimar o número de categorias (FREY; DUECK, 2007), mas estudos mostram que o número estimado é fortemente dependente do modelo (WALLACH et al., 2010). Também não é conveniente verificar medidas de ajuste do modelo (CHANG et al., 2009).

37 É importante apontar que as validações realizadas aqui são apenas um subconjunto do que pode ser empregado.

Federais em seus discursos no Pequeno Expediente ao longo da 54ª legislatura (2011-2014).

1. **Definição do número de tópicos**:
   Uma vez realizado o pré-processamento dos dados, de acordo com os procedimentos apresentados anteriormente[38], o primeiro desafio imposto pelo modelo é a definição do número *k* de tópicos presente no *corpus*, ou seja, a quantidade de temas abordados em cada uma das legislaturas analisadas.
   Para a definição do número *k* de tópicos, utilizaram-se duas estratégias: (1) o uso de um modelo não paramétrico para clusterização de texto baseado no *Dirichlet process prior* (BLEI; LAFFERTY, 2006; GRIMMER, 2010); e (2) a estimação de diferentes modelos.
   O modelo não paramético resultou em 36 tópicos contidos no acervo. No entanto, dadas as ponderações já apresentadas, esse resultado não foi considerado de forma definitiva e a estipulação da quantidade *k* de tópicos contou com uma avaliação qualitativa do resultado de diferentes modelos para cada legislatura. A avaliação qualitativa permite que o valor *k* seja definido pela coesão substantiva identificada pelo analista através da análise dos *stems* mais associados a cada tópico em diferentes modelos e da leitura de amostras aleatórias de documentos presentes nas categorias estimadas por cada modelo.
   Foram estimados e analisados os resultados de modelos que variaram de 5 a 80 tópicos. Por um lado, comparados entre si, quanto menor o número de tópicos de um modelo, maior é a diversidade de discursos classificados em cada um, resultando em categorias muito genéricas. Por outro, quanto maior o número de tópicos do modelo, maior é a quantidade de tópicos tratando sobre o mesmo tema. Por essa razão, com o auxílio da evidência estatística do modelo não paramétrico, foi possível analisar de forma qualitativa os resultados dos 75 modelos estimados para a definição de um resultado de 39 tópicos para a 54ª legislatura. O resultado do modelo com as 39 categorias podem ser encontrados na Tabela 3. Na primeira coluna apresenta-se o rótulo dado a cada tópico após a leitura de uma amostra de ao menos dez discursos aleatoriamente selecionados de cada um deles. Na segunda coluna, é possível verificar até o quinto *stem* com maior informação mútua[39] em cada tópico. Na terceira, é apresentado o percentual de documentos do *corpus* classificado em cada um dos tópicos.

2. **Validação**:
   Para que o resultado apresentado fosse considerado consistente, distintas formas de validação foram adotadas para averiguar se os resultados são substantivamente relevantes. Os tópicos foram validados por meio de quatro procedimentos: (1) dado que a matéria-prima para a análise dos tópicos são os *stems* das palavras contidas nos discursos, verificou-se quais os dez *stems* mais associados a cada um pelo do cálculo de sua informação mútua (GRIMMER, 2010, 2013); (2) foram lidos ao menos dez discursos aleatoriamente selecionados para rotulação de cada tópico; (3) sendo cada discurso pertencente a um tópico, é

38 Além da DTM desenvolvida e já apresentada, o *expressed agenda model* também solicita uma matriz adicional que informe o índice do primeiro e do último discurso de cada um dos deputados na DTM, pressupondo que as linhas desta estejam organizadas por orador com mais de um discurso.

39 A informação mútua entre um tópico e um *stem* mede a quantidade de informação que este provê sobre a possibilidade de um tópico gerar um documento aleatoriamente selecionado no *corpus*.

analisada sua pertinência temporal conforme a frequência dos tópicos ao longo da legislatura; (4) é qualitativamente analisada a dedicação de parlamentares selecionados a tópicos específicos, de modo que seja possível identificar se há coerência entre a classificação temática dos discuros e perfis parlamentares amplamente conhecidos e difundidos na sociedade e na ciência política brasileira.

Tabela 3
Temas dos discursos proferidos na legislatura 54

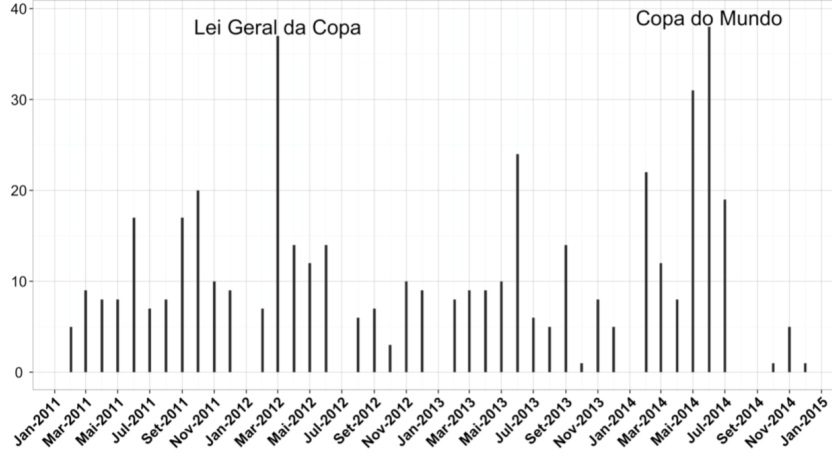
Rótulo	Stems	%
Votação	vot, mat, votaca, sim, projet	6,1
Datas comemorativas	dia, registr, jos, vid, jornal	6,0
Trabalho	trabalh, direit, empreg, dia, projet	4,9
Questões municipais	municipi, prefeit, nov, cidad, recurs	4,9
Protestos e corrupção	pov, polit, trabalh, precis, dilm	4,8
Projetos de lei	projet, lei, aprov, trabalh, comissa	4,5
Questões regionais	cidad, prefeit, municipi, regia, trabalh	4,0
Educação	educaca, escol, professor, ensin, alun	3,9
Economia	econom, setor, polit, ano, invest	3,8
Sistema político	polit, reform, vot, eleitoral, campanh	3,4
Agentes de saúde	saud, agent, atend, recurs, trabalh	3,4
Medida Provisória	med, provisor, vot, emend, lei	3,1
Questões regimentais	comissa, lid, titul, art, suplent	3,1
Segurança pública	polic, seguranc, milit, trabalh, crim	2,9
Empresas	empres, petrobr, milho, trabalh, servic	2,8
Transporte	transport, rodov, port, sant, sul	2,6
Agropecuária	agricultur, produtor, produca, produt, famili	2,6
Sistema de Justiça	justic, tribunal, suprem, defensor, process	2,5
Recursos e investimentos	milho, mil, invest, municipi, recurs	2,3
Direitos Humanos	direit, human, comissa, pesso, trabalh	2,3
Gênero	mulh, violenc, polit, trabalh, dia	2,2
Pessoa com deficiência	pesso, deficienc, trabalh, direit, vid	2,1
Crédito financeiro	banc, nord, econom, jur, trabalh	1,9
Meio ambiente	ambient, mei, ambiental, are, codig	1,8
Programas federais	program, famil, bols, social, rend	1,7
Educação superior	univers, educaca, ensin, curs, estud	1,7
Questão hídrica	agu, sec, nord, regia, municipi	1,6
Saúde	saud, hospital, atend, canc, doenc	1,5
Esporte	cop, mund, esport, futebol, estadi	1,4
Medicina	medic, saud, trabalh, program, atend	1,4
Criança e adolescente	crianc, adolescent, violenc, direit, sexual	1,3
Questão indígena	indigen, terr, indi, pov, direit	1,2
Servidor público	servidor, trabalh, servic, pec, direit	1,2
Energia	energ, eletr, consumidor, tarif, cont	1,1
Amazônia	amazon, manaus, zon, franc, regia	1,1
Drogas e violência	drog, crack, pesso, usuari, saud	0,8
Idoso e Previdência	idos, aposent, pesso, trabalh, projet	0,8
Questão racial	negr, dia, polit, populaca, racial	0,7
Estatuto da Juventude	juventud, jovens, polit, direit, trabalh	0,5

Os resultados de dois dos quatro procedimentos de validação: a leitura atenta de uma amostra aleatória dos discursos presentes em cada tópico para rotulagem adequada e a análise de raízes com a maior informação mútua em cada um dos tópicos podem ser encontrados na Tabela 3. Avançamos a seguir, portanto, no sentido de avaliar a pertinência temporal dos tópicos e a ênfase temática esperada de alguns parlamentares de perfil amplamente conhecido pela sociedade brasileira e a ciência política nacional.

Pertinência temporal de tópicos selecionados. É possível averiguar a validade de um tópico por meio de sua pertinência temporal. Verifica-se se os discursos relacionados a cada tópico estão em acordo com debates desenvolvidos ao longo da legislatura e, em especial, se condizem com a ocorrência de eventos exógenos à instância de produção do discurso.

A paixão brasileira pelo futebol somada à escolha do país para sediar a Copa do Mundo FIFA de 2014 e a Copa as Confederações FIFA de 2013 produziu efeitos sobre toda atividade política nacional. Além da previsão de investimentos em transporte e infraestrutura, a atividade legislativa federal contou com a necessidade de produzir ordenamento jurídico específico para a realização do evento. Ainda assim, a atividade parlamentar a seu respeito não se restringiu à aprovação de dispositivos legais para sua execução. Conforme a Figura 2, apresenta, em 2012, para aprovação da Lei Geral da Copa (Lei Ordinária 12.663/12), mas, sobretudo, em 2014 – ano do evento – as falas proferidas no Pequeno Expediente trataram de destacar a temática do esporte, a relevância da Copa do Mundo em território nacional e seus efeitos políticos.

Figura 2
Pronunciamentos classificados na categoria Esporte ao longo da 54ª Legislatura

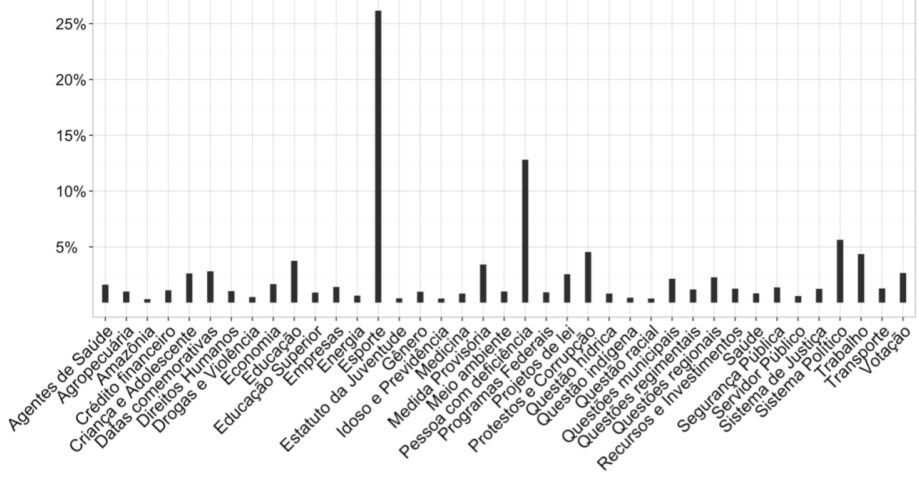


Ênfase temática de deputados federais selecionados. A principal contribuição do *expressed agenda model*, em comparação com as demais metodologias utlizadas na classificação de conteúdo de forma não supervisionada, é sua estrutura hierárquica

que permite identificar a ênfase de temática de autores[40]. Por tal razão, como última estratégia de validação dos resultados obtidos pelo modelo para as legislaturas analisadas, foram averiguadas as ênfases temáticas de deputados federais cujo perfil é amplamente difundido e conhecido na ciência política nacional.

O então deputado federal Romário, do Partido Socialista Brasileiro (PSB-RJ), era conhecido em todo o país em função de sua atuação como jogador de futebol. Durante seu mandato como deputado federal, o ex-jogador elegeu dois grandes temas para sua atuação: o esporte e as pessoas com deficiência. Foi autor de 21 projetos de lei, entre eles o PL4129/12, que institui a Semana Olímpica nas escolas públicas, e o PL 7916/14, que dispõe sobre a contratação de Apaes e Pestalozzis – entidades sem fins lucrativos – como prestadoras de serviços do Poder Público, com especialização em educação especial. Ao longo da 54ª legislatura, o deputado federal Romário realizou 26 pronunciamentos no Pequeno Expediente, e a Figura 3 apresenta a ênfase temática de seus discursos estimada em conjunto com a dos demais oradores dessa legislatura.

Figura 3
Ênfase temática dos pronunciamentos realizados pelo deputado federal Romário (PSB-RJ) na 54ª legislatura



40 No primeiro *Hackathon* da Câmara dos Deputados, realizado no mês de outubro de 2013, foi desenvolvida uma aplicação *web* que permite ao usuário saber qual tema cada deputado federal mais enfatizou em seus discursos e, ainda, qual a importância daquele tema para determinado parlamentar quando comparado aos demais. O projeto Retórica Parlamentar, assim denominado, foi desenvolvido por Davi Moreira, pelo doutor em Ciência Política, Manoel Galdino, e por Luis Carli, então doutorando em Visualização de Dados pela Faculdade de Arquitetura e Urbanismo, ambos pela Universidade de São Paulo. A equipe do Laboratorio Hacker e do Departamento de Taquigrafia, Revisão e Redação da Câmara dos Deputados deu continuidade ao projeto e ele pode ser visto em: <https://bit.ly/2i9kOEu> e <https://bit.ly/2OYDOnJ>. Acesso em: 7 jun. 2018.

Como se pode constatar, as falas proferidas pelo deputado federal Romário possuem nítida relação com sua atuação parlamentar.

Em conjunto com os outros dois procedimentos de validação, a ênfase temática dos deputados federais identificada pelo *expressed agenda model* e ilustrada com a apresentação dos resultados para o tema do esporte ao longo da legislatura e os pronunciamentos do deputado federal Romário indica que o modelo estimado foi satisfatório.

Atribuir textos a categorias é o uso mais comum de métodos de análise de conteúdo na ciências sociais. Vimos nesta seção que os métodos automatizados podem mitigar o custo da atribuição de documentos às categorias e amplificar a quantidade de classificação que os humanos executam de forma manual. Graças ao desenvolvimento tecnológico e o baixo custo dos equipamentos informáticos domésticos, tarefas antes impossíveis e questões de pesquisas antes restritas a um grande montante de recursos humanos e financeiros podem agora ser acessadas com sucesso.

Métodos de escalonamento

A utilização da análise quantitativa de textos para a extração de posições políticas/ideológicas de partidos, políticos e eleitores é uma área extremamente promissora na ciência política. O teste de modelos de competição partidária, por exemplo, depende do conhecimento das posições dos principais atores envolvidos no jogo político.

Se, por um lado, existem diversos modelos que estimam os "pontos ideais" desses atores a partir de votações nominais (CLINTON; JACKMAN; RIVERS, 2004; POOLE; ROSENTHAL, 2007) e pesquisas de opinião (ALDRICH; MCKELVEY, 1973; POWER; ZUCCO, 2009), por outro, a utilização e disponibilidade desses dados são limitadas. Votações nominais representam apenas uma pequena amostra daquilo que é discutido e decidido no interior de um parlamento (CARRUBBA et al., 2006). Nesse sentido, as posições políticas extraídas a partir delas são mais um produto do que a causa do processo político sob investigação. Além disso, apenas membros do corpo legislativo participam de votações nominais. Importantes atores como presidentes, ministros e eleitores não podem ter suas posições políticas estimadas. Pesquisas de opinião também são problemáticas. Elas possuem problemas de comparabilidade interpessoal (BRADY, 1985; KING et al., 2004) e são limitadas temporalmente. Não podemos voltar ao passado e conduzir pesquisas com atores políticos de outras épocas.

O emprego de textos como fonte primária para a estimação de posições políticas não é apenas uma alternativa devido à limitação de outras fontes. Em realidade, a utilização de palavras – sejam elas escritas ou faladas – é a forma mais básica de como o conflito político é expressado (GRIMMER; STEWART, 2013; MONROE; SCHRODT, 2009). Antes de qualquer votação nominal, parlamentares discutem a matéria em plenário. Partidos apresentam suas ideias em programas. Cidadãos discutem política nas redes sociais. Em resumo, é por meio de palavras que os indivíduos podem expressar suas preferências políticas e é através de seu uso estratégico que a política se materializa.

Uma das iniciativas pioneiras no uso de textos para inferir posições políticas foi o *Comparative Manifesto Project* (BUDGE et al., 2001). Esse projeto utiliza técnicas de análise de conteúdo para codificar os programas de mais de mil partidos políticos desde 1945 até hoje em mais de cinquenta países. Essa codificação é empreendida manualmente por uma equipe formada por um grande número de indivíduos treinados. Como é possível imaginar, esse projeto envolve grande quantidade de recursos financeiros que dificilmente estão disponíveis para a maior parte dos pesquisadores.

No entanto, o emprego de técnicas que dependem quase exclusivamente de recursos computacionais tem tornado a tarefa de estimar posições políticas a partir de textos acessíveis a qualquer pesquisador[41]. Atualmente, as duas técnicas mais populares são o *Wordscores* (LAVER; BENOIT; GARRY, 2003) e o *Wordfish* (SLAPIN; PROKSCH, 2008).

Wordscores

O *Wordscores* é um algoritmo supervisionado para estimar posições políticas (LAVER; BENOIT; GARRY, 2003). Nessa família de algoritmos, são apresentados ao computador alguns dados de entrada e as saídas esperadas. Chamamos esse conjunto de entradas e saídas de *training set*. A partir desse conjunto de informações o algoritmo "aprende" a classificar novos documentos. Esses formam o conjunto do *test set*. Assim, no *Wordscores* temos dois conjuntos de textos. O primeiro é formado pelos textos de referência (*training set*). Nele temos documentos cujas posições políticas são definidas, *a priori*, em uma dimensão conhecida pelo analista. Na grande maioria dos casos, essa dimensão está associada à escala ideológica esquerda-direita. O segundo conjunto é formado pelos textos cujas posições políticas são desconhecidas (*test set*), mas gostaríamos de conhecer. Observamos apenas o número de vezes que cada palavra aparece em cada texto. De modo intuitivo, o algoritmo classifica os documentos do *test set* em um contínuo entre os documentos de referência a partir da similaridade da frequência relativa de palavras.

A primeira etapa para implementar o *Wordscores* é escolher os textos de referência e definir quais são suas posições políticas. Essa etapa é fundamental e envolve o conhecimento substantivo do contexto no qual os dados são gerados. Os autores fornecem algumas diretrizes para a escolha desses textos. Em primeiro lugar, é importante que os textos de referência utilizem o mesmo léxico que os textos virgens. Por exemplo, se queremos classificar discursos de parlamentares, é recomendável que os textos de referência também sejam discursos de parlamentares. Textos de diferente natureza, como programas de partidos, utilizam um conjunto de palavras muito diferente do empregado em discursos, portanto, trazem poucas informações. A segunda orientação é selecionar textos que cubram todo o espectro ideológico. Idealmente, é recomendável escolher textos que ocupem os extremos da escala, além da posição central. Por fim, a última recomendação é que os textos de referência possuam um conjunto diversificado de palavras. Assim, devemos evitar o uso de documentos curtos como textos de referência, porque os do *test set* serão analisados no contexto do universo de palavras dos textos de referência (*training set*).

Após a escolha dos textos devemos atribuir valores às suas posições políticas. Por exemplo, se nossos textos de referência são discursos de um partido de esquerda, um de direita e um de centro, podemos atribuir os valores de –1, 1 e 0 para cada partido, respectivamente. Outra possibilidade é utilizar as posições estimadas a partir de outros dados, como pesquisas de opinião. Com isso, completamos nosso *training set*.

A segunda etapa é gerar *scores* para as palavras dos textos de referência. Esse *score* é a média da posição política atribuída *a priori* (–1, 1 ou 0, no exemplo acima) ponderada pela probabilidade de observar um documento,

41 É possível aplicar o *Wordscores* e o *Wordfish* utilizando o pacote "austin" do R (LOWE, 2015). Para mais detalhes, ver: <https://bit.ly/2MsqgnN>. Acesso em: 7 jul. 2018.

dado que estamos analisando uma palavra em particular. Isto é, $S_{pd} = \sum_r (P(r|p) \times A_{pd})$, em que $P(r|p) = \frac{P(r|p)\,P(r)}{P(p)}$ é a probabilidade de observar um documento de referência *r*, dado que estamos observando a palavra *p*; e A_{pd} é a posição política da palavra *p* na dimensão política *d*.

Por exemplo, suponha que os textos de esquerda, centro e direita tenham mil palavras cada. Suponha também que a palavra "copa" apareça dez vezes no texto de esquerda, vinte no de direita e trinta no de esquerda. A probabilidade de observarmos o documento de esquerda, dado que observamos a palavra "copa", é de 17% (0,01 / 0,06). Para o documento de direita temos 33% (0,02 / 0,06) e para o documento de centro temos 50% (0,03 / 0,06).

Assim, o *score* da palavra "copa" será: 0,17 (–1) + 0,5 (0) + 0,33 (1) = 0,16. Isto é, dada a frequência relativa de palavras nos textos de referência, se soubéssemos apenas que a palavra "copa" aparecia em um documento qualquer, esperaríamos que sua posição política seria de 0,16.

Com o *score* para todas as palavras no universo dos textos de referência podemos estimar a posição política dos textos virgens. Essa terceira etapa nada mais é do que calcular o *score* médio das palavras ponderando pela frequência relativa de palavras em cada documento virgem. Ou seja, $S_{vd} = \sum_p P(p|v) \times S_{pd})$, em que $P(p|v)$ é a probabilidade de observar a palavra *p* no documento virgem *v*.

Para colocar os textos do *test set* na mesma escala que os textos de referência, podemos aplicar a seguinte transformação: $S^*_{vd} - (S_{vd} - S_{\overline{vd}})\left(\frac{SD_{rd}}{SD_{vd}}\right) + S_{\overline{vd}}$, em que $S_{\overline{vd}}$ é o *score* médio dos textos do *test set* e SD_{rd} e SD_{vd} são os desvios-padrão amostrais dos textos de referência e virgens, respectivamente.

Podemos também calcular medidas de incerteza para os *scores*, como a variância: $V_{vd} = \sum_p P(r|p)\ (S_{pd} - S_{vd})^2$. Com isso é possível implementar testes como diferenças de médias e avaliar se as diferenças entre as posições políticas estimadas para dois documentos são estatisticamente significativas.

Embora o *Wordscores* constitua um grande avanço na análise quantitativa de textos, ele não é livre de problemas. O principal é o fato de ele depender fortemente da escolha dos textos de referência (*training set*). Em situações extremas é possível que, com a escolha diferente de textos de referência, um mesmo pesquisador encontre resultados diferentes para um mesmo conjunto de dados. A segunda limitação, como apontado por Lowe (2008), é a possibilidade de as diferenças entre os textos estarem mais relacionadas com o estilo linguístico do autor do que com as posições políticas. Como todas as palavras adicionam a mesma quantidade de informação sobre o documento, temos que palavras politicamente relevantes em um contexto sejam igualmente ponderadas a palavras pouco informativas.

Wordfish

A segunda técnica mais popular para estimar posições políticas a partir de textos é o *Wordfish* (SLAPIN; PROKSCH, 2008). Ao contrário do *Wordscores*, esse é um algoritmo não supervisionado pois não depende da escolha de textos de referência, ou seja, da construção de um *training set*. Dessa forma, não há a limitação de diferentes pesquisadores chegarem a resultados diferentes a partir do mesmo conjunto de dados. Outra vantagem em relação ao *Wordscores* é o fato de ele não atribuir o mesmo peso para todas as palavras. O *Wordfish* estima a importância das palavras para discriminar as posições políticas. Assim, palavras

politicamente relevantes em uma dimensão têm peso maior na tarefa de localizar os documentos no espectro político.

O *Wordfish* é baseado em modelos da TRI, tal como aqueles utilizados para estimação de pontos ideais a partir de votações nominais (CLINTON; JACKMAN; RIVERS, 2004). Mas, em vez de utilizar votos dados a projetos, ele opera com a frequência relativa de palavras. Nesse sentido, acredita-se que o uso relativo das palavras forneça informações relevantes sobre as posições políticas dos atores.

O modelo assume uma distribuição de Poisson para a contagem de palavras. Isto é, cada palavra *j* de um documento *i*, Y_{ij}, é gerada a partir de uma distribuição de Poisson com parâmetro $\lambda_{ij} > 0$, $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$. A escolha dessa distribuição foi determinada por sua simplicidade. Ela possui apenas um parâmetro (λ) que, ao mesmo tempo, representa a média e a variância. Outra coisa importante de se notar é a existência um pressuposto de que a probabilidade de observamos uma palavra em um documento é independente da posição das outras palavras no mesmo documento. Embora esse pressuposto seja falso, ele é frequentemente utilizado na análise quantitativa de textos.

A forma funcional do modelo é dada por $\lambda_{ij} = \exp(\alpha_i + \psi_j + \beta_j \times \omega_i)$, em que α é um conjunto de efeitos fixos por documento. A inclusão desses parâmetros ocorre porque alguns documentos são mais longos do que outros; ψ é um conjunto de efeitos fixos por palavra. Sua inclusão é importante porque algumas palavras são mais frequentes do que outras. Já β é o parâmetro de discriminação, uma estimativa de quanto a palavra *j* é importante para distinguir as posições políticas. Por fim, ω é a estimativa da posição política. Esse é o principal parâmetro de interesse.

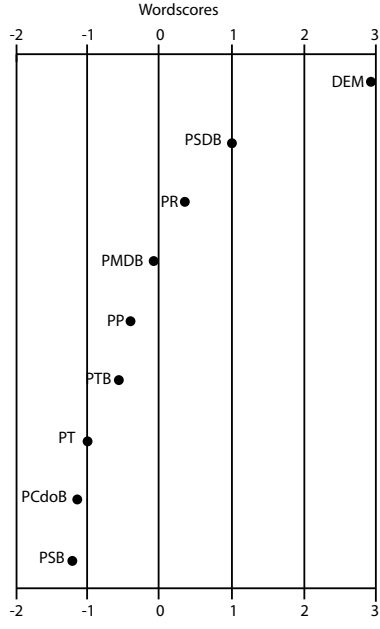
O principal interesse dos autores foi estimar as posições políticas dos partidos a partir de seus programas. Em geral, os programas apresentam as posições dos partidos em um grande leque de assuntos. Portanto, a expectativa é a de que, ao aplicar o algoritmo, o resultado deve ser uma escala ideológica. Se o analista está interessado em extrair as posições dos partidos no que diz respeito a um tema específico, a primeira etapa deve ser selecionar os textos sobre esse tema.

A limitação desse modelo é o fato de ele necessitar que os documentos cubram uma grande quantidade de temas para extrair as posições ideológicas. Assim, se os parlamentares focarem seus discursos em determinadas áreas temáticas, provavelmente não teremos resultados consistentes ao aplicarmos o *Wordfish*. Como a variação no uso das palavras não será determinada pelas preferências políticas, mas pelos tópicos, a diferença nas posições estimadas também refletirá essa diferença (LAUDERDALE; HERZOG, 2016).

Aplicação para o caso brasileiro – Wordscores

Nesta seção aplicaremos o *Wordscores* aos discursos dos deputados durante o ano de 2011, utilizando a DTM obtida pelo pré-processamento apresentado anteriormente. Para essa aplicação, agrupamos os discursos por partido. Assim, todos os documentos de parlamentares de um mesmo partido compõem um único documento. O objetivo desse procedimento é tentar garantir que os textos tratem de uma grande quantidade de temas. Como individualmente cada parlamentar tende a focar seus esforços em determinadas áreas temáticas, ao agruparmos os discursos, temos uma variedade maior. Selecionamos também apenas os partidos com pelo menos 300 discursos. Ao todo temos nove partidos: Partido dos Trabalhadores (PT), Partido Trabalhista Brasileiro (PTB), Partido Progressista (PP), Movimento Democrático Brasileiro (PMDB), Partido da República (PR), Partido da Social Democracia Brasileira (PSDB), PSB, PCdoB e DEM.

Gráfico 1
Posições políticas dos partidos políticos brasileiros (2011)



Para essa aplicação, selecionamos como textos de referência os discursos do PT e do PSDB. Essa escolha parece ser adequada, já que no período atual os dois partidos têm polarizado as disputas políticas. Como é possível observar pelo Gráfico 1, o *Wordscores* classificou os principais partidos políticos em uma escala ideológica. Do lado esquerdo (negativo) temos PSB, PCdoB e PT, do lado direito (positivo) temos DEM, PSDB e PR, e no centro temos PTB, PP e PMDB.

Manter a metodologia das ciências sociais brasileiras na fronteira do conhecimento humano para que técnicas contemporâneas possam ser utilizadas para questões de pesquisa sobre o presente, o passado e o futuro é um dos objetivos desse artigo. Por isso, para além dos métodos apresentados até aqui, na próxima seção indicamos desenvolvimentos e avanços recentes que apontam para onde caminha o uso do texto como dado nas ciências sociais.

Desenvolvimentos e aplicações mais recentes

O volume de conteúdo publicado cresce rapidamente. Porém a forma de publicação não se restringe mais ao mundo físico como jornais, livros, panfletos etc., mas também – e especialmente – atinge o mundo virtual em mídias sociais e outros meios baseados na internet que reduzem custos de publicação e expressão social por meio do uso do texto como dado, ampliando as fontes de acervo para análises em ciências sociais (BARBERÁ, 2012). Não há sinal de que essa tendência mude e, como aponta Simon Jackman (MAGALHÃES et al., 2013), ao ser perguntado sobre as perspectivas em metodológica em ciência política:

> "Eu acredito que tratar textos como dados é algo grande. Isso significa que tudo é dado. Há dados em todos os lugares. Textos vem juntos com análise dimensional. O que é uma lei? Como leis se

> relacionam entre si? O que é um discurso político? Qual é a retórica de um político? Como você faria um trabalho quantitativo sobre isso? Eu acredito que isso vai ser grande.""

As ciências sociais se mantêm nessa fronteira do conhecimento e avanços não param de ocorrer. A seguir, apresentamos uma visão geral da tendência no campo de análise do texto como dado.

Olhar o passado com as lentes do presente

Sem dúvida alguma uma das áreas mais promissoras de uso e aplicação dos métodos de análise automatizada de conteúdo é a análise de processos históricos[42]. Como o título dessa subseção apresenta, há, hoje, a oportunidade de abordar questões de pesquisa sobre o passado com a possibilidade de aplicação do ferramental metodológico contemporâneo. A seguir, veremos dois exemplos dessa aplicação.

Medindo novidade, transitoriedade e ressonância

Em trabalho recente, Barron et al. (2018) analisaram mais de 40 mil discursos durante os debates no interior do parlamento da Revolução Francesa. Os autores traçaram a criação, destruição e propagação de ideias a partir dos padrões de uso das palavras. Os resultados corroboram evidências qualitativas de que parlamentares da esquerda traziam inovações aos debates, ao passo que parlamentares de direita agiam de modo a preservar os padrões anteriores. Esse processo foi dirigido em parte por algumas importantes figuras da época, como Robespierre e Pétion de Villeneuve. Esses políticos radicais não apenas introduziram novas estratégias discursivas e padrões de uso de palavras em taxas maiores que os outros parlamentares, mas também o fizeram de modo que suas ideias se mantiveram ao longo do tempo.

Discursos do presente, que são muito diferentes de discursos proferidos no passado, indicam que novas ideias são trazidas ao debate político. Por outro lado, grandes desvios dos padrões discursivos de hoje em relação a padrões do futuro indicam a natureza transitória dessas ideias. Discursos que trazem novos padrões ao debate e levam a discussão para uma nova direção são aqueles que têm ressonância. Ou seja, o desequilíbrio entre alta novidade e baixa transitoriedade é o que caracteriza a ressonância de uma ideia.

A partir de uma definição bayesiana de surpresa (ITTI; BALDI, 2006), uma medida de divergência entre as distribuições a priori e a posteriori é utilizada para análise. Em outras palavras, dizemos que novos dados observados (D) trazem uma surpresa quando a distribuição, *a posteriori* resultante da observação desses dados, é significantemente diferente da distribuição *a priori*[43]. Assim, desenvolvem medidas de novidade, transitoriedade e ressonância.

Os autores classificaram os discursos em tópicos (K = 100) utilizando o LDA e analisaram como a combinação desses tópicos se desviam de discursos anteriores (novidade) e posteriores (transitoriedade). Grandes desvios comparando padrões de hoje com os do passado indicam que os tópicos são novos (novidade). Grandes desvios comparando

42 Uma iniciativa interessante com este enfoque pode ser encontrada em: http://www.history-lab.org/. Acessado em 09 de julho de 2018.

43 Essa ideia é operacionalizada por meio da uma medida de entropia relativa ou divergência de Kullback-Leibler (BARRON et al., 2018).

padrões de hoje com os do futuro indicam que os padrões não foram retidos (transitoriedade). Por sua vez, discursos que trazem novos padrões ao debate e levam a discussão a uma nova direção são aqueles que possuem ressonância.

Medindo complexidade linguística

Spirling (2015) estimou o impacto da expansão do sufrágio sobre a complexidade linguística dos discursos dos membros do parlamento inglês na era vitoriana. Com a *Second Reform Act*, de 1867, houve a inclusão de grande contingente de novos eleitores. Além disso, houve redução dos requerimentos necessários, em termos de propriedade, para participar do processo eleitoral. Com isso, o resultado da Reforma foi a extensão do sufrágio para uma grande massa de trabalhadores urbanos pouco educados, quando não analfabetos. Em reação a esse novo cenário eleitoral, os parlamentares reduziram a complexidade de seus discursos para atingir esses eleitores.

Para medir essa mudança de graus na compreensão dos discursos o autor lançou mão de métricas comumente utilizadas em pesquisas na área de educação. Essas medidas levam em consideração a relação entre o número de sílabas e o número de palavras em um documento (FLESCH, 1948).

Os autores analisaram mais de 650 mil discursos proferidos por mais de 3.500 parlamentares entre 1832 e 1915. Os resultados apontam que, logo após a Reforma, os parlamentarem alteraram seus discursos de modo a torná-los de mais fácil compreensão para o novo eleitor mediano, alguém mais pobre e menos educado do que o eleitor mediado anterior.

Big Data, o texto como dado e causalidade

Enquanto lê esta sentença, tente imaginar quantas buscas no Google foram realizadas e registradas ao redor do planeta. Na era do *Big Data* a todo instante volumes massivos de dados são produzidos (LAZER et al., 2009) e as ciências sociais, com o uso da análise automatizada de conteúdo, têm papel crucial na transformação desses dados em informação e conhecimento. O *Big Data* oferece a oportunidade de produzir conhecimento a partir de um volume de dados inviável há apenas alguns anos.

Diante desse desafio, para além de conhecimentos computacionais, analisar habilmente a massa de conteúdo que tem sido intensamente produzida também exige uma avaliação rigorosa (PATTY; PENN, 2015), um desenho de pesquisa cuidadoso e a implementação criativa de técnicas estatísticas (GRIMMER, 2015). Como Grimmer (Idem) aponta, para que a análise do *Big Data* realmente forneça respostas aos problemas da sociedade, deve-se reconhecer que ela é tanto sobre ciência social quanto sobre ciência da computação.

O papel da descrição. Vimos neste artigo que o ganho de escala obtido por meio do uso dos métodos apresentados se coloca como ferramenta promissora diante dos desafios do *Big Data*. Assim, as oportunidades para inferências descritivas são abundantes em *Big Data*, têm potencial enorme para reorientar teorias e questões de inferência causal estabelecidas nas ciências sociais[44]. Conforme Grimmer (2015), a análise de coleções de conteúdo, de acervos de notícias, de postagens em mídias sociais, entre outros, podem contribuir para respostas a importantes questões das ciências

44 No âmbito da ciência política, o projeto *VoteView* talvez seja o melhor exemplo de como projetos puramente descritivos afetam teorias e as questões de inferência causal da literatura (MCCARTY; POOLE; ROSENTHAL, 2006; POOLE; ROSENTHAL, 2007).

sociais como: o conhecimento sobre a agenda da grande mídia, o quanto a política está presente nas redes sociais, o posicionamento do público a respeito de um tema específico, ou reações diante de propostas de campanha, bem como as próprias campanhas eleitorais realizadas em âmbito virtual. A chance de inferência descritiva de conteúdo em tamanha escala cria para as ciências sociais a oportunidade de fazerem perguntas causais e criarem teorias anteriormente impossíveis (MONROE et al., 2015).

Causalidade. Diante dos desafios do *Big Data*, ser capaz de aplicar técnicas estatísticas a conjuntos de dados massivos para a obtenção de informações descritivas do conteúdo publicado é apenas o princípio da contribuição que as ciências sociais podem dar a essa revolução computacional. Somada a esse princípio, está a oportunidade do uso de metodologias de análise automatizada de conteúdo para a identificação de efeitos causais baseados no texto como dado. Combinados a experimentos, o uso do texto como dado pode ser útil para a descoberta de medidas que testem teorias de interesse das ciências sociais a partir de grandes coleções de texto. Trabalhos que conectam a literatura do texto como dado (LAVER; BENOIT; GARRY, 2003; PENNEBAKER; MEHL; NIEDERHOFFER, 2003; QUINN et al., 2010), com a crescente literatura sobre inferência causal nas ciências sociais (HERNAN; ROBINS, 2018; IMBENS; RUBIN, 2015; PEARL, 2009) nesse sentido, já têm sido publicados.

Egami et. al. (2018) apresentam uma estrutura conceitual para fazer inferências causais com tratamentos obtidos a partir da análise automatizada de conteúdo, fornecendo uma estrutura rigorosa para inferências causais baseadas em texto. Em linha semelhante, Fong e Grimmer (2016) apresentam um novo modelo experimental e um modelo estatístico para, simultaneamente, descobrir tratamentos presentes num acervo e, ainda, estimar seus efeitos causais. De forma adicional, focados em surveys com questões de respostas abertas, Roberts et al. (2014) mostram como o fato de o STM admitir a inclusão de covariáveis pode ser útil para a análise de respostas a experimentos em survey, concluindo que tal abordagem possui grau razoável de sucesso quando comparada à codificação manual.

A combinação do pontencial da analise automatizada de conteúdo com a era do *Big Data* e a oportunidade de produção de inferências causais é, sem dúvida alguma, um dos campos mais promissores para a área de metodologia em ciências sociais. Certamente, as ciências sociais brasileiras têm muito a contribuir com esse processo e deve se manter na fronteira do conhecimento.

Considerações finais

O principal objetivo deste artigo foi apresentar ao leitor um leque atualizado das principais metodologias de análise automatizada de conteúdo. Sem esgotar a atual variedade de métodos, técnicas e modelos, trata-se de um guia inicial para essa intensa e instigante área de pesquisa. No Quadro 3 apresentamos uma visão geral dos principais métodos para análise quantitativa de textos revisados neste trabalho.

Como buscamos destacar ao longo de todo o artigo, o uso de métodos automatizados para análise do texto como dado (*text as data*) é algo ainda muito recente na história da humanidade. Mesmo sob a contribuição de diferentes áreas do conhecimento, sendo a análise de conteúdo um campo tradicional de dedicação das ciências sociais, não há melhor área para guiar e contribuir com esse avanço. Muito ainda será feito e consideramos importante que as ciências sociais brasileiras acompanhem a fronteira desse processo.

Quadro 3
Visão geral dos métodos para análise quantitativa de textos

Família	Técnica	Objetivo
Semelhança entre textos	Similaridade de cosseno	Medir quão similares são dois documentos.
	Algoritmo de Smith-Waterman	Encontrar quais os trechos mais similares entre dois documentos.
Métodos de classificação em categorias conhecidas	Dicionário (Análise de sentimentos)	Classificar documentos em categorias conhecidas com auxílio de um dicionário anotado.
	Supervisionado (classificador de Naive Bayes)	Classificar documentos em categorias conhecidas a partir de um conjunto de treinamento.
Métodos de classificação em categorias desconhecidas	Não supervisionado (LDA, Dynamic Multitopic Model, Expressed Agenda Model, STM)	Classificar documentos quando não se conhece as categorias previamente.
Métodos de escalonamento	*Wordscores*	Estimar posições políticas em uma dimensão predeterminada a partir de documentos de referência.
	Wordfish	Estimar posições políticas quando não se conhece previamente referências da dimensão.

Os trabalhos de Moreira (2016) e Izumi (2017), apresentados no início deste artigo, são um exemplo notável de que muito ainda pode e deve ser feito pelas ciências sociais no país. Se, de um lado, no Pequeno Expediente na Câmara dos Deputados há evidências consistentes para concluir que os temas enfatizados não são governados pela relação governo-oposição (MOREIRA, 2016), de outro, no Senado Federal o posicionamento político resgata a importância dessa variável (IZUMI, 2017). Afinal de contas, há um padrão geral de fala dos parlamentares no Congresso Nacional? Para além dos temas, qual é o posicionamento político dos deputados federais nos discursos proferidos no Pequeno Expediente? Ademais, qual seria a ênfase temática presente nos discursos dos senadores?

As diferentes conclusões apresentadas nesses trabalhos e as inúmeras lacunas entre elas demonstram quão rica é esta agenda de pesquisa e quão importante é a dedicação das ciências sociais brasileiras a essa temática. A paisagem no horizonte é deslumbrante e está pronta para ser explorada.

Referências

ALDRICH, J.; MCKELVEY, R. A method of scaling with applications to the 1968 and 1972 presidential elections. *The American Political Science Review*, Washington, DC, v. 71, n. 1, p. 11-130, 1973.

BARBERÁ, P. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, Cambridge, UK, v. 23, n. 1, p. 76-91, 2015.

BARRON, A. et al. Individuals, institutions, and innovation in the debates of the French Revolution. *Proceedings of the National Academy of Sciences*, Washington, DC, v. 115, n. 18, p. 4607-4612, 2018.

BERINSKY, A.; HUBER, G.; LENZ, G. Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis*, Cambridge, UK, v. 20, n. 3, p. 351-368, 2012.

BISHOP, C. *Neural networks for pattern recognition*. Gloucestershire: Clarendon Press, 1995.

BLEI, D. M.; LAFFERTY, J. D. Dynamic topic models. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 23., 2006, New York. *Proceedings…* New York: ACM, 2006. pp. 113-120.

BLEI, D.; NG, A.; JORDAN, M. Latent dirichlet allocation. *Journal of Machine Learning Research*, Cambridge, MA, v. 3, n. 1, p. 993-1022, 2003.

BRADY, H. The perils of survey research: inter-personally incomparable responses. *Political Methodology*, Oxford, UK, v. 11, n. 3-4, p. 269-291, 1985.

BREIMAN, L. Random forests. *Journal of Machine Learning Research*, Cambridge, MA, v. 45, n. 1, p. 5-32, 2001.

BUDGE, I. et al. *Mapping policy preferences*: estimates for parties, electors, and governments, 1945-1998. Oxford, UK: Oxford University Press, 2001.

CAMPBELL, S. PENNEBAKER, J. The secret life of pronouns flexibility in writing style and physical health. *Psychological Science*, Washington, DC, v. 14, n. 1, p. 600-65, 2003.

CAMPOS, L. A., FERES JR., J.; GUARNIERI, F. 50 Anos da Revista DADOS: uma análise bibliométrica do seu perfil disciplinar e temático. *Dados*, Rio de Janeiro, v. 60, n. 3, p. 623-661, 2017.

CARRUBBA, C. et al. Off the record: unrecorded legislative votes, selection bias and roll-call vote analysis. *British Journal of Political Science*, Cambridge, UK, v. 36, n. 4, p. 691-704, 2006.

CHANG, J. et al. Reading tea leaves: how humans interpret topic models. In: BENGIO, Y. et al. *Advances in neural information processing systems*. Cambridge, MA: MIT Press, 2009. p. 288-296.

CLINTON, J.; JACKMAN, S.; RIVERS, D. The statistical analysis of roll call data. *American Political Science Review*, Washington, DC, v. 98, n. 2, p. 355-370, 2004.

EFRON, B.; GONG, G. A leisurely look at the bootstrap, the jacknife, and cross-validation. *American Statistician*, Abingdon, v. 37, n. 1, p. 36-48, 1983.

EGAMI, N. et al. *How to make causal inferences with text*. Working paper. 2018. Disponível em: <https://bit.ly/2M-tXMdq>. Acesso em: 21 jul. 2018.

FEINERER, I. HORNIK, K. tm: Text Mining Package. *R package*, [s.l.], 2018. Disponível em: <https://bit.ly/2K-cAx2w>. Acesso em: 21 jul. 2018.

FLESCH, R. A new readability yardstick. *Journal of Applied Psychology*, Washington, DC, v. 32, n. 3, p. 221-233, 1948.

FOKKENS, A. et al. Offspring from reproduction problems: what replication failure teaches us. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 51., 2013, Sofia. *Proceedings…* Sofia: Association for Computational Linguistics, 2013. (Volume 1: Long Papers). p. 1691-1701.

FONG, C.; GRIMMER, J. Discovery of treatments from text corpora. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 54., 2016, Berlin. *Proceedings…* Berlin: Association for Computational Linguistics, 2016. p. 1-10. FREY, B.; DUECK, D. Clustering by passing messages between data points. *Science*, Washington, DC, v. 315, n. 5814, p. 972-976, 2007.

GARRETT, K.; JANSA, J. Interest group influence in policy diffusion networks. *State Politics & Policy Quarterly*, Thousand Oaks, v. 15, n. 3, p. 387-417, 2015.

GRIMMER, J. A Bayesian hierarchical topic model for political texts: measuring expressed agendas in Senate press releases. *Political Analysis*, Cambridge, UK, v. 18, n. 1, p. 1-35, 2010.

_____. 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.; KING, G. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, Washington, CD, v. 108, n. 7, p. 2643-2650, 2011.

GRIMMER, J.; STEWART, B. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, v. 21, n. 3, p. 267-297, 2013.

GRÜN, B.; HORNIK, K. Topicmodels: AN R Package for fitting topic models. *Journal of Statistical Software*, Innsbruck, v. 40, n. 13, p. 1-30, 2011.

HAND, D. Classifier technology and the illusion of progress. *Statistical Science*, Bethesda, v. 21, n. 1, p. 1-14, 2006.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning*. New York: Springer, 2001.

HERNAN, M.; ROBINS, J. *Causal inference*. Boca Raton: CRC Press, 2018.

HOPKINS, D. KING, G. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, Washington, DC, v. 54, n. 1, p. 229-247, 2010.

HOPKINS, D. et al. ReadMe: software for automated content analysis. *Gari King*, Cambridge, MA, 2017. Disponível em: <https://bit.ly/2Mq7HRl>. Acesso em> 21 jul. 2018.

IMBENS, G.; RUBIN, D. *Causal inference in statistics, social, and biomedical sciences*. Cambridge, UK: Cambridge University Press, 2015.

ITTI, L.; BALDI, P. Bayesian surprise attracts human attention. In: JORDAN, M. I.; LECUN, Y.; SOLLA, S. A. (Eds.). *Advances in neural information processing systems*: proceedings of the first 12 conferences. Cambridge, MA: The MIT Press, 2006.

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

JURAFSKY, D.; MARTIN, J. *Speech and natural language processing*: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River: Prentice Hall, 2009.

KING, G. et al. Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, Cambridge, UK, v. 98, n. 1, p. 191-207, 2004.

KRIPPENDORFF, K. *Content analysis: an introduction to its methodology*. New York: Sage, 2004.

KROEGER, M. *Plagiarizing policy*: model legislation in state legislatures. Working paper. 2015. Disponível em: <https://bit.ly/2o0lpf5>. Acesso em: 21 jul. 2018.

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

LAVER, M.; BENOIT, K.; GARRY, J. Extracting policy positions from political texts using words as data. *American Political Science Review*, Washington, DC, v. 97, n. 2, p. 311-331, 2003.

LAZER, D. et al. Life in the network: the coming age of computational social science. *Science*, New York, v. 323, n. 5915, p. 721, 2009.

LI, W.; LAROCHELLE, D.; LO, A. *Estimating policy trajectories during the financial crisis*. Working paper. 2014. Disponível em: <https://bit.ly/2MtZfjN>. Acesso em: 21 jul. 2018.

LIU, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, London, v. 5, n. 1, p. 1-167, 2012.

LOWE, W. Understanding wordscores. *Political Analysis*, Cambridge, UK, v. 16, n. 4, p. 356-371, 2008.

_____. Austin: do things with words. *Conjugateprior*, Princeton, 2015. Disponível em: <https://bit.ly/2BCFGAY>. Acesso em: 21 jul. 2018.

LUCAS, C. et al. Computer-assisted text analysis for comparative politics. *Political Analysis*, Cambridge, UK, v. 23, n. 2, p. 254-277, 2015.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: LE CAM, L. M.; NEYMAN, J. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1967. (Volume 1: Statistics). p. 281-297.

MAGALHÃES, R. et al. Perspectives on political methodology: interview with Simon Jackman. *Leviathan*, São Paulo, n. 7, p.158-175, 2013.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to information retrieval*. Cambridge: Cambridge University Press, 2008.

MARON, M.; KUHNS, J. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, New York, v. 7, n. 3, p. 216-244, 1960.

MCCARTY, N. POOLE, K. ROSENTHAL, H. *Polarized America*: the dance of ideology and unequal riches, Cambridge: MIT University Press, 2006.

MONROE, B.; SCHRODT, P. Introduction to the special issue: the statistical analysis of political text. *Political Analysis*, Cambridge, UK, v. 16, n. 4, p. 351-355, 2008.

MONROE, B. et al. No! Formal theory, causal inference, and big data are not contradictory trends in political science. *PS: Political Science & Politics*, Cambridge, UK, v. 48, n. 1, p. 71-74, 2015.

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) – Universidade de São Paulo, SP, 2016.

NEUENDORF, K. *The content analysis guidebook*. Thousand Oaks: Sage, 2002.

OOMS, J. Tesseract: Open Source OCR Engine. *R package*, [s.l.], 2018. Disponível em: <https://bit.ly/2whiySw>. Acesso em 21 jul. 2018.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, Hanover, v. 2, n. 1-2, p. 1-135, 2008.

PATTY, J.; PENN, E. Analyzing big data: social choice and measurement. *PS: Political Science and Politics*, Cambridge, UK, v. 48, n. 1, p. 95-101, 2015.

PEARL, J. *Causality*. Cambridge, UK: Cambridge University Press, 2009.

PENNEBAKER, J. W.; MEHL, M. R.; NIEDERHOFFER, K. G. Psychological aspects of natural language use: our words, ourselves. *Annual Review of Psychology*, Palo Alto, v. 54, n. 1, p. 547-577, 2003.

POOLE, K.; ROSENTHAL, H. *Ideology and congress*. New Brunswick: Transaction Publishers, 2007.

PORTER, M. F. An algorithm for suffix stripping. *Program*: Electronic Library and Information Systems, Belfast, v. 14, n. 3, p. 130-137, 1980.

POWER, T.; ZUCCO, C. Estimating ideology of Brazilian legislative parties, 1990-2005: a research communication. *Latin American Research Review*, Pittsburgh, v. 44, n. 1, p. 218-246, 2009.

QUINN, K. et al. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, Washington, DC, v. 54, n. 1, p. 209-228, 2010.

ROBERTS, M. E. Introduction to the Virtual Issue: recent innovations in text analysis for social science. *Political Analysis*, Cambridge, UK, v. 24, n. 10, p. 1-5, 2016.

ROBERTS, M.; STEWART, B.; TINGLEY, D. stm: R Package for Structural Topic Models. *R package*, [s.l.], 2018. Disponível em: <https://bit.ly/2wc0rOT>. Acesso em: 3 jul. 2018.

ROBERTS, M. E. et al. *The structural topic model and applied social science*. Advances in neural information processing systems workshop on topic models: computation, application, and evaluation. Cambridge, MA: Harvard University, 2013.

_____. Topic models for open-ended survey responses with applications to experiments. *American Journal of Political Science*, Washington, DC, v. 58, n. 4, p. 1064-1082, 2014.

SLAPIN, J.; PROKSCH, S.-O. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, Washington, DC, v. 52, n. 3, p. 705-722, 2008.

SMITH, T.; WATERMAN, M. Identification of common molecular subsequences. *Journal of Molecular Biology*, Amsterdam, v. 147, n. 1. p. 195-197, 1981.

SOUZA, M.; VIEIRA, R. Sentiment analysis on Twitter data for Portuguese language. In: INTERNATIONAL CONFERENCE COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 10., 2012, Coimbra. *Proceedings…* Coimbra: University of Coimbra, 2012. p. 241-247.

SOUZA, M. et al. Construction of a Portuguese opinion lexicon from multiple resources. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 8., 2011, Uberlândia. *Proceedings…* Uberlândia: Federal University of Uberlândia, 2011. pp. 59-66.

SPIRLING, A. Democratization and linguistic complexity: the effect of franchise extension on parliamentary discourse, 1832-1915. *The Journal of Politics*, Chicago, v. 78, n. 1, p. 120-136, 2015.

TABOADA, M. et al. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, Cambridge, MA, v. 37, n. 2, p. 267-307, 2011.

VENABLES, W. N.; RIPLEY, B. D. *Modern applied statistics with S*. 4. ed. New York: Springer, 2002.

WALLACH, H. et al. An alternative prior for nonparametric Bayesian Clustering. In: International CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 13., 2010, Sardinia. *Proceedings…* Sardinia: Chia Laguna Resort, 2010. p. 892-999, 2010.

WEBER, R. P. *Basic content analysis*. Newbury Park: Sage, 1990. (University Paper Series on Quantitative Applications in the Social Sciences).

WELBERS, K.; VAN ATTEVELDT, W.; BENOIT, K. Text analysis in R. *Communication Methods and Measures*, Abingdon, v. 11, n. 4, p. 245-265, 2017.

WICKHAM, H. httr: Tools for Working with URLs and HTTP. *R package*, [s.l.], 2016. Disponível em: <https://bit.ly/2PwgzT0>. Acesso em: 20 jul. 2018.

_____. rvest: Easily Harvest (Scrape) Web Pages. *R package*, [s.l.], 2018. Disponível em: <https://bit.ly/2wee0fI>. Acesso em: 21 jul. 2018.

WICKHAM, H.; HESTER, J.; OOMS, J. xml2: Parse XML. *R package*, [s.l.], 2018. Disponível em: <https://bit.ly/2MrMzdi>. Acesso em: 20 jul. 2018.

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

WILKERSON, J.; SMITH, D.; STRAMP, N. Tracing the flow of policy ideas in legislatures: a text reuse approach. *American Journal of Political Science*, Washington, DC, v. 59, n. 4, p. 943-956, 2015.

Resumo

O texto como dado: desafios e oportunidades para as ciências sociais

A comunicação é instrumento fundamental para as relações humanas. É por meio dela, por exemplo, que valores são construídos, símbolos sociais são estabelecidos, tradições são repassadas, debates são concretizados, a política se materializa e o conflito político se expressa. Foco de análises dos cientistas sociais há séculos, a análise do conteúdo transmitido na comunicação sempre esteve restrita à necessidade de volumes relevantes de recursos para a avaliação manual de grandes acervos. Revertendo esse quadro limitado, recentes desenvolvimentos tecnológico, computacional e científico permitem que as ciências sociais potencializem sua investigação reduzindo drasticamente os custos envolvidos na análise de grandes acervos. Por intermédio de novos métodos desenvolvidos, atualmente, é possível verificar comportamentos que antes não eram observáveis, medir quantidades anteriormente imensuráveis e testar hipóteses até então impossíveis de serem testadas. Nesse escopo, o principal objetivo deste artigo é manter as ciências sociais brasileiras na fronteira desse processo e apresentar ao leitor um leque atualizado das principais metodologias de análise automatizada de conteúdo. Sem esgotar suas inúmeras possibilidades, este artigo é um guia para a inovadora e instigante área de pesquisa do texto como dado.

Palavras-chave: Análise Automatizada de Conteúdo; Semelhança entre Textos; Métodos de Classificação; Métodos de Escalonamento; *Big Data*.

Abstract

The text as data: challenges and opportunities for Social Sciences

Communication is a fundamental tool for human relations. It is through communication that values are constructed, social symbols are established, traditions are passed on, debates are realized, politics are materialized and political conflict is expressed. A focus in analyses of social scientists, the analysis of the content trasmitted in communication has always been restricted to the need for a great amount of research funds for the manual assessment of large collections. Changing this limited scenario, recent technological, computational and scientific developments allowed social scientists to analyse larger collections of documents with low cost. Currently, through the development of new methods, it is now possible to identify behaviors that could not be observed, to measure quantities that could not be quantified, and to test hypothesis that could not be tested. In this sense, the main objective of this study is to maintain Brazilian Social Sciences at the frontier of this process and present to the reader the latest methodologies for automated content analysis. Without exhausting its several possibilities, this article is a guide to the innovative area of researching text as data.

Keywords: Automated Content Analysis; Similarity between Texts; Classification Methods; Scheduling Methods; Big Data.

Résumé

Le texte en tant que donné : défis et opportunités pour les Sciences Sociales

La communication est un outil fondamental pour les relations humaines. C'est par la communication que des valeurs sont construites, des symboles sociaux sont établis, des traditions sont transmises, des débats sont réalisés, des politiques sont materialisées et des conflits politiques sont exprimés. Un accent dans les recherches des sociologues, l'analyse du contenu transmis dans la communication a toujours été limitée au besoin d'une grande quantité de fonds de recherche pour l'évaluation manuelle de grandes collections. En changeant cet scénario limité, les récents développements technologiques, informatiques e scientifiques ont permis aux sociologues d'analyser des plus grandes collections de documents à bas prix. Actuallement, grâce au développement de nouvelles methodes, il est désormais possible de identifier comportements qui étaient inobservables, de mesurer des quantités auparavant incommensurables et tester des hypothèses jusqu'alors impossibles. Dans ce sens, l'objectif de cet article est de mantenir les Sciences Sociales brésiliennes à la frontier de ce processus et de présenter au lecteur les méthodologies le plus récentes pour l'analyse de contenu automatisée. Sans épuiser ses nombreuses possibilités, cet article est un guide sur le domaine innovant de la recherche des textes en tant que donnés.

Mots-clés: Analyse de Contenu Automatisée ; Similarité entre textes ; Méthodes de Classification ; Methodes de Planification ; Big Data.