# PARLA: OS TEMAS MAIS DEBATIDOS NA CÂMARA DOS DEPUTADOS

**Ricardo Modesto Vieira** *

**Palavras-chaves:** Análise do Discurso. Texto como Dado. Análise Automatizada de Conteúdo.

## RESUMO

A revolução do *big data* e da inteligência artificial forneceu uma grande oportunidade para novas formas de disponibilização dos discursos e dos debates parlamentares. Como os deputados expressam suas opiniões e defendem suas posições por meio de palavras, é possível utilizar a frequência dessas palavras e expressões no discurso parlamentar para mostrar quais foram os temas mais debatidos na Câmara dos Deputados. Com esse propósito surgiu o Parla, que utiliza os discursos proferidos no Plenário para mostrar quais foram os temas mais debatidos durante a legislatura. O objetivo é fornecer um retrato fiel do que os seus representantes falaram na Casa de todos os brasileiros. Para garantir um retrato fiel do que fala cada deputado, foi necessária a utilização de dois diferentes métodos: o saco de palavras e o *Naive Bayes/Decision Tree*, pois não há um método global para a análise automatizada de conteúdo (GRIMMER; STEWART, 2013).

O saco de palavras mede apenas a frequência das palavras no discurso do Deputado. Para manter a importância da ordem das palavras no discurso, o Parla mostra apenas as expressões de uso mais frequente entre duas a cinco palavras. Como as palavras únicas não traziam conteúdo expressivo, elas foram retiradas do modelo e cada conjunto de até cinco palavras foi considerado como uma "palavra única". O objetivo é criar uma matriz de documentos e termos (*Document Term Matrix* – DTM) na qual cada linha representa um documento e cada coluna representa um termo – bigrama, trigrama ou até 5-gram – único. Nesse sentido, cada célula da matriz denota o número de vezes que cada um desses termos linguísticos de até cinco palavras, indicados na coluna, aparece no documento indicado na linha; consequentemente, cada documento é representado por um vetor único.

Já o *Naive Bayes/Decision Tree* é um método supervisionado de aprendizado de máquina. Os métodos de aprendizagem supervisionada usam a frequência em que as palavras aparecem em um texto para classificar os documentos em categorias predeterminadas. O algoritmo então "aprende" como classificar os documentos nessas categorias usando um conjunto de treinamento. Ou seja, o algoritmo usa características dos documentos para classificá-los nas categorias (GRIMMER; STEWART, 2013). O Parla utilizou 6.200 sentenças, classificadas manualmente pelos servidores da indexação de discursos, como conjunto de teste para aprender a identificar qual foi o tema mais debatido dentro de um conjunto pré-estabelecido de 31 temas.

Baseado no teorema de *Bayes*, o *Naive Bayes* é um dos métodos supervisionados de classificação mais utilizados na literatura de Ciência Política. Embora parta de um pressuposto ingênuo — o modelo assume que as palavras são geradas de forma independente para uma dada categoria (*the naive assumption*), quando na verdade o uso de palavras é altamente correlacionado em qualquer conjunto de dados —, o modelo fornece

* Câmara dos Deputados. E-mail: ricardomodestovieira@gmail.com

um método alternativo útil para atribuir documentos a categorias predeterminadas (GRIMMER; STEWART, 2013; IZUMI; MOREIRA, 2018). No caso de grandes acervos, o classificador *Decision Tree* pode ser utilizado em conjunto com o *Naive Bayes* para aumentar a precisão (KOHAVI, 2011). Esse algoritmo "faz perguntas" sobre os dados até conseguir filtrar a informação o suficiente para fazer uma predição. No caso dos discursos, as ramificações da árvore são definidas pelos sacos de palavras de todas as sentenças e organizadas de forma otimizada para que, quando receber um novo saco palavras o algoritmo consiga filtrar e predizer a qual tema pertence.

Para aplicar o *Naive Bayes/Decision Tree* foi necessário primeiro definir os temas e, depois, ensinar o algoritmo a classificar nesses temas. Em princípio, foi utilizada a tabela de classificação e indexação de temas da Câmara dos Deputados. Depois, foram retirados os temas "Homenagens e Datas Comemorativas" e "Processo Legislativo e Atuação Parlamentar", por conterem as atividades de representação parlamentar, não contribuindo com conteúdo ao debate temático. Outros dois temas – "Administração Pública" e "Política, Partidos e Eleições" – foram considerados muito amplos, mas, como esses temas apresentam conteúdo significativo, eles foram divididos. Assim, o tema "Administração Pública" foi separado em "*Impeachment*", "Corrupção" e "Serviço Público" e o tema "Política, Partidos e Eleições" foi separado em "Reforma Política" e "Eleição". Dessa forma, o Parla apresenta 40% de precisão no classificador macro temático e em média 70% nos classificadores menores.

Para que a aplicação dos dois algoritmos fosse possível, foi necessário realizar uma séria de etapas de pré-processamento com os discursos. A fim de reduzir a complexidade e o tamanho do vocabulário, bem como focar no que é usual e significativo no texto, o Parla utiliza apenas as palavras de média frequência no discurso. Por isso foi preciso retirar as palavras de conteúdo desnecessário – as que aparecem em 90% dos discursos - e as pouco frequentes — em menos de 1%. As mais frequentes normalmente não geram conteúdo significativo e correspondem ao vocabulário fechado da língua Portuguesa – como conjunções, preposições, artigos, pronomes e verbos, como os de ligação. Também são também retiradas as palavras e expressões que são utilizadas muito comumente no processo legislativo, mas que não geram conteúdo significativo, como as falas procedimentais do Presidente, a leitura da ata e da agenda, eleições da Mesa, orações e homenagens. Essas listas de palavras são chamadas de listas de *stopwords*. No caso do Parla, também foram retirados os nomes dos Estados e o nome dos deputados federais.

Após a remoção das *stopwords*, é necessário reduzir a variabilidade das palavras por meio de *stemming*. *Stemming* é a redução da palavra a seu radical por meio da remoção de seu final, como em plurais ou conjugações verbais, a fim de reduzir as palavras à sua forma básica e agrupá-las. Esse processo de reduzir a palavra ao seu radical não necessariamente vai resultar na raiz exata. Para isso é necessário utilizar um algoritmo mais complexo que identifica a origem da palavra e retorna apenas sua *lemma* ou raiz. Como a língua Portuguesa é bastante complexa, aplicar todas as regras e exceções do processo de reduzir um termo ao seu radical tornaria o algoritmo muito lento. Após o *stemming*, as ocorrências individuais de cada palavra são chamadas de *tokens* e o conteúdo do discurso está finalmente pronto para ser convertido em dados quantitativos.

Para aumentar a usabilidade da ferramenta, o Parla disponibiliza filtros que possibilitam a separação dos discursos por partido, por Estado ou por homens e mulheres. Assim, é possível conhecer o que falam apenas as deputadas, ou os deputados de um Estado, ou comparar o que falam os deputados de diferentes partidos. Outra opção é conhecer o que

falam naturalmente os deputados, sem a interferência da agenda parlamentar ou do controle partidário. Para isso, o Parla disponibiliza um filtro que captura apenas os discursos de Pequeno e Grande Expediente, nos quais cada deputado é livre para debater sua própria agenda.

## REFERÊNCIAS

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

KOHAVI, R. **Scaling Up the Accuracy of Naive-Bayes Classifiers**: a Decision-Tree Hybrid. Mountain View (CA): Data Mining and Visualization Silicon Graphics, 2011.

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