Disponível em: www.univali.br/periodicos ISSN: 1983-716X

Revista Alcance

# MINERAÇÃO DE TEXTO PARA ANÁLISE DE DISCURSO: TEMÁTICAS E ARGUMENTOS DA DECISÃO DE VOTO DE DEPUTADOS DURANTE A VOTAÇÃO DO *IMPEACHMENT*

## TEXT MINING FOR DISCOURSE ANALYSIS: THEMES AND ARGUMENTS OF THE DEPUTIES' VOTING DECISION DURING THE IMPEACHMENT VOTE

## MINERÍA DE TEXTO PARA ANÁLISIS DE DISCURSO: TEMÁTICAS Y ARGUMENTOS DE LA DECISIÓN DE VOTO DE DIPUTADOS DURANTE LA VOTACIÓN DEL IMPEACHMENT

**CARLA BONATO MARCOLIN**
Doutora
Universidade Federal de Uberlândia - Brasil
cbmarcolin@gmail.com
ORCID: http://orcid.org/0000-0003-0260-5073

**FERNANDA DA SILVA MOMO**
Doutoranda
Universidade Federal do Rio Grande do Sul – Brasil
fernandamomo@yahoo.com.br
ORCID: http://orcid.org/0000-0002-6512-5280

**JOÃO LUIZ BECKER**
Doutor
Universidade Federal do Rio Grande do Sul – Brasil
jlbecker@ea.ufrgs.br
ORCID: http://orcid.org/0000-0003-4176-7374

**ARIEL BEHR**
Doutor
Universidade Federal de Uberlândia - Brasil
ariel.behr@ufrgs.br
ORCID: http://orcid.org/0000-0002-9709-0852

Submetido em: 28/08/2018
Aprovado em: 18/03/2019

Doi: alcance.v26n1(Jan/Abr).p004-012

**RESUMO**

O avanço de técnicas para análise de dados não estruturados pode auxiliar a compreender melhor o posicionamento e os votos dos políticos que representam uma população. O objetivo do presente artigo é analisar a relação semântica latente das temáticas presentes nos argumentos da decisão de voto dos parlamentares de diferentes partidos políticos. Para tal, foram utilizados dados de discurso de todos os deputados durante a votação do *impeachment*, ocorrida em 2015. Nesse sentido, utilizaram-se como base teórica para a realização das análises a perspectiva de Weiss (1983) sobre a tomada de decisão de políticos e a teoria da dissonância cognitiva de Festinger (1957). Adicionalmente, a partir do uso da técnica LSA (*Latent semantic analysis*), técnica de mineração de texto baseada em decomposição matricial, buscou-se contribuir com as análises ao trazer resultados relacionados aos principais termos associados e uso de determinadas palavras no contexto político. Como

resultados, verificou-se que, para o caso apresentado, o discurso dos deputados não é um elemento que permite separar os diferentes grupos votantes, o que indica que, para compreender a posição de um político e escolher melhor seu representante, os cidadãos precisam ir além do seu discurso.

**Palavras-Chave**: Mineração de Texto. LSA. Discurso político.

**ABSTRACT**

The advances in techniques for analyzing unstructured data can help to better understand the positioning and votes of politicians who represent a population. This article analyses the underlying semantic relationship between the themes present in the arguments for the voting decision of parliamentarians of different political parties. For this, it uses discourse data from all the deputies during the impeachment voting, which took place in 2015. Weiss's (1983) perspective on the decision-making of politicians, and Festinger's (1957) theory of cognitive dissonance were used as the theoretical basis for the analysis. Additionally, using the technique of LSA (Latent semantic analysis) — a text mining technique based on matrix decomposition — it aims to contribute to the analyses by bringing results related to the main associated terms, and the use of certain words in the political context. It was found that for the case presented, the deputies' discourse is not an element that enables the different voting groups to be distinguished, indicating that in order to understand the position of a politician, and better choose their representative, citizens need to go beyond the politicians' discourse.

**Keywords**: Text Mining. LSA. Political speech.

**RESUMEN**

El avance de técnicas para análisis de datos no estructurados puede ayudar a comprender mejor el posicionamiento y los votos de los políticos que representan una población. El objetivo del presente artículo es analizar la relación semántica latente de las temáticas presentes en los argumentos de la decisión de voto de los parlamentarios de diferentes partidos políticos. Para esto, se utilizaron datos de discurso de todos los diputados durante la votación del impeachment, ocurrida en 2015. En ese sentido, se utilizó como base teórica para la realización de los análisis la perspectiva de Weiss (1983) sobre la toma de decisión de políticos y la teoría de la disonancia cognitiva de Festinger (1957). Además, a partir del uso de la técnica LSA (Latent semantic analysis), técnica de minería de texto basada en descomposición matricial, se buscó contribuir con los análisis al traer resultados relacionados a los principales términos asociados y uso de determinadas palabras en el contexto político. Como resultados, se constató que, para el caso presentado, el discurso de los diputados no es elemento que permite separar a los diferentes grupos votantes, lo que indica que para comprender la posición de un político y elegir mejor su representante, los ciudadanos deben ir más allá de su discurso.

**Palabras Clave*:*** Minería de texto. LSA. Discurso político.

## 1. INTRODUÇÃO

As organizações enfrentam muitos desafios devido à complexidade e à diversidade do ambiente. Risco, tempo e informação envolvem grande parte dos processos gerenciais. Neste processo, as tecnologias têm se apresentado como capazes de apoiar e trazer resultados de forma a minimizar o risco, em menos tempo e de uma forma mais precisa. Entretanto, diferentes organizações se deparam com um grande volume de dados disponível, o que se pode caracterizar como Big Data (McAfee & Brynjolfsson, 2012; Wu, Zhu, Wu, & Ding, 2014).

Este grande volume de dados pode se constituir em um desafio, gerando sobrecarga de informação. Mesmo sem um consenso na literatura, relaciona-se tal sobrecarga com o fato de receber muita informação (Eppler & Mengis, 2004; Popovič, Hackney, Tassabehji & Castelli, 2018), o que pode potencialmente diminuir a qualidade da análise dos dados e levar a decisões equivocadas. Além disso, a sobrecarga de informação está relacionada ao sentimento de perda de controle da situação, uma vez que, de toda esta informação disponível, nem tudo (e por vezes muito pouco) pode ser considerado útil e relevante para o processo decisório (Bawden & Robinson, 2009).

Ao mesmo tempo, o Big Data mostra-se uma grande oportunidade para diferentes organizações. Junto com o crescimento do volume de informação, especula-se um crescente potencial para seu uso, agregando importância à análise e à compreensão dos dados de processos organizacionais, uma vez que há possibilidade de melhor entender os diferentes atores como clientes, mercados, produtos, ambiente organizacional, impactos da tecnologia, dentre outros (Marchand & Peppard, 2013; Mayer-Schonberger & Cukier, 2013).

Dado que informação e conhecimento estão intrinsicamente relacionados, há um desafio emergente em transformar todo volume de dados em formas efetivas de conhecimento com valor para diferentes instituições. Assim, faz-se necessário discutir tecnologias, ferramentas e soluções que permitam extrair valor efetivo de todas informações disponíveis de forma adequada e eficiente (Freitas Júnior, Maçada, Oliveira & Brinkhues, 2016).

Um dos principais desafios, dado o cenário de Big Data, é a variabilidade. Apesar de muitas informações estarem estruturadas, grande parte do volume crescente de dados não é estruturado, em formato de vídeos, imagens ou textos, estando, portanto, fora de bancos de dados. Trabalhar com este tipo de dados é um desafio, uma vez que não é possível utilizar processos tradicionais de ETL (*extract, transform, load*, do inglês extrair, transformar e carregar), que já foram muito pesquisados e estão implementados em grande parte das ferramentas de análise de dados (Chen & Zhang, 2014)

Porém, não se pode desconsiderar o potencial de uso desses dados somente por conta de não serem análises tradicionais. Especialmente após a explosão das redes sociais, milhares de pessoas comentam, publicam e compartilham dados o tempo todo através de equipamentos móveis, como *tablets*, *smartphones* e *notebooks*. E todos estes canais de expressão de opinião acabam criando um ambiente sobrecarregado de dados em texto: uma simples busca no Google, por exemplo, pode produzir milhares de resultados em apenas alguns segundos. Nesse sentido, ainda, verifica-se que, assim como termômetros são sensores que representam uma informação da realidade (a temperatura), e um GPS pode especificar uma determinada localização, os dados em texto podem ser sensores análogos, da percepção sobre a realidade, vinda de um indivíduo, uma comunidade ou até mesmo uma região (Aggarwal & Zhai, 2012).

Em relação à tomada de decisão no âmbito político, destaca-se que esta é preponderante para a determinação das ações do estado. Nesse sentido, Weiss (1983) descreve que essa tomada de decisão nesse contexto não é baseada apenas em evidências, e sim em informação, ideologia e interesse. Assim, em um ambiente como esse, é comum a existência de um elevado grau de contradições e, portanto, há uma maior propensão à dissonância cognitiva (Festinger, 1957), que pode ser expressa no discurso do tomador de decisão e gerada a partir das perspectivas destacadas por Weiss (1983) como base para a tomada de decisão política. Percebe-se, assim, que a possibilidade de analisar de forma objetiva um conjunto de textos políticos pode auxiliar a compreender se esse fenômeno existe e de que forma ele é expresso no posicionamento e nos votos dos políticos que representam uma população.

Diante do contexto exposto, tomando como base de dados o conjunto de discursos dos 513 deputados da Câmara Federal com suas justificativas para voto sobre o processo de *impeachment* da Presidenta Dilma Rousseff, o objetivo do presente artigo é analisar a relação semântica latente das temáticas presentes nos argumentos da decisão de voto dos parlamentares de diferentes partidos políticos.

Adicionalmente, busca-se contribuir com a análise de uma temática em um grande conjunto de textos, ao trazer resultados relacionados aos principais termos associados e uso de determinadas palavras no contexto político, a partir de uma técnica de mineração de texto baseada em decomposição matricial (Visinescu & Evangelopoulos, 2014; Deerwester *et al*., 1990). Além disso, enfatiza-se a contribuição potencial das técnicas de mineração de texto como ferramentas de análise que possam ilustrar elementos da tomada de decisão e da dissonância cognitiva (Kladis & Freitas, 1996; Festinger, 1957).

Este artigo traz na próxima seção a abordagem aos temas da Tomada de Decisão e da Dissonância Cognitiva, revisadas por sua pertinência em relação ao contexto estudado. Após, a seção trata sobre modelo utilizado para mineração de texto. A seção seguinte traz os procedimentos de tratamento executados na base de dados da pesquisa e as escolhas adotadas para a pesquisa, inerentes ao processo de mineração. A seção posterior apresenta os principais resultados encontrados e, por fim, a última seção traz a conclusão, as principais limitações e as sugestões de direções futuras ao campo de pesquisa.

## 2. TOMADA DE DECISÃO E DISSONÂNCIA COGNITIVA

Os ambientes institucionais, sejam eles relativos ao setor público ou privado, carecem em seu funcionamento da realização de suscetíveis processos de escolhas. Nesse sentido, torna-se "impossível pensar na organização sem considerar a ocorrência constante no processo decisório" (Freitas *et al.*, 1997, p. 37). No âmbito político brasileiro, tendo em vista a democracia como regime político, o papel da decisão dos representantes eleitos pela população é preponderante no rumo das ações do estado.

Além disso, identifica-se uma cultura política em que há uma relação clientelista entre representante e representado a partir das práticas de gestão pública (Pinho, 1998), o que pode ser vislumbrado na posição política adotada pelos representantes e seus discursos. Assim, conforme Weiss (1993), põe-se em questionamento o pressuposto de racionalização das políticas que enfoca como base para tomada de decisão evidências acima dos interesses. Nesse sentido, Weiss (1983) afirma que se deve considerar, além das evidências, outras fontes de influência como ideologia e interesses dos representantes, sendo, portanto, sua desconsideração uma "[má leitura da] natureza da tomada de decisão democrática" (Weiss, 1983, p.224).

Segundo Coelho *et al.* (2016, p. 6), no que tange à tomada de decisão democrática, um de seus imperativos está relacionado a "acolher diferentes interesses e ideologias representadas na sociedade". Portanto, ainda conforme Weiss (1983), para se compreender a tomada de decisão política, é necessário observar a interação constante entre informação, ideologia e interesses. Por fim, destaca-se ainda que, quando há a discussão e a negociação entre diferentes grupos de atores, é importante se considerar influências em relação a estruturas e procedimentos (Coelho *et al*., 2016; Weiss, 1983).

Tendo em vista essas considerações, no que se refere ao contexto político brasileiro, cabe ressaltar que o poder legislativo se apresenta como principal representante dos interesses do povo, de forma que seu funcionamento e relacionamento com a sociedade refletem, de certa forma, o nível de desenvolvimento de um sistema político (Mainwaring, 2001; Amaral & Pinho, 2016). Assim, nesse ambiente importante para definições diretivas do país, é inegável a existência de um elevado grau de contradições, tendo em vista a pluralidade crescente de fragmentação partidária e ideológica presentes no congresso brasileiro (Braga, 2013).

Nesse ambiente, em que há um elevado grau de contradições, há uma maior propensão à dissonância cognitiva (Festinger, 1957). Essas contradições ocorrem, uma vez que os políticos são afeitos à opinião pública e, portanto, podem mudar de ideia perante uma pressão externa. A teoria da dissonância cognitiva, cunhada por Festinger, está baseada em três pressupostos: i) "pode existir relações dissonantes entre elementos cognitivos"; ii) a "existência de dissonância origina pressões para reduzir a dissonância e evitar aumentos nela"; iii) "manifestações da operação destas pressões incluem mudanças de comportamento, mudanças de cognição e exposição atenta a novas informações e opiniões" (Festinger, 1957, p. 31).

Portanto, nessa teoria é destacada a pressão psicológica a qual o tomador de decisão está exposto durante o processo de escolha, de forma que, inconscientemente, esse indivíduo privilegie informações que justifiquem suas atitudes e diminuam as possíveis dissonâncias existentes. Portanto, um decisor buscará inconscientemente informações que privilegiem a alternativa escolhida (Kladis & Freitas, 1996) e, mesmo assim, as dissonâncias podem estar expressas no seu discurso explicitando em muitos casos os componentes da decisão política destacados por Weiss (1983), que são: informação, ideologia e interesses.

## 3. DADOS EM TEXTO: O MODELO LSA

Dados em texto não são novidade advinda dos tempos atuais. As áreas de biblioteconomia e ciência da informação já endereçam questões de indexação e organização de dados de texto há bastante tempo. De uma forma mais recente, a ciência da computação vem apoiando, por meio de técnicas e modelos específicos nestas tarefas e em outras como recuperação e relevância dos textos trabalhados (Manning, Rhagavan & Schutze, 2009). Paralelamente, a área de mineração de dados vem se modernizando cada vez mais para atender ao grande volume de dados existentes, especialmente nas empresas, e como resolver questões de manipulação e análise de forma a acompanhar a crescente dinâmica e a velocidade dos processos (Aggarwal & Zhai, 2012).

Dados em texto também têm crescido de forma expressiva em diversas outras áreas, muito devido à Internet. O formato livre e simples de texto para usuários, aliado às diferentes e às crescentes plataformas (como *blogs*, redes sociais, fóruns, comunidades, etc.), aumentou expressivamente o volume de texto disponível, o que traz a necessidade de ferramentas que busquem formas de análise e padrões de uma maneira dinâmica e escalável, acompanhando a crescente geração de textos por parte dos usuários (Aggarwal & Zhai, 2012).

A dificuldade de dados neste formato é justamente a sua falta de estrutura. Enquanto que dados estruturados podem ser mais facilmente manipulados por meio de sistemas de gerenciamento de banco de dados, garantindo integridade e fornecendo agilidade para análise, os dados em texto não dispõem desse tipo de ferramenta por serem desestruturados (Visinescu & Evangelopoulos, 2014). Esta falta de estrutura está na sua origem: em vez de serem processados e gerados por códigos e funções, os dados de texto são criados diretamente pelos usuários, o que dificulta seu tratamento e sua análise em larga escala.

Assim, o objetivo da área de mineração de texto é permitir não apenas processar uma grande quantidade de texto, mas facilitar sua compreensão, buscando padrões e novos olhares para dados de texto. Por ser automatizado, permite ainda que as análises por meio da mineração de texto sejam escaláveis em quantidade e possíveis de serem repetidas sempre da mesma forma (Debortoli *et al.*, 2016).

Sendo um conjunto de técnicas de processamento da área de linguagem natural (NLP, *natural processing language*) combinado com técnicas de estruturação e descoberta de padrões em dados da mineração de dados, a área de mineração de texto é desafiadora em suas metodologias, uma vez que trabalha com semântica e significado de termos dependentes de contexto e interpretação. Alguns modelos buscam trabalhar estes desafios, sendo um deles o modelo de Análise Semântica Latente, conhecido como LSA (Landauer, 2007).

O modelo trabalha com a decomposição de valores singulares (SVD, de *singular value decomposition*). O principal objetivo de usar esta técnica é descobrir, a partir dos dados, uma estrutura escondida e latente, que revela e minimiza os efeitos de sinonímia e polissemia. Tudo começa a partir de uma matriz que reúne termos e documentos de um conjunto de textos (que podem ser artigos, livros, *e-mails*, comentários de redes sociais, etc.), o qual se denomina *corpus* (Manning, Rhagavan & Schutze, 2009). Esta matriz apresenta, nas linhas, os termos deste conjunto de documentos (palavras ou expressões), enquanto que cada coluna representa um documento do *corpus*. Após, a decomposição em valores singulares é realizada, conforme a Figura 1.

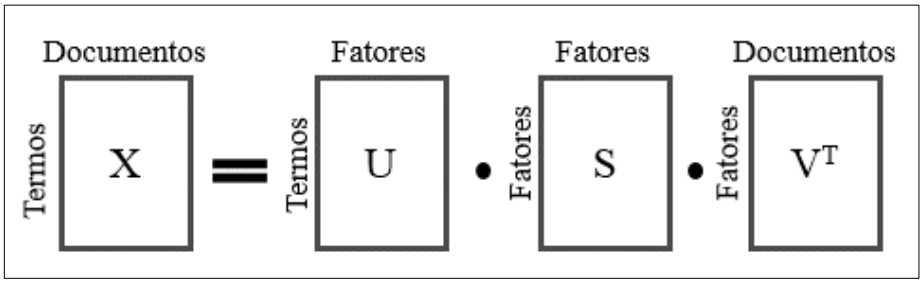


**Figura 1.** SVD no contexto de LSA
Fonte: Adaptada de Ashton, Evangelopoulos e Prybutok (2014).

O principal objetivo é descobrir a estrutura semântica de determinado conjunto de documentos, revelando quais deles estão mais próximos e qual a relação com os diferentes termos encontrados. O processo consiste em decompor a matriz X, contendo os termos e os documentos do *corpus* em outras três matrizes. A matriz U traz os autovetores da matriz X à direita, enquanto que a matiz $V^T$ traz os autovetores à esquerda, transpostos. A matriz S, por fim, é uma matriz diagonal contendo os valores singulares de X, que são a raiz quadrada dos seus autovalores. A decomposição de valores singulares possui diversas aplicações, uma vez que trabalha com os autovalores e autovetores de uma matriz, que carregam muita informação sobre o conjunto de dados por meio dela expresso (Lay, 2007).

No contexto de LSA, é possível identificar termos que estão inter-relacionados e seus correspondentes documentos associados. Da mesma forma, permite verificar documentos com ideias próximas e seus termos correspondentes. Neste sentido, os fatores dali gerados podem ser vistos como tópicos, ou seja, como as principais temáticas presentes em um determinado conjunto de documentos analisados. Assim, pode-se trabalhar com o LSA alternativamente a técnicas como análise de conteúdo, de forma mais precisa e objetiva (Kulkarni, Apte & Evangelopoulos, 2014).

Desta forma, o modelo LSA já foi aplicado em outros contextos. Pode-se encontrar aplicações em diferentes áreas que trabalham com dados de texto, como revisões de literatura, auxiliando a identificar as principais temáticas dentro de um campo de pesquisa (Kulkarni *et al.*, 2014). Também este modelo já foi utilizado como forma de reunir conhecimento sobre determinado assunto a partir de dados de mercado, como anúncios de

emprego (Debortoli, Muller & Brocke, 2014) ou seguindo tendências de pesquisa ao longo do tempo (Ashton, Evangelopoulos & Prybutok, 2014). Mais do que resumir, o LSA permite explorar um grande conjunto de dados, o que o posiciona não apenas como forma de minerar textos, mas sim como forma de analisar textos (*Text Analysis*).

## 4. MÉTODO

Nesta seção, estão descritos todos os procedimentos realizados, bem como algumas escolhas metodológicas e suas justificativas. Assim como em processos de mineração de dados, toda mineração de texto precisa iniciar com um procedimento adequado de coleta, limpeza e filtragem dos dados, de forma a resultar em uma matriz de termos e documentos fidedigna aos dados. Para os procedimentos de limpeza e filtragem dos dados, bem como para a criação e manipulação da matriz de termos e documentos, foi utilizado um *script* automatizado em R, por meio do *software* RStudio. R é uma linguagem de *software open-source*, que permite trabalhar com uma grande quantidade de ferramentas, métodos e técnicas (Zhao, 2013). O R é comumente utilizado na academia e na indústria, já sendo considerado uma das principais ferramentas para se trabalhar com análise e mineração de dados (Computerworld, 2015).

O primeiro procedimento foi a coleta dos dados, realizada por meio do *site* da câmara dos deputados (Portal da Câmara dos Deputados, 2016). A votação utilizada como modelo em questão está disponível transcrita, por deputado, e também no YouTube. De forma a garantir a veracidade dos dados, a coleta foi realizada a partir da transcrição disponível no *site* da Câmara por um integrante do grupo de pesquisa, e conferida a partir da gravação disponível no YouTube por outro integrante.

Os dados coletados caracterizam um *corpus* relevante para a presente pesquisa. Ao contrário do que outras votações na Câmara dos Deputados, a votação do *impeachment* contou com uma presença expressiva dos deputados (511 de 513). Para um parâmetro comparativo, as sessões com votação exigem um mínimo de 257 parlamentares, sendo que a presença neste dia foi quase o dobro do mínimo. A natureza da matéria e a ampla cobertura da mídia possivelmente foram os motivos pelos quais o voto não foi um mero "sim" ou "não": muitos deputados apresentaram diversos motivos pelos quais estavam expressando sua opinião, o que proporcionou um discurso que refletiu um posicionamento político perante a matéria.

O segundo procedimento foi a remoção de caracteres especiais e acentos, sendo necessário o desenvolvimento de uma função "removeacento". Exemplos de caracteres especiais removidos são &, *, #, entre outros. Considera-se que estes caracteres podem ter sido resultantes de erros de registro ou de coleta. Após, o terceiro procedimento foi a unificação do formato dos discursos para letras minúsculas e remoção de números, permitindo uma análise agregadora, uma vez que as ferramentas de mineração de texto são *case sensitive*, ou seja, tratam caracteres – e por consequência termos – maiúsculos e minúsculos de forma diferente. Por fim, o quarto procedimento dentro da fase de limpeza e filtragem dos dados foi a remoção de *stopwords*, palavras com alta frequência na base de dados, porém sem valor de significado. Tal procedimento seguiu a recomendação na literatura de remoção das *stopwords* de forma a melhorar o processo de análise de dados em texto (Altszyler, Ribeiro, Sigman & Slezak, 2017; Manning, Rhagavan & Schutze, 2009). A remoção das *stopwords* foi realizada com base em um dicionário do pacote tm, do R, de forma a manter o padrão e permitir a replicabilidade do estudo. A Figura 2 apresenta alguns exemplos das *stopwords* removidas.

| | | | |
|---|---|---|---|
| porque | nas | em | nada |
| de | para | com | são |
| no | o/a | ou | sob |

**Figura 2.** Exemplo de *stopwords* removidas
Fonte: Dados da pesquisa

Após a limpeza e a filtragem dos dados, resultou-se em uma base com 511 discursos válidos (dado que houve 02 ausências), contendo ainda o partido, o voto e o estado de cada deputado. O total de termos únicos computados foi de 3.353. A partir destes, uma matriz de termos e documentos foi criada. A ponderação desta matriz (o valor numérico constante em cada das células $a_{ij}$) pode ser realizada por meio da frequência dos termos (TF, do inglês *term frequency*) ou por meio da composição da frequência dos termos (TF) com o inverso da quantidade de documentos que possuem o termo (IDF, do inglês *inverse document frequency*) (Manning, Rhagavan & Schutze, 2009), sendo este último o índice adotado neste estudo.

Quanto ao índice TF, existe uma crítica devido ao seu baixo poder discricionário: há uma alta probabilidade de que em uma coleção de documentos sobre a indústria automobilística, por exemplo, a palavra "carro" apareça em todos os documentos. Porém, usualmente os pesquisadores conhecem previamente, ainda de forma superficial, o assunto de qual o *corpus* em questão trata. Portanto o índice TF acabaria por sobrevalorizar este tipo de termo, o que pode não auxiliar para a análise.

Assim, o índice mais utilizado é o TF-IDF (Equação 01). O índice TF-IDF considera uma composição da frequência do termo (TF) com o inverso da frequência dos documentos que possuem o termo (IDF). Dessa forma, é possível valorizar termos raros, que diferenciam os documentos, e que, quando ocorrem, possuem uma frequência significativa perante os demais. Além disso, este índice desvaloriza termos que ocorrem em muitos documentos, penalizando-os inclusive com o valor 0 quando ocorrem virtualmente em todos os documentos (Altszyler *et al.*, 2017; Crain *et al*., 2012).

$$(tf - idf)_{t,d} = tf_{t,d} \times log \frac{N}{df_t} \quad (01)$$

Em seguida aos procedimentos de filtragem e limpeza, a matriz de termos e documentos foi gerada no R, utilizando como parâmetro a ponderação pelo índice TF-IDF. Após, esta matriz foi utilizada como *input* para criação do espaço semântico latente, por meio da biblioteca "LSA" do R. A ideia de criar um espaço semântico é permitir a representação da estrutura dos textos, conjuntamente, no formato de vetores (através dos autovalores e dos autovetores associados), de forma a conseguir descobrir relações para além da semelhança entre os caracteres das palavras (Valle-Lisboa & Mizraji, 2007).

Seguindo a documentação da biblioteca LSA (Wild, 2015), foi definido o cálculo da decomposição em relação à retenção dos valores singulares (ou seja, as raízes dos autovalores), objetivando a redução de dimensionalidade dado o LSA. O parâmetro, chamado *dimcalc*, permite definir como serão escolhidos os valores singulares a serem retidos. A definição da quantidade de valores singulares retidos para posterior análise é um problema em aberto na literatura (Efron, 2007; Ashton *et al.,* 2014), portanto para definição da retenção foram testados alguns dos parâmetros disponíveis. A escolha foi por permanecer com a definição do parâmetro que trouxe os autovalores correspondentes à melhor relação custo-benefício, uma vez que a adição de novos autovalores (e por consequência novos vetores no espaço semântico) acabava por agregar marginalmente, em relação ao custo de inclusão de uma nova dimensão. A Figura 3 apresenta a relação gráfica dos valores singulares, demonstrando a participação marginal daqueles na ponta da cauda.

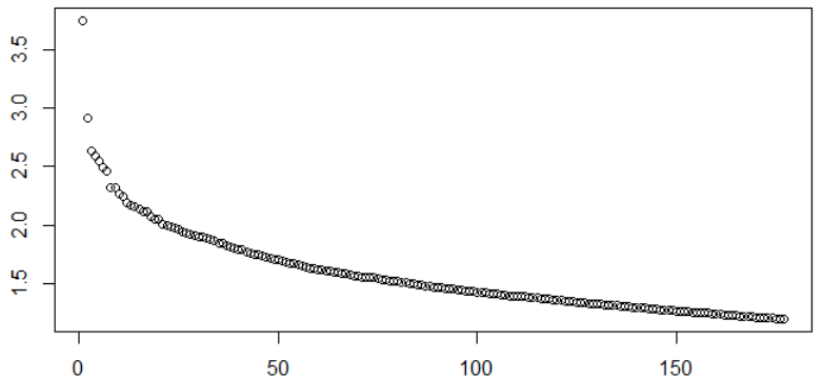


**Figura 3.** Valores singulares da Matriz X
Fonte: Dados da pesquisa

Assim, o resultado após a criação do espaço latente foram as três matrizes, U, S e $V^T$ (Equação 02). Estas representam, de forma aproximada, a matriz X original, de termos e documentos, ponderadas pelo índice TF-IDF. A matriz U contém, nas linhas, os termos resultantes após todos os procedimentos aqui descritos, e nas colunas, os fatores associados a cada termo. A matriz S contém os valores singulares na diagonal e zeros. Já a matriz $V^T$, por fim, contém os documentos nas linhas e os fatores nas colunas já no seu formato transposto. Os

valores dos fatores são os responsáveis por conectar os termos com os documentos, por conterem os autovetores, à esquerda e à direita, da matriz X.

$$X \cong U.S.V^T \quad (02)$$

A próxima seção apresenta a análise dos resultados a partir da matriz U, contendo os autovetores dos termos em relação aos principais fatores (ou tópicos) encontrados. A escolha de analisar a matriz U se deve ao objetivo do trabalho, que foi de explorar o poder de síntese de um grande volume de dados em texto em alguns principais tópicos por meio do seu conjunto de termos.

## 5. RESULTADOS

Para melhor leitura dos dados, apresentam-se aqui alguns dados quantitativos importantes em relação à estrutura política que derivou o *corpus* de documentos trabalhados. Primeiro, a Figura 4 apresenta a distribuição de partidos em relação aos presentes – ou seja, aqueles cujo discurso está contido nos dados analisados. Percebe-se que o partido com maior quantidade de representantes foi o PMDB (66 votantes presentes), seguido pelo PT (60 votantes presentes).

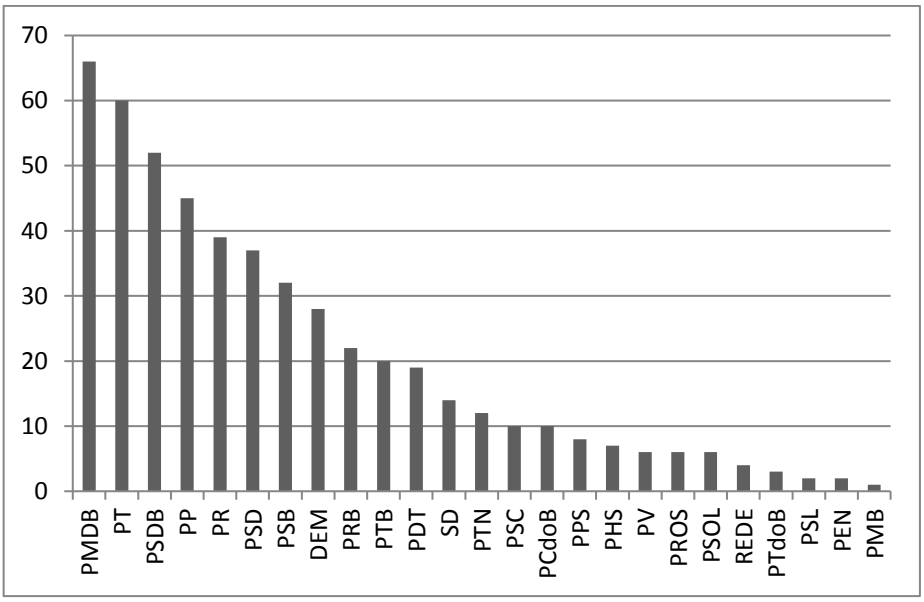


**Figura 4.** Distribuição de partidos na base de dados
Fonte: Base de dados

Tal diversidade entre os partidos com posição institucional divergente representa uma característica importante da base. Como os textos estão sendo analisados de forma integrada, caso houvesse uma participação majoritária de um mesmo partido, a análise poderia ser prejudicada na sua capacidade de discricionariedade. É importante citar, ainda, que houve diferença entre a posição institucional do partido e o voto dos respectivos deputados. Mesmo assim, nos partidos citados, não houve ocorrência de tal diferença (todos votantes apresentaram a mesma opção dos seus partidos).

Na Tabela 1 apresenta-se a distribuição de votos em relação a cada partido. Pela natureza delicada da matéria, destaca-se que existem textos de 511 dos 513 parlamentares em exercício na época, o que traz para o *corpus* uma variedade importante de posicionamentos. Estão destacados os 12 partidos cujos deputados votaram todos da mesma maneira. Destaca-se, além da homogeneidade da votação, que estes partidos representam juntos aproximadamente 43% do total de votos.

Tabela 1.
**Relação de voto por partido**

| Partidos | Sim | Não | Abstenção | Ausente | Total Presentes | %, Sim | %, Não |
|---|---|---|---|---|---|---|---|
| *PSDB* | *52* | *0* | *0* | *0* | *52* | *100%* | *0%* |
| *DEM* | *28* | *0* | *0* | *0* | *28* | *100%* | *0%* |
| *PRB* | *22* | *0* | *0* | *0* | *22* | *100%* | *0%* |
| *SD* | *14* | *0* | *0* | *0* | *14* | *100%* | *0%* |
| *PSC* | *10* | *0* | *0* | *0* | *10* | *100%* | *0%* |
| *PPS* | *8* | *0* | *0* | *0* | *8* | *100%* | *0%* |
| *PV* | *6* | *0* | *0* | *0* | *6* | *100%* | *0%* |
| *PSL* | *2* | *0* | *0* | *0* | *2* | *100%* | *0%* |
| *PMB* | *1* | *0* | *0* | *0* | *1* | *100%* | *0%* |
| PSB | 29 | 3 | 0 | 0 | 32 | 90,63% | 9,38% |
| PMDB | 59 | 7 | 0 | 1 | 66 | 89,39% | 10,61% |
| PHS | 6 | 1 | 0 | 0 | 7 | 85,715 | 14,29% |
| PP | 38 | 4 | 3 | 0 | 45 | 84,44% | 8,89% |
| PSD | 29 | 8 | 0 | 0 | 37 | 78,38% | 21,62% |
| PTB | 14 | 6 | 0 | 0 | 20 | 70% | 30,00% |
| PR | 26 | 10 | 3 | 1 | 39 | 66,67% | 25,64% |
| PTN | 8 | 4 | 0 | 0 | 12 | 66,67% | 33,33% |
| PROS | 4 | 2 | 0 | 0 | 6 | 66,67% | 33,33% |
| PTdoB | 2 | 1 | 0 | 0 | 3 | 66,67% | 33,33% |
| REDE | 2 | 2 | 0 | 0 | 4 | 50% | 50% |
| PEN | 1 | 1 | 0 | 0 | 2 | 50% | 50% |
| PDT | 6 | 12 | 1 | 0 | 19 | 31,58% | 63,16% |
| *PT* | *0* | *60* | *0* | *0* | *60* | *0%* | *100%* |
| *PCdoB* | *0* | *10* | *0* | *0* | *10* | *0%* | *100%* |
| *PSOL* | *0* | *6* | *0* | *0* | *6* | *0%* | *100%* |
| ***Total*** | ***367*** | ***137*** | ***7*** | ***2*** | ***511*** | | |

Fonte: Dados da pesquisa

Após o pré-processamento descrito na seção anterior, o total de palavras na base, considerando repetições, foi de 131.410 termos. O discurso mais longo consistiu de 894 palavras, o que representa menos de 1% do total da base. Ainda assim, a parcela de 80% dos termos foi fruto do discurso de 300 deputados, o que representa que grande parte das palavras foi dita por mais da metade dos presentes. Percebe-se assim que, apesar das diferenças no tamanho de cada um dos discursos, o *corpus* utilizado é consistente com a lei de Zipf (Zipf, 1949), que atesta que se utiliza um pequeno conjunto de palavras para expressar a maioria das ideias, independente do idioma. A Figura 5 apresenta a distribuição dos termos em relação à quantidade de deputados.

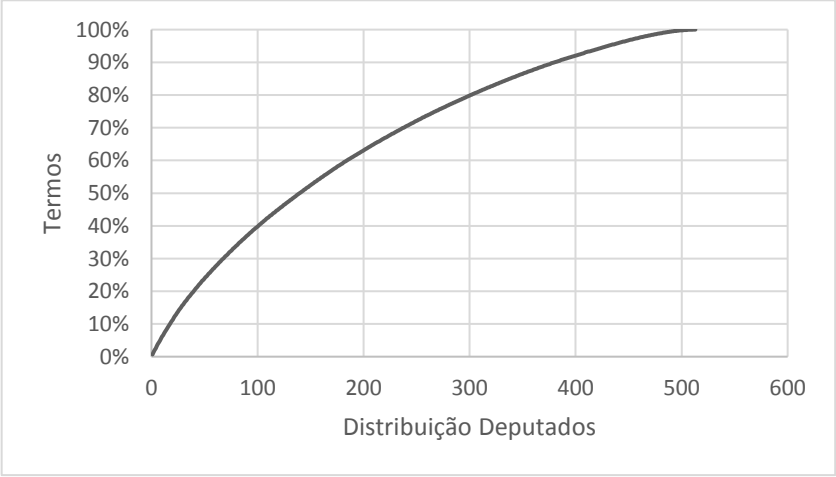


**Figura 5.** Distribuição Deputados x % Termos
Fonte: Base de Dados

Dado a pauta em questão quando da votação, uma primeira exploração dos dados foi em relação aos termos mais fortemente associados com a palavra “corrupção”, por meio de uma análise de correlação diretamente na matriz de termos e documentos. A medida de correlação identifica a força da relação linear entre os termos, variando de 0 a 1 (Hair *et al*., 2010). A Figura 6 apresenta as palavras cujo índice de correlação com o termo corrupção era maior ou igual a 0,2, em ordem decrescente. Alguns termos representam uma aversão, relatada no discurso, à corrupção, como “combate”, “contra” e “caçar”. Outras significam repúdio, como “ridícula” e “envergonha”. Há ainda termos interessantes, como “ladroagem”, “DNA” e “revanchista”.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.41 | combate | 0.25 | instante | 0.23 | consequencia | 0.23 | rasgar |
| 0.35 | administrativo | 0.25 | espaços | 0.23 | contribuir | 0.23 | representada |
| 0.35 | araujo | 0.25 | nazare | 0.23 | convidar | 0.23 | revanchista |
| 0.35 | campina | 0.24 | aceita | 0.23 | detras | 0.23 | ridicula |
| 0.35 | coalizao | 0.24 | assassinado | 0.23 | DNA | 0.23 | reletiva |
| 0.35 | combatia | 0.24 | direita | 0.23 | enriquecer | 0.23 | rransforma |
| 0.35 | decidisse | 0.24 | implantado | 0.23 | envergonha | 0.22 | chama |
| 0.35 | inteira | 0.24 | modelo | 0.23 | estara | 0.21 | paraiba |
| 0.35 | segmentos | 0.24 | pede | 0.23 | grades | 0.21 | tanta |
| 0.35 | unir | 0.24 | pedimos | 0.23 | legislatura | 0.21 | catarina |
| 0.35 | vereador | 0.24 | daria | 0.23 | logo | 0.21 | duas |
| 0.32 | varios | 0.23 | acham | 0.23 | Moro | 0.21 | ladroagem |
| 0.31 | contra | 0.23 | anualmente | 0.23 | niveis | 0.21 | projeto |
| 0.3 | combater | 0.23 | beneficiario | 0.23 | poeiras | 0.21 | salvar |
| 0.27 | impunidade | 0.23 | caçar | 0.23 | pagando | 0.2 | contra |
| 0.26 | combate | 0.23 | capitalismo | 0.23 | permanecer | 0.2 | poder |
| 0.26 | filhas | 0.23 | categoria | 0.23 | policial | | |

**Figura 6.** Termos correlacionados com a palavra “corrupção”
Fonte: Dados da pesquisa

Com este primeiro olhar abrangente, foi possível perceber que muitos termos eram distantes da discussão do mérito quando da votação. A palavra com maior correlação, “combate”, demonstra que havia na declaração do voto dos deputados um posicionamento de enfrentamento, e que era importante a expressão nesse sentido. Para ampliar essa percepção, a próxima análise partiu da criação do espaço vetorial a partir do LSA, no qual foi possível identificar um conjunto de tópicos formados por termos que os descrevem. A identificação dos tópicos no LSA tem um papel importante na relação semântica das palavras. Ao minimizar significativamente os efeitos de polissemia e sinonímia, o LSA permite compreender, para além de uma análise léxica, como as palavras se relacionam nos textos do *corpus*. Assim, as diferentes ideias são reveladas, contendo as principais palavras que formam o seu conceito. Por trabalhar com decomposição matricial a partir de uma matriz ponderada pelo TF-IDF, cada palavra possui um grau de pertencimento em relação a cada um dos autovalores da matriz, que representam a informação total dos dados. Assim, é possível uma razoável importância a respeito do grau de pertencimento daquela palavra em relação à ideia geral do tópico.

Para seleção dos tópicos a serem analisados, foram considerados aqueles contendo o próprio termo “corrupção”. A análise dos tópicos permite compreender, a partir de um grande volume de dados, aqueles termos que possuem uma relação latente, ou seja: que tendem, em relação ao seu sentido, a aparecer em conjunto de forma significativa. Dos 177 tópicos resultantes da decomposição matricial, 52 continham a palavra corrupção positivamente associada (índice acima de 0,1), demonstrando uma presença significativa deste assunto dentre os discursos dos deputados (aproximadamente 30% dos discursos).

Destes 52 tópicos, foram escolhidos os três com os maiores autovalores associados e cujos índices da palavra corrupção foram os maiores, respectivamente. As nuvens de palavras foram geradas considerando os termos responsáveis por 65% ou mais da representatividade. Assim como em outros trabalhos com LSA (Ashton

*et al.*, 2014), a definição do limite deste índice foi realizada por meio de testes empíricos para buscar o melhor ponto entre a contribuição marginal e a relevância da adição de mais termos.

A primeira nuvem, apresentada na Figura 7, traz o tópico cuja maior carga é o próprio termo "corrupção", responsável sozinho por aproximadamente 10% de representatividade nesta dimensão, dentre todos os 3.353 termos únicos. Percebe-se que as palavras "combater" e "ensinou" apresentam respectivamente os maiores índices. Há também algumas expressões curiosas, como "irmão" e memória".


**Figura 7.** Nuvem de palavras tópico 1
Fonte: Dados da pesquisa

O próximo tópico traz a palavra corrupção com uma representatividade menor (aproximadamente 4%), porém sendo o terceiro maior índice. Este tópico chama a atenção, pois trouxe de forma próxima o termo "contra" com as palavras "esquerda", "ricos" e "direita". Uma análise dos demais termos (como "pobre", "boquinha" e "ditadura") indica discursos com expressões bastante fortes e populares. A Figura 8 traz a nuvem representativa deste tópico.


**Figura 8.** Nuvem de palavras tópico 2
Fonte: Dados da pesquisa

Por fim, o tópico três (Figura 9) apresentou uma diversidade maior de termos, que compuseram 65% da explicação total, sendo 50 termos, contra a média de 25 termos dos tópicos anteriores. Este tópico apresenta os termos "votos" e "corrupção" com maior representação dentre as demais palavras. Um aspecto interessante foi que esse tópico reuniu a maior quantidade de nomes de cidade (como "Criciúma" e "Joinville"), bem como diversos nomes próprios ("Pedro", "Bianca", "Bruno", entre outros), indicando que neste conjunto de discursos houve diversas menções nominais.


**Figura 9.** Nuvem de palavras do tópico 3
Fonte: Dados da pesquisa

Analogamente, houve também outros 52 tópicos que apresentaram carga negativa para a palavra corrupção, com índice acima de 0,1. Estes tópicos apresentam conjuntos de palavras que se afastam do termo no espaço vetorial, indicando uma relação de rejeição com determinada carga. Para apresentação, foram selecionados os dois tópicos com maiores autovalores e maiores cargas negativas na palavra corrupção. Por este motivo, a palavra corrupção não está presente em nenhuma das nuvens subsequentes.

O primeiro tópico desta seção apresenta o conjunto de palavras que mais se apresentou afastado, no espaço vetorial criado após a decomposição matricial do termo “corrupção”. Para construção da nuvem, foram utilizados os mesmos parâmetros descritos anteriormente. Percebe-se, na Figura 10, que há uma forte presença de palavras religiosas, sendo os três principais termos deste tópico “caminhos”, “ilumine” e “senhor”, nesta ordem.

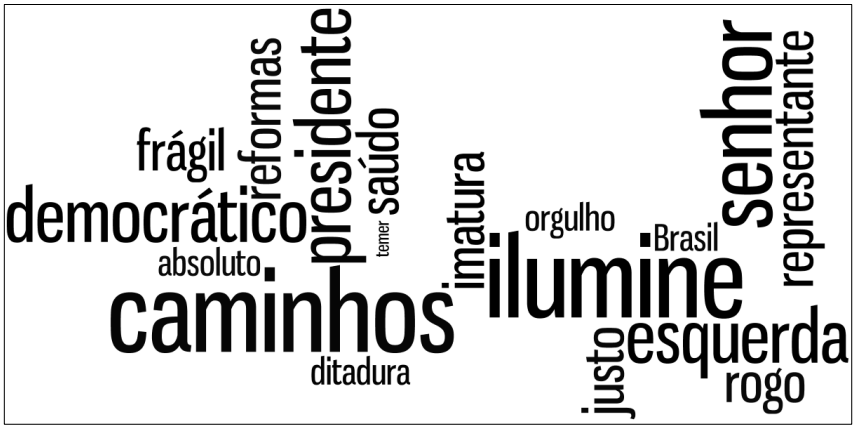


**Figura 10.** Nuvem de palavras tópico 4
Fonte: Dados da pesquisa

O outro tópico selecionado (Figura 11) apresentou a palavra “corrupção” com o 10º mais negativo índice, dentre os 3.353 termos únicos apropriados por dimensão. Destaca-se a presença conjunta dos termos “sergio” e “moro”, claramente uma indicação de citação, neste conjunto de discursos, ao juiz Sérgio Moro. Os demais termos, na sequência, foram “ditadura”, “esquerda” e “país”.


**Figura 11.** Nuvem de palavras tópico 5
Fonte: Dados da pesquisa

Os tópicos correlacionados também agruparam termos que poderiam denotar argumentos dissonantes do processo em votação. Portanto, avaliou-se a correlação de Pearson entre os discursos que eram a favor e contra o *impeachment* da presidenta, encontrando uma correlação de 0,866. A partir desse valor, pode-se verificar uma forte associação linear entre os discursos dos deputados que votaram sim e os que votaram não. Assim, reforça-se ainda mais o entendimento de que o discurso dos deputados não pode ser considerado uma base confiável para distinguir sua posição quanto ao *impeachment*. Essa não associação de discurso e votos pode estar fortemente associada a dissonâncias cognitivas que os deputados possuem em um momento de tomada de decisão relevante, uma vez que a dissonância se caracteriza por esse conflito cognitivo ao qual passa o tomador de decisão (Festinger, 1957).

A seguir, apresenta-se a Tabela 2, que apresenta os 50 termos mais frequentes nos discursos entre os deputados que votaram a favor e contra e os 50 termos mais frequentes que são específicos do voto a favor ou contra. Apesar de os discursos que votaram a favor serem mais numeroso dos que os que votaram contra (367 e 144, respectivamente), a média de palavras por voto não foi muito distinta (52,29 para votos "sim" e 58,31 para votos "não"), o que demonstra que ambos grupos utilizaram uma quantidade de termos muito próxima para expor seus argumentos.

Tabela 2.
**50 palavras mais frequentes**

| **Apenas Sim** | | | **Apenas Não** | | | **Sim ∩ Não** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Palavra** | **Freq** | **%** | **Palavra** | **Freq** | **%** | **Palavra** | **Freq Sim** | **%** | **Freq Não** | **%** |
| Sim | 386 | 6,58% | Democracia | 95 | 1,62% | Voto | 383 | 6,53% | 154 | 2,63% |
| Família | 106 | 1,81% | Golpe | 87 | 1,48% | Presidente | 368 | 6,27% | 120 | 2,05% |
| Esperança | 51 | 0,87% | Defesa | 46 | 0,78% | Senhor | 303 | 5,17% | 85 | 1,45% |
| Deus | 50 | 0,85% | Crime | 32 | 0,55% | Brasil | 270 | 4,60% | 46 | 0,78% |
| Partido | 47 | 0,80% | Cunha | 30 | 0,51% | Povo | 192 | 3,27% | 54 | 0,92% |
| Governo | 45 | 0,77% | Trabalhadores | 26 | 0,44% | Estado | 134 | 2,28% | 23 | 0,39% |
| Cidade | 44 | 0,75% | Processo | 24 | 0,41% | País | 124 | 2,11% | 36 | 0,61% |
| Querida | 43 | 0,73% | Eduardo | 24 | 0,41% | Nome | 104 | 1,77% | 30 | 0,51% |
| Futuro | 42 | 0,72% | Temer | 23 | 0,39% | *Impeachment* | 92 | 1,57% | 38 | 0,65% |
| Momento | 41 | 0,70% | Presidenta | 21 | 0,36% | Aqui | 84 | 1,43% | 39 | 0,66% |
| Filhos | 41 | 0,70% | Luta | 19 | 0,32% | Brasileiro; brasileiros | 130 | 2,22% | 30 | 0,51% |
| Rio | 40 | 0,68% | Ruas | 19 | 0,32% | Todos | 74 | 1,26% | 16 | 0,27% |
| Senhoras | 39 | 0,66% | Porque | 16 | 0,27% | Respeito | 68 | 1,16% | 38 | 0,65% |
| Minas | 36 | 0,61% | Ser | 15 | 0,26% | Contra | 62 | 1,06% | 86 | 1,47% |

(Continua)

(Conclusão)

| Apenas Sim | | | Apenas Não | | | Sim ∩ Não | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Palavra** | **Freq** | **%** | **Palavra** | **Freq** | **%** | **Palavra** | **Freq Sim** | **%** | **Freq Não** | **%** |
| Paulo | 35 | 0,60% | Vai | 15 | 0,26% | Dilma | 56 | 0,95% | 46 | 0,78% |
| Grande | 33 | 0,56% | Deputado | 14 | 0,24% | Deputados | 52 | 0,89% | 18 | 0,31% |
| Também | 32 | 0,55% | Mulher | 14 | 0,24% | Hoje | 49 | 0,84% | 13 | 0,22% |
| Favor | 31 | 0,53% | Vida | 14 | 0,24% | Corrupção | 46 | 0,78% | 16 | 0,27% |
| Nação | 31 | 0,53% | Michel | 13 | 0,22% | Senhores | 46 | 0,78% | 17 | 0,29% |
| Gerais | 31 | 0,53% | Querem | 13 | 0,22% | Quero | 42 | 0,72% | 27 | 0,46% |
| Milhões | 28 | 0,48% | Deste | 13 | 0,22% | Casa | 39 | 0,66% | 27 | 0,46% |
| Anos | 28 | 0,48% | História | 13 | 0,22% | Neste | 38 | 0,65% | 15 | 0,26% |
| **Total** | **1260** | **21,48 %** | Brasileira | 13 | 0,22% | Responsabilidade | 35 | 0,60% | 15 | 0,26% |
| | | | **Total** | **599** | **10,21 %** | Homenagem | 32 | 0,55% | 28 | 0,48% |
| | | | | | | Constituição | 32 | 0,55% | 40 | 0,68% |
| | | | | | | Dizer | 29 | 0,49% | 16 | 0,27% |
| | | | | | | Estão | 29 | 0,49% | 20 | 0,34% |
| | | | | | | **Total** | **2913** | **49,67%** | **1093** | **18,64%** |

Fonte: Dados da pesquisa

Entendendo que a pauta em questão tratava de ações autorizadas pela presidência da república não previstas na legislação, que implicariam em irregularidade fiscal, seria esperado que os argumentos dos deputados pudessem denotar as razões pelas quais tais atos poderiam, ou não, ser enquadrados como crime de responsabilidade. Percebe-se, pelo terceiro bloco da Tabela 2, que há diversos termos conectados com possíveis argumentos, como "corrupção", "responsabilidade", "contra" e "constituição". Porém este bloco se refere aos termos mais utilizados em ambos os votos, o que pode indicar que razões semelhantes foram utilizadas para expressar posições contrárias.

Com base nos termos exclusivos dos votos de cada lado, buscou-se perceber se havia alguma dissonância cognitiva em relação à pauta em votação. Percebe-se, porém, que os termos não refletem motivos conectados com o julgamento de um crime de responsabilidade fiscal. No primeiro bloco da Tabela 2, que contém os termos exclusivos dos votos "sim", verifica-se que a família foi um argumento evocado, por meio das palavras "família" e "filhos", que se conectam com "momento" e "futuro". Outro argumento utilizado foram os estados federativos, por meio dos termos "rio", "paulo" e "minas", que fazem referência aos estados do Rio de Janeiro, São Paulo e Minas Gerais. Ilustram-se esses argumentos a partir dos trechos transcritos a seguir:

> *"Senhor Presidente, em respeito a minha mulher, aos meus filhos e aos meus netos, pelo povo do querido estado do Pará, por um futuro melhor para o Brasil, eu voto sim."*
>
> *"Senhor Presidente, senhoras e senhores deputados, eu voto aqui hoje a favor das nossas crianças, da nossa juventude, das nossas famílias, e da minha Paraíso, do meu sul de Minas."*

Movimento semelhante pode ser percebido no segundo bloco do Quadro 3, que apresenta as palavras exclusivas dos votos contrários à abertura do processo de *impeachment*. Nota-se que os termos "cunha", "eduardo", "michel" e "temer" fazem alusão a figuras políticas da época, e que têm um peso significativo com argumento dos votos "não". Ainda, expressões como "golpe", "trabalhadores" e "luta" parecem caracterizar mais um posicionamento político do que uma avaliação sobre o tema em questão. A seguir, destacam-se dois trechos que ilustram esse ideia:

> *"Eu não reconheço a legitimidade de Eduardo Cunha para presidir esse processo, não reconheço legitimidade de o conspirador Michel Temer para presidir esse país. Não acredito*

> *em político demagogo, que fala em combater a corrupção e se alia com corruptos, aquilo que de pior o Brasil tem. Contra o golpe, contra os golpistas, eu voto não."*
>
> "Eu quero falar em nome da democracia, em homenagem a todos os que estão nas redes sociais e nas ruas, lutando pela democracia e contra o golpe, que voto contra o golpe, contra os golpistas, contra Eduardo Cunha, contra Michel Temer."

Portanto, há indícios da presença de dissonâncias cognitivas entre a decisão expressa (sim ou não) e o entendimento dos deputados sobre o tema expresso no discurso, tendo em vista a não possibilidade de caracterização do voto do deputado como a favor ou contra o impeachment a partir do seu discurso. Assim, vale destacar que essa fala dos deputados representa uma tomada de decisão norteada, conforme Weiss (1983), por informação, ideologia e interesses. Nesse sentido, cabe destacar ainda que, mesmo com discursos homogêneos e, portanto, contraditórios, o tomador de decisão, seguindo Kladis e Freitas (1996) e Festinger (1957), o faz tendo em vista privilegiar a alternativa escolhida.

## 6. CONSIDERAÇÕES FINAIS

Este trabalho buscou explorar uma análise de dados em texto de forma alternativa a outras análises mais tradicionais, buscando desmistificar e ampliar as possibilidades do uso de dados em texto. Além de altamente disponíveis, os dados em texto têm origem diretamente do interlocutor, seja um político (como neste artigo), um consumidor, uma organização ou um órgão governamental. Portanto, tais dados possuem informações ricas, uma vez que passam por pouco ou nenhum processamento, diferente de dados tradicionalmente utilizados em ferramentas de descoberta de conhecimento.

Percebeu-se, após essa compilação de resultados, que houve muitos tópicos que agruparam palavras antagônicas, como "esquerda" e "direita", "ricos" e "pobres", "ditadura" e "democrática". Dentre os termos com maior frequência entre os deputados que votaram "sim" e "não", houve termos semelhantes e bastante frequentes, proporcional à quantidade de votos, como "democracia", "povo", "Brasil", "defesa" e "constituição". Essas indicações podem ser compreendidas como uma baixa discricionariedade entre os textos, mesmo ao separar os grupos.

Pode-se compreender, a partir destes resultados, que o discurso dos deputados não é um elemento com capacidade de prever seu voto, o que poderia indicar que os discursos são inadequados para descrever a opinião e o posicionamento dos políticos. Tal fato demonstra-se preocupante, pois significa que, para conhecer um candidato e votar com mais consciência, de forma a evitar mais casos de corrupção, apenas compreender e conhecer seus discursos não é eficiente. Assim, o discurso dos deputados e, portanto, a escolha tomada por eles, não necessariamente está relacionada à racionalização, conforme destacado por Weiss (1983). Nesse sentido, é possível destacar nesse recorte de tomada de decisão dos políticos brasileiros a presença de outros componentes desse processo de decisão política destacados por Weiss (1983), que são informação, ideologia e interesses.

Além disso, cabe ressaltar que não é possível identificar o voto de um deputado pelo seu discurso, o que pode ser um indício de dissonâncias cognitivas existentes entre a decisão tomada pelo deputado e seu entendimento sobre o tema. Entretanto, mesmo com essas dissonâncias, o tomador da decisão explicita sua escolha e outras informações que, por vezes, podem ser argumentos contraditórios à sua tomada de decisão, mas que são argumentos que podem indiretamente resultar em apoio dos eleitores. Portanto, mesmo com discursos contraditórios, o tomador de decisão está utilizando informações que privilegiem a alternativa escolhida, conforme destacado por Kladis e Freitas (1996) e Festinger (1957).

Como possibilidade de pesquisa futura, sugere-se procurar confirmar esta conclusão a partir de expansão dos textos ao longo do tempo, uma vez que a constituição da própria base pode ser elencada como uma das limitações deste trabalho. Ao longo de sua exploração, percebeu-se que o discurso composto de poucas palavras pode limitar a análise mais completa de similaridade. Assim, sugere-se, conforme Dumais e Nielsen (1992) e Zelikowitz e Hirsh (2001), que seja realizada uma expansão dos textos e uma nova análise, verificando as diferenças entre usar textos mais longos ou mais curtos para a tarefa em questão.

Outra possibilidade de pesquisa futura, já em andamento, é compreender a distribuição destes argumentos ao longo do tempo. Com uma base de dados mais completa e indexada, pode-se compreender como

os tópicos se distribuem ao longo do tempo, identificando tendências e permitindo, possivelmente, entender o posicionamento de similaridade entre os mesmos indivíduos dentro, por exemplo, do período do mandato.

Como contribuição prática, este artigo possibilitou a criação de um *script* que automatiza todos os tratamentos, conforme as boas práticas indicadas na literatura de mineração de texto (Manning, Rhagavan & Schutze, 2009), disponível a partir de solicitação ao primeiro autor. O avanço de técnicas de Analytics tem posicionado as empresas com uma orientação forte aos dados, e ser capaz de trabalhar com textos em volume pode fornecer ao gestor interessantes *insights* em diferentes bases. Com isto, este trabalho busca também desmistificar e procurar acessar o valor de ferramentas e técnicas que permitam pesquisadores e executivos analisarem um grande volume de dados não estruturados por uma perspectiva objetiva e automatizada, aproximando a perspectiva de negócios ao uso e ao desenvolvimento de ferramentas desta natureza, ponto ainda passível de maior desenvolvimento (Popovič *et al*., 2018; Cheng & Zhang, 2014). Além disso, foi possível perceber a carência de ferramentas voltadas para análise de texto em português, especialmente em relação ao *stemming*. *Stemmers* são algoritmos que reduzem o termo ao seu radical, conforme regras gramaticais (Orengo & Huyck, 2001). São poucas implementações desenvolvidas para a língua portuguesa, o que pode limitar o uso de gestores brasileiros, especialmente. Assim, como pesquisa futura, planeja-se trabalhar agregando este conceito, de forma a aprimorar ainda mais os possíveis resultados.

## REFERÊNCIAS

Aggarwal, C. C., & Zhai, C. X. (2012). *Mining Text Data*. Berlin: Springer.

Altszyler, E., Ribeiro, S., Sigman, M., & Slezak, D. F. (2017). The interpretation of dream meaning: Resolving ambiguity using Latent Semantic Analysis in a small corpus of text. Consciousness and cognition, 56, 178-187.

Amaral, M.S., & Pinho, J.A.G. (2016). Tuitando por Votos: Congressistas Brasileiros e o Uso do Twitter nas Eleições de 2014. In Anais do XL Encontro da Associação Nacional de Pós-Graduação e Pesquisa em Administração. Costa do Sauípe: *ENANPAD*.

Ashton, T., Evangelopoulos, N., & Prybutok, V. (2014) Extending monitoring methods to textual data: a research agenda. *Quality & Quantity Journal*, *48*, pp. 2277-2294.

Bawden, D., & Robinson, L. (2009). The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science*, *35* (2), 180-191.

Braga, M. S. S. (2013). A Agenda dos Estudos sobre Partidos Políticos e Sistemas Partidários no Brasil. *Revista de Discentes de Ciência Política da UFSCAR*, 1(1), 1-25.

Chen, P., & Zhang, C. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, *Information Sciences*, 275, pp. 314-347.

Coelho, C., Terrasêca, M., & Correia, J. A. (2016). Tipos de expertise na tomada de decisão política em educação: Contribuições de uma discussão teórico-concetual. *Arquivos Analíticos de Políticas Educativas*, 24(3). http://dx.doi.org/10.14507/epaa.v24.2103

Computerworld. (2015). *6 R Resources to improve your data skills*. Disponível em http://www.computerworld.com/article/2497464/business-intelligence/business-intelligence-60-r-resources-to-improve-your-data-skills.html.

Crain, S. P., Zhou, K., Yang, S., & Zha, H. (2012). Dimensionality Reduction and Topic Modeling. In Aggarwal, C. C., & Zhai, C. X. (Eds), *Miningr Text Data,* (pp. 129-156). Berlin: Springer.

Debortoli, S., Junglas, I., Muller, O., & Brocke, J. (2016) Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems*, forthcoming. Recuperado em: https://www.researchgate.net/profile/Jan_vom_Brocke/publication/294406698Text_Mining_for_Information_Systems_Researchers_An_Annotated_Topic_Modeling_Tutorial/links/56c3351208ae8a6fab59f478.pdf.

Debortoli, S., Muller, O.; & Brocke, J. (2014). Comparing Business Intelligence and Big Data Skills: a Text Mining Study Using Job Advertisements. *Business & Information Systems Engineering, 5,* pp. 289-300.

Deerwester, S., Dumais, S.T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science, 41* (6), 391-407.

Dumais, S., & Nielsen, J. (1992). Automating the Assignment of Submitted Manuscripts to Reviewers, in 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Efron, M. (2007). Eigenvalue-Based Model Selection During Latent Semantic Indexing. *Journal of the American Society for Information Science and Technology, 56* (9), 969-988.

Eppler, M. J., & Mengis, J. (2004). The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society*, *20*, pp. 325-344.

Festinger, L. (1962). A theory of cognitive dissonance. California: Stanford University Press.

Freitas, H., Becker, J. L., Hoppen, N., & Kladis, C. M. (1997). *Informação e decisão*: Sistemas de apoio e seu impacto. Porto Alegre: Ortiz, 74.

Freitas Junior, J. C. S, Maçada, A. C. G., Oliveira, M., Brinkhues, R. A. (2016). Big Data e Gestão do Conhecimento: Definições e Direcionamentos de Pesquisa. *Revista Alcance*, 23 (4), 529-546.

Hair, J. F., Anderson, R. E., Babin, B. J., & Black, W. C. (2010). Multivariate data analysis: A global perspective (Vol. 7). Upper Saddle River, NJ: Pearson.

Kladis, C. M., & Freitas, H.M.R. O gerente nas organizações: funções, limitações e estilos decisõesrios. *Revista Ser Humando***,** 109.

Kulkarni S. S.; Apte, U. M., & Evangelopoulos, N. E. (2014). The Use of Latent Semantic Analysis in Operations Management Research. *Decision Sciences*, *45* (5), 971-993.

Landauer, T. K. (2011). LSA as a Theory of meaning. In Landauer, T. K.; McNamara, D. S.; Dennis, S. & Kintsch, W. (Eds), *Handbook of Latent Semantic Analysis*, (pp. 3-34). New York: Routledge.

Lay, D. (2007). *Linear Algebra and its applications*. São Paulo: LTC.

Mainwaring, S. P. (2001). *Sistemas Partidários em Novas Democracias*: o caso do Brasil. Rio de Janeiro: Editora FGV.

Manning, C. D., Rhagavan, P., & Schutze, H. (2009). *An Introduction to Information Retrieval*. Cambrigde: Cambridge University Press.

Marchand, D. A. & Peppard, J. (2013). Why IT fumbles analytics. *Harvard Business Review*, *91* (1), 104–112.

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana* (Vol. 1). São Paulo: Campus.

McAfee, A., & Brynjolfsson, E. (2012, October). Big Data: The Management Revolution. *Harvard Business Review*, pp. 60-68.

Orengo, V. M., & C.R. Huyck. (2001) A Stemming Algorithm for the Portuguese Language. *Proceedings of the 8th International Symposium on String Processing and Information Retrieval (SPIRE)*. Laguna de San Raphael, Chile. pp. 183-193

Pinho, J. A. G. (1998). Reforma do Aparelho do Estado: limites do gerencialismo frente ao patrimonialismo. *Organizações e Sociedade*, 12(5), 59-79.

Popovič, A., Hackney, R., Tassabehji, R., & Castelli, M. (2018). The impact of big data analytics on firms' high value business performance. *Information Systems Frontiers*, 20(2), 209-222.

Portal da Câmara dos Deputados. http://www2.camara.leg.br/. Acesso em Maio 2016.

Valle-Lisboa, J. C., & Mizraji, E. (2007). The uncovering of hidden structures by Latent Semantic Analysis, *Information Sciences 177*, pp. 4122-4147.

Visinescu, L. L., & Evangelopoulos, N. (2014). Orthogonal rotations in latent semantic analysis: An empirical study, *Decision Support Systems 62*, pp. 131-143.

Wegba, K., Lu, A., Li, Y., & Wang, W. (2017). Interactive Movie Recommendation Through Latent Semantic Analysis and Storytelling. arXiv preprint arXiv:1701.00199.

Weiss, C. H. (1983). Ideology, interests, and information. In D. Callahan, & B. Jennings (Eds.), *Ethics, the Social Sciences, and Policy Analysis* (pp. 213-245). New York: Plenum Press. http://dx.doi.org/10.1007/978-1-4684-7015-4_9

Wild, F. (2015). *Latent Semantic Analysis*. Disponível em https://cran.r-project.org/web/packages/lsa/lsa.pdf.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1), 97–107. http://doi.org/10.1109/TKDE.2013.109

Zelikowitz, S., & Hirsh, H. Using LSI for Text Classification in the Presence of Background Text, in Proceedings of the 10th International Conference on Information and Knowledge Management. 2001

Zhao, Y. (2013). *R and Data Mining: Examples and Case Studies*. Philadelphia: Eselvier.

Zipf, G. K. (1949). Human behavior and the principle of least effort. Oxford, England: Addison-Wesley Press.