UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

KEVIN PERONDI REGIS

MODELAGEM DE TÓPICOS EM REDES SOCIAIS DE POLÍTICOS BRASILEIROS

MONOGRAFIA

CAMPO MOURÃO
2019

KEVIN PERONDI REGIS

MODELAGEM DE TÓPICOS EM REDES SOCIAIS DE POLÍTICOS BRASILEIROS

Trabalho de Conclusão de Curso de Graduação apresentado à disciplina de Trabalho de Conclusão de Curso 2, do Curso de Bacharelado em Ciência da Computação do Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. André Luis Schwerz

CAMPO MOURÃO
2019

ATA DE DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO

Às **10:00** do dia **29 de novembro de 2019** foi realizada na sala **E101** da UTFPR-CM a sessão pública da defesa do Trabalho de Conclusão do Curso de Bacharelado em Ciência da Computação do(a) acadêmico(a) **Kevin Perondi Regis** com o título **Modelagem de Tópicos em Redes Sociais de Políticos Brasileiros**. Estavam presentes, além do(a) acadêmico(a), os membros da banca examinadora composta por: **Prof. Dr. André Luis Schwerz** (orientador(a)), **Prof. Dr. Rafael Liberato Roberto e Prof. Dr. Andre Luiz Satoshi Kawamoto**. Inicialmente, o(a) acadêmico(a) fez a apresentação do seu trabalho, sendo, em seguida, arguido(a) pela banca examinadora. Após as arguições, sem a presença do(a) acadêmico(a), a banca examinadora o(a) considerou _____ na disciplina de Trabalho de Conclusão de Curso 2 e atribuiu, em consenso, a nota _____ (_____). Esse resultado foi comunicado ao (à) acadêmico(a) e aos presentes na sessão pública. A banca examinadora também comunicou ao (à) acadêmico(a) que este resultado fica condicionado à entrega da versão final dentro dos padrões e da documentação exigida pela UTFPR ao professor responsável do TCC no prazo de **onze dias**. Em seguida foi encerrada a sessão e, para constar, foi lavrada a presente Ata que segue assinada pelos membros da banca examinadora, após lida e considerada conforme.

Observações:

_____

_____

_____

_____

Campo Mourão, **29 de novembro de 2019**

_____
Prof. Dr. Rafael Liberato Roberto
Membro 1

_____
Prof. Dr. Andre Luiz Satoshi Kawamoto
Membro 2

_____
Prof. Dr. André Luis Schwerz
Orientador

A ata de defesa assinada encontra-se na coordenação do curso.

Resumo

Regis, Kevin Perondi. Modelagem de Tópicos em Redes Sociais de Políticos Brasileiros. 2019. 52. f. Monografia (Curso de Bacharelado em Ciência da Computação), Universidade Tecnológica Federal do Paraná. Campo Mourão, 2019.

Com o surgimento das redes sociais, tornou-se possível reduzir a distância entre as pessoas e até mesmo estabelecer relações entre pessoas desconhecidas que partilham de um ideal em comum. Aproveitando-se dessa ideia, o cenário político brasileiro tem se adaptado a essa tecnologia com a atuação dos candidatos nas mídias sociais desde meados de 2010. As redes sociais se tornaram uma ferramenta fundamental para a divulgação de campanhas eleitorais e definições de estratégias para agradar o seu eleitorado. A popularidade do uso das redes sociais tem como consequência a geração de uma infinidade de dados complexos que podem ser analisados e explorados. Com o grande volume de dados, surge a necessidade de aplicar métodos e ferramentas a fim de processar e sintetizar as informações, facilitando a sua interpretação. Dentre os métodos utilizados, a modelagem de tópicos visa encontrar tópicos latentes contidos em um documento levando em consideração a ocorrência das palavras. Assim, o presente trabalho tem como objetivo verificar a eficiência do algoritmo para modelagem de tópicos conhecido como *Latent Dirichlet Allocation* (LDA). Para tanto, foi definido como público alvo os Senadores da República em exercício a partir do ano de 2015, visando coletar informações de suas publicações em páginas públicas na rede social Facebook. Após a coleta dos dados e aplicação do algoritmo LDA, verificou-se por meio de uma classificação manual os resultados obtidos, buscando avaliar o seu desempenho em termos de acertos e erros. Nesse processo, verificou-se um total de 7.694 publicações com classificação adequada ($\geq 0{,}7$) a partir das 112.584 publicações. Tal taxa representa 6,8% das publicações analisadas. A partir deste montante de publicações, foi realizada uma verificação manual com o intuito de apurar a quantidade de acertos e erros, o qual observou-se que o algoritmo obteve 5.239 acertos (4,7%) e 2.455 erros (2,2%). Assim, no presente trabalho a aplicação do algoritmo LDA não foi capaz de sintetizar precisamente os assuntos relevantes em um determinado conjunto de dados, levando em consideração a sua eficiência de padrão médio para a classificação de tópicos.

Palavras-chaves: Modelagem de Tópicos. LDA. Latent Dirichlet Allocation. Análise de Redes Sociais. Política.

Abstract

Regis, Kevin Perondi. Topic Modeling in Social Media of Brazilian Politicians. 2019. 52. f. Monograph (Undergradute Program in Computer Science), Federal University of Technology – Paraná. Campo Mourão, PR, Brazil, 2019.

With the emergence of social networks, it has become possible to reduce the distance between people and even establish relationships between unknown people who share a common ideal. Taking advantage of this idea, the Brazilian political scene has been adapting to this technology with the acting of candidates in social media since mid-2010. Social networks have become a fundamental tool for the dissemination of election campaigns and definitions of strategies to please the your electorate. The popularity of using social networks results in the generation of a multitude of complex data that can be analyzed and exploited. With the large volume of data, there is a need to apply methods and tools in order to process and synthesize information, facilitating its interpretation. Among the methods used, topic modeling aims to find latent topics contained in a document taking into account the occurrence of words. Thus, the present work aims to verify the efficiency of topic modeling algorithm known as LDA. Therefore, it was defined as target audience the Senators of the Republic in exercise from 2015, aiming to collect information from their publications on public pages on the social network Facebook. After data collection and application of the LDA algorithm, the results obtained were verified through a manual classification, trying to evaluate their performance in terms of hits and errors. In this process, there were a total of 7.694 publications with appropriate classification (≥ 0.7) from the 112.584 publications. This rate represents 6.8% of the publications analyzed. From this amount of publications, a manual verification was performed to determine the number of hits and errors, which showed that the algorithm obtained 5,239 hits (4.7%) and 2.455 errors (2.2%). Thus, in the present work the application of the gls LDA algorithm was unable to precisely synthesize the relevant subjects in a given data set, taking into account their average standard efficiency for the classification of topics.
Keywords: Topic Modeling. LDA. Latent Dirichlet Allocation. Social Network Analysis. Policy.

Lista de figuras

Lista de tabelas

Siglas

ABCP: Associação Brasileira de Ciência Política
API: *Application Programming Interface*
ARENA: Aliança Renovadora Nacional
BoW: *Bag-of-Words*
LDA: *Latent Dirichlet Allocation*
LSA: *Latent Semantic Analysis*
Mallet: *MAchine Learning for LanguagE Toolkit*
MDB: Movimento Democrático Brasileiro
NLTK: *Natural Language Toolkit*
PDS: Partido Democrático Social
PDT: Partido Democrático Trabalhista
PMDB: Partido do Movimento Democrático Brasileiro
PODE: Podemos
PP: Partido Progressista
PSC: Partido Social Cristão
PSDB: Partido da Social Democracia Brasileira
PT: Partido dos Trabalhadores
PTB: Partido Trabalhista Brasileiro
TSE: Tribunal Superior Eleitoral

Sumário

Capítulo 1

Introdução

O uso intensivo da Internet no dia a dia tem possibilitado a divulgação de informações em tempo real e facilitado o acesso a uma ampla variedade de conteúdo. Inicialmente, as redes sociais surgiram com a finalidade de reduzir a distância entre as pessoas e até mesmo estabelecer relações entre desconhecidos que partilham de um ideal em comum. Assim, não demorou muito para o meio político utilizar as redes sociais como meio de comunicação visando atingir a maior quantidade de pessoas, divulgar suas atividades políticas e gerar debates de assuntos relevantes no cotidiano.

De acordo com Vieira (2016), as redes sociais têm sido amplamente adotadas para comunicação política e divulgação de campanhas eleitorais no mundo todo nos últimos anos. No Brasil, este instrumento popularizou-se na corrida presidencial de 2010 e, desde então, tornou-se uma ferramenta fundamental como estratégia de comunicação pelos candidatos.

Por outro lado, as redes sociais se tornaram uma fonte de dados, em que é possível realizar diversas análises qualitativas e quantitativas, de grande interesse corporativo e para o desenvolvimento de pesquisas. A grande capacidade de disseminação de informações pelas redes sociais impulsionou a uma nova abordagem para as campanhas eleitorais, quando comparadas com eleições anteriores.

A aplicação da análise de dados em redes sociais, permite a identificação dos principais assuntos abordados pelo público alvo, bem como, demais atividades que descrevem o comportamento do usuário na rede social. Atualmente, existem diversos trabalhos abordando a análise de dados em diferentes redes sociais como: Facebook (LEV-ON; HALEVA-AMIR, 2016);(Alashri et al., 2016); (MURTA et al., 2017); (FREELON, 2017); (STIER et al., 2018), Twitter ((JüRGENS; JUNGHERR, 2015); (JUNGHERR, 2015); (KARLSEN; ENJOLRAS, 2016); (SOUZA et al., 2017); (KREISS; MCGREGOR, 2018), Instagram ((FILIMONOV et al., 2016); (BOSSETTA, 2018), entre outros.

Atualmente, o Facebook destaca-se como a maior rede social da atualidade com

aproximadamente 2,375[1] bilhões usuários. Nela, o usuário pode criar páginas públicas para a divulgação de informações de fácil acesso. Além disso, o Facebook disponibiliza uma *Application Programming Interface* (API) que dá acesso aos conteúdos, possibilitando sua extração. Diante disso, torna-se viável a utilização desta rede social para a manipulação de dados, por meio de aplicações, como *crawlers*.

Todavia, o uso massivo das redes sociais tem como consequência a geração diária de grande quantidade de informação. O excesso de dados, muitas vezes dificulta uma avaliação efetiva, uma vez que há conteúdos inconsistentes ou insignificantes. Assim, torna-se fundamental a aplicação de medidas e modelos a fim de sintetizar as informações, evidenciando apenas os conteúdos relevantes.

Dentre os métodos de análise de redes sociais, podemos citar a Modelagem de Tópicos utilizando o algoritmo LDA, proposto por Blei et al. (2003). Basicamente, este modelo probabilístico visa encontrar tópicos latentes contidos em um documento levando em consideração a ocorrência das palavras. Na Modelagem de Tópicos são utilizadas medidas de coerência para a definição de um tópico, medindo o grau de similaridade semântica entre as palavras que obtiveram maior pontuação.

No entanto, como qualquer outro método de aprendizado de máquina, a utilização desse modelo está sujeita ao erro de classificação ou definição de tópicos. Neste contexto, o presente trabalho tem como objetivo verificar a eficiência do algoritmo LDA em identificar os principais assuntos abordados por um determinado público alvo. Dessa forma, o presente trabalho teve como hipótese verificar a precisão do algoritmo LDA em sintetizar assuntos relevantes abordados em uma determinada classe política com diferentes ideologias partidárias.

Para tanto, foi definido como público alvo os Senadores da República em exercício a partir do ano de 2015, visando coletar informações de publicações em páginas públicas dos respectivos Senadores na rede social Facebook.

Os dados coletados foram submetidos à modelagem de tópicos com a aplicação do algoritmo LDA, seguido de uma classificação manual dos resultados obtidos que buscou avaliar o seu desempenho em termos de acertos e erros.

A partir dos resultado obtidos, verificou-se que aproximadamente metade do conteúdo extraído das publicações classificadas pelo algoritmo, dentro dos parâmetros estabelecidos, se confirmaram pertencentes ao tópico identificado.

Nos próximos capítulos serão abordados a fundamentação teórica contendo o embasamento dos principais assuntos pertinentes ao presente trabalho, bem como, a descrição de trabalhos relacionados, a metodologia aplicada para o desenvolvimento do objetivo proposto, os resultados obtidos com a respectiva discussão dos dados e, por fim, a conclusão apresentando o desfecho das análises realizadas.

[1] <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>

Capítulo 2

Fundamentos

Neste capítulo serão abordados assuntos relevantes deste trabalho, tendo como base informações encontradas na literatura . Inicialmente será apresentado uma breve explicação referente ao cenário da política brasileira, associando-os ao uso das redes sociais. Em seguida será discutido o conceito de *Big Data*, o uso de coletores de informações (Web Crawler) e por fim, a contextualização da Modelagem de Tópicos.

2.1. Contextualização da Política Brasileira

Em 1979, com a reforma da Lei nº 6.767, deu-se a abertura para a criação de novos partidos políticos no Brasil. Assim, surgiu um sistema partidário composto por: Partido do Movimento Democrático Brasileiro (PMDB), Partido Democrático Social (PDS), Partido dos Trabalhadores (PT), Partido Democrático Trabalhista (PDT), Partido Trabalhista Brasileiro (PTB) e Partido Progressista (PP). Tais partidos tiveram como referência política a Aliança Renovadora Nacional (ARENA) e o Movimento Democrático Brasileiro (MDB). A liberdade total para as organizações partidárias veio com a Ementa Constitucional de 1985 (BRAGA; BOURDOUKAN, 2009).

Segundo Madeira e Tarouco (2011), inicialmente as frentes ideológicas eram divididas de acordo com a relação com o antigo regime, ou seja, aqueles que apoiavam o governo autoritário eram considerados de direita e os opositores, denominados de esquerda, eram representados por uma parcela significativa de grupos que foram reprimidos durante o regime militar. Em seguida, alguns partidos políticos de direita decidiram desvincular a sua imagem do antigo regime formando a terceira frente, denominada centro (POWER; Zucco Jr., 2009).

Tarouco e Madeira (2015) realizaram uma comparação de trabalhos científicos que classificavam partidos de acordo com escalas, sendo os menores valores indicativos de um partido de extrema esquerda e os maiores valores indicativos de um partido de extrema

direita, como pode ser observado na Tabela 2.1. Os dados da primeira coluna foram extraídos durante um encontro realizado pela Associação Brasileira de Ciência Política (ABCP) no ano de 2010, em que os participantes foram convidados a classificar os partidos políticos em uma escala de 1 a 7. Já a segunda coluna estão representados os valores reportados por Wiesehomeier e Benoit (2007) no qual os autores aplicaram questionários por uma página[1] em que solicitavam à analistas brasileiros que classificassem os partidos políticos em uma escala de 1 a 20.

Tabela 2.1. Posição dos partidos nas duas escalas Fonte: (TAROUCO; MADEIRA, 2015)

Partido	Posição na escala do survey ABCP 2010	Posição na escala Wiesehomeier e Benoit 2007
Psol	1,4	2,95
PCdoB	2,3	4,96
PT	2,9	6,37
PSB	3,0	7,50
PDT	3,3	8,38
PV	3,5	7,36
PPS	4,0	10,38
PMDB	4,2	11,50
PSDB	4,6	13,46
PTB	5,0	13,60
PSC	5,2	15,62
PP	6,0	16,78
DEM	6,2	17,33

Codato et al. (2018) realizaram uma distribuição dos partidos políticos de acordo com seus grupos ideológicos, avaliando os candidatos à Câmara Federal nas eleições de 1998, 2002, 2006, 2010 e 2014, de acordo com a Tabela 2.2.

Tabela 2.2. Distribuição dos partidos políticos brasileiros por grandes grupos ideológicos. Adaptado de (CODATO et al., 2018)

Esquerda	Centro	Direita
PC do B, PCB, PCO, PDT, PPL, PSB, PSOL, PSTU, PT, Rede	(P)MDB, PMN, PPS, PROS, PSDB, PV, PHS	PAN, PEN, DEM, PGT, PL, PR, PPB/PP, PRB, PRN, PRONA, PRP, PRTB, PSC, PSD, PSDC, PSL, PSN, PST, PT do B, PTB, PTC, PODE, SD

[1] <http://www.wiesehomeier.net/>

De acordo com a página do Tribunal Superior Eleitoral (TSE)[2], em 2019 há um total registrado de 32 partidos políticos no Brasil com diferentes ideologias partidárias. Em tempos de maciço uso da internet e mídias sociais, muitos políticos adotaram essas ferramentas como meio de comunicação para divulgação de campanhas eleitorais e interação com os cidadãos.

Segundo Baquero et al. (2016), o surgimento de modernas tecnologias envolvendo a internet tornou mais fácil o acesso à informação e comunicação por meio das redes sociais. Assim, tais ferramentas tornaram-se fundamentais para o desenvolvimento da cultura política atingindo principalmente o público juvenil que se comunica e vive suas relações sociais de modo diferenciado quando comparado com gerações anteriores.

Braga e Carlomagno (2018) mostram que a partir de 2010 surgem os primeiros conteúdos de cunho político veiculados pelas mídias sociais de candidatos como instrumento de mobilização e agendamento de campanhas, desconstrução da imagem dos adversários, debates, surgimento das chamadas *fake news*, entre outros.

Dessa forma, a socialização da política em redes sociais tem contribuído para a construção de um novo senso crítico de cidadania e uma cultura política entre os eleitores de modo mais participativo.

2.2. Redes Sociais

A rede social surgiu da necessidade humana de se comunicar e interagir. Segundo Marteleto (2001), a rede (*network*) pode ser definida como uma estrutura sem fronteiras, uma comunidade não geográfica, um sistema de apoio ou até mesmo um espaço físico que se pareça com uma árvore. Na realidade, o trabalho pessoal como redes de conexões existe há muito tempo, mas somente nas últimas décadas que analisaram como uma ferramenta organizacional.

De acordo com Levato (2013), as redes sociais são consideradas espontâneas, com comunicação dinâmica, simples e rápida entre os seus usuários no qual é possível um melhor aproveitamento dos recursos oferecidos. Ainda, pode ser definida como um grupo de indivíduos, tanto pessoas quanto empresas, que estão ligados uns aos outros por algum interesse em comum.

Dentre as redes sociais existentes, o Facebook e o Twitter têm se destacado por sua popularidade entre os usuários. Criado em fevereiro de 2004, o Facebook tornou-se uma das maiores redes sociais, capaz de criar milhões de interações sociais diariamente.

Segundo Wilson et al. (2012), o Facebook possuía mais de 845 milhões de usuários ativos em meados de 2012, mais de 4 bilhões de compartilhamentos de conteúdos por dia, incluindo o carregamento de aproximadamente 250 milhões de fotos e integrado com mais de sete milhões de páginas e aplicativos. De acordo com o *site* Statista[3], em agosto de 2017 o

[2] <http://www.tse.jus.br>
[3] <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>

Facebook ultrapassou 2 bilhões de contas ativas e o Twitter possui aproximadamente 328 milhões.

De modo geral, os usuários do Facebook apresentam comportamento de busca por informações sociais de pessoas, empresas e grupos, bem como, criam laços estreitos com pessoas desconhecidas sem nenhuma conexão fora das redes sociais (ELLISON et al., 2011).

Valenzuela et al. (2009) verificaram que a intensidade do uso do Facebook pode estar relacionada com a participação política e civil da população, a satisfação da vida e a confiança social dos usuários. Por meio de um perfil privado, os usuários podem tornar público "*posts*" ou mensagens que expressam seus sentimentos, opiniões, atividades comuns entre amigos ou detalhes sobre *sites* ou eventos externos em seu "Mural". Nesses "*posts*" os contatos ou "amigos" podem deixar comentários, curtidas e até mesmo realizar compartilhamento do mesmo, tal como, acompanhar todas as publicações feitas por páginas ou demais "amigos" através do "*Feed* de notícias".

Por outro lado, o Facebook permite que os usuários criem e se juntem a grupos ou páginas com base em interesses e atividades comuns incorporando-os aos seus perfis. Assim, uma parcela importante do impacto cívico e político do Facebook ocorre dentro de grupos desenvolvidos por usuários e organizações. Ao mesmo tempo, o aumento da participação em grupos *on-line* e *off-line* geralmente ajuda a criar relações de confiança entre os membros (KOBAYASHI et al., 2006).

Tendo em vista o poder de divulgação que uma página no Facebook pode conter, figuras políticas adotaram esse meio como forma de divulgar seu trabalho ou propostas. E assim, estar em constante participação com seu círculo social virtual.

Braga e Carlomagno (2018) afirmam que com a crescente disponibilidade de grandes quantidades de informação no mundo virtual (os chamados *big data*), o emprego das tecnologias digitais daria lugar a uma nova fase das campanhas eleitorais, distinta das anteriores.

Segundo McAfee e Brynjolfsson (2012), as redes sociais são responsáveis por uma grande massa de dados, também conhecidos como "*Big Data*". O mesmo é válido para *smartphones* e os outros dispositivos móveis que são capazes de fornecer enormes fluxos de dados vinculados a pessoas, atividades e locais. O *Big Data* em redes sociais são compostos por dados de mensagens, atualizações, imagens postadas, comentários, sinais de GPS, entre outras, em que os usuários podem ser considerados como geradores de dados.

Tais dados são diariamente armazenados e processados para fins de interesse corporativo ou até mesmo utilizado para desenvolvimento de pesquisas e análise de dados, a partir de informações públicas disponíveis nas redes sociais.

2.3. Big Data

De acordo com Morales et al. (2016), o termo *Big Data* passou a ser difundido no contexto tecnológico por cientistas e executivos em meados de 2008. Atualmente o *Big Data* é considerado não só um grande volume e variedades de informações armazenados, mas também uma fonte de recursos e valores ocultos.

Com o início da era *Big Data* as empresas foram afetadas de maneira significativa com o crescimento surpreendente dos dados, uma vez que os bancos de dados tradicionais e o seu gerenciamento não estavam suportando o grande volume de informações. Assim, surgiu uma nova geração de tecnologias mais complexas e capazes de armazenar um conjunto de dados amplamente maiores, tendo como exemplo, "NoSQL" (MARZ; WARREN, 2014).

O termo "NoSQL" pode ser interpretado como a abreviação de "*NotOnly*SQL", no qual foi introduzido por Carlo Strozzi em 1980. O principal motivo para o desenvolvimento do NoSQL foi a Web 2.0, que aumentou a quantidade de uso e dados armazenada nos bancos de dados. Este tipo de banco de dados é projetado para lidar com todos os tipos de falhas, em que as variedades de falhas de hardware são funcionalmente consideradas eventuais ocorrências do que eventos excepcionais (ABRAMOVA; BERNARDINO, 2013).

Han et al. (2011) afirmam que as principais vantagens do NoSQL consistem em leitura e gravação de dados rapidamente, suporte de armazenamento em massa, fácil expansão e baixo custo.

De acordo com Strauch (2011), algumas redes sociais como Facebook e Twitter utilizam o Cassandra como forma de armazenamento de dados. O Cassandra é considerado um banco de dados NoSQL implementado em Java, o qual foi desenvolvido pela Fundação Apache Software juntamente com o Facebook e tornou-se código aberto em meados de 2008.

Segundo Lobo (2017), nos últimos anos os dados passaram a ser armazenados na internet em sistemas de *warehouses* ou em redes de computadores na forma de "nuvem". Atualmente, estima-se que o volume de dados seja de 2,5 exabytes, com um crescimento anual previsto de 57% no período de 2014 a 2019, resultando em um volume de aproximadamente 24,3 exabytes em 2019.

Considerando que as redes sociais são grandes fontes de informações, a aquisição dos dados são realizadas por rastreadores Web que são amplamente utilizados em aplicativos baseados em páginas Web. Tais rastreadores atuam como mecanismos de pesquisa apresentando inúmeras soluções eficientes em diversas áreas de conhecimento (CHEN et al., 2014). Dentre os mecanismos de mineração de dados em páginas web, é possível citar o Web Crawler.

2.4. Web Crawler

De acordo com Baeza-Yates e Ribeiro-Neto (2008), um Web Crawler é um software usado para download automático de páginas Web. Sua execução solicita como entrada uma ou mais páginas de inicialização ou páginas semente que são baixadas, analisadas e digitalizadas para novos links. Os links apontando para páginas que ainda não foram baixadas são adicionados a uma fila central de URLs para recuperação posterior. Em seguida, o rastreador seleciona uma nova página da fila para baixar e o processo é repetido até que um critério de parada seja cumprido.

Na Figura 2.1 é apresentado um fluxograma descrevendo a funcionalidade de um Web Wrawler básico.

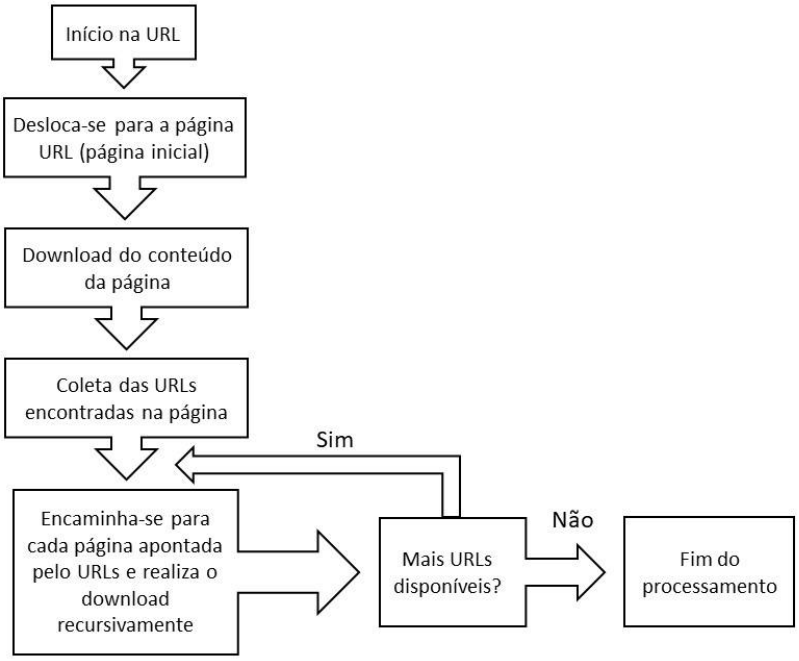

Figura 2.1. Fluxograma sobre as etapas para a implementação de um Web Crawler. Adaptado de (BAEZA-YATES; RIBEIRO-NETO, 2008)

Segundo Sreeja e Chaudhari (2014), os Web Wrawlers podem ser divididos em 3 categorias:

- *Deep* web crawler: este tipo de web crawler rastreia páginas da web sem considerar a relação entre as páginas. Eles começam a partir de um "URL de sementes" e rastreia páginas da web recursivamente até que não haja mais links a seguir.
- *Focused* crawler: é um tipo de web crawler que rastreia páginas da web que são específicas de um tópico ou domínio pré-definido. Quando comparado ao *Deep* web crawler, os *focused* crawler consomem menos quantidade de recursos do sistema e

utilizam uma variedade de técnicas para recuperação específica do domínio.

- *Forum* crawler: é um software específico utilizado para o rastreamento de *sites* de fórum em que os usuários compartilham conhecimentos, postam suas consultas ou encontram diversas informações relacionadas às suas consultas. Este tipo de crawler pode ser genérico ou personalizado e utilizam diferentes métodos para a execução do rastreamento.

De modo geral, os tipos de coletas de dados realizadas pelos pesquisadores são baseadas na utilização de Crawlers que usam uma API fornecida pela aplicação. Dependendo da API utilizado pelo Crawler podem apresentar algumas restrições na coleta de dados, no qual o desenvolvedor deve estar ciente quanto a interpretação e o manuseio da informação obtida. (RECUERO, 2014)

A aplicação do Web Crawler é capaz de gerar uma gigantesca base de dados fundamentais para a análise de redes sociais, possibilitando a exploração e interpretação das informações extraídas. Dessa forma, a análise de redes sociais tem sido aplicada no desenvolvimento de pesquisas em diversas áreas de conhecimento e tipos de intercâmbios de informações (QUAN-HAASE; MCCAY-PEET, 2016).

A partir da extração de dados realizada pelo Web Crawler, o conjunto de informações permitirá a análise de rede social, mediante a aplicação da modelagem de tópicos para compactar a quantidade de dados.

2.5. Modelagem de Tópicos

De acordo com Qin et al. (2016), a modelagem de tópicos é um dos modelos bayesianos hierárquicos mais conhecidos para o tratamento de linguagem natural e análise de documentos. É capaz de executar com êxito a classificação de textos, no qual o conteúdo é representado por palavras, desconsiderando a gramática e até a ordem das palavras por meio do modelo de *Bag-of-Words* (BoW).

Para Griffiths et al. (2007), afirmam cada tópico é uma distribuição de probabilidade por palavras, e o conteúdo do tópico é refletido nas palavras às quais atribui alta probabilidade. Por exemplo, a alta probabilidade de ocorrência das palavras “Madeira” e “Riacho” sugerem que o tópico se refere a natureza, enquanto as altas probabilidades das palavras “Federal” e “Reserva” sugerem que um tópico se refere a Finanças.

Segundo Souza et al. (2017), as principais dificuldades encontradas na modelagem de tópicos em redes sociais são o processamento da grande quantidade de informações geradas, bem como, os ruídos existentes, uma vez que, nem todo conteúdo compartilhado ou gerado é considerado relevante.

Nesse contexto, o algoritmo LDA realiza o processamento de grande quantidade de dados com o intuito de reduzi-los por meio do modelo BoW. Basicamente, o LDA é

um modelo probabilístico generativo de um corpus. A ideia básica é que os documentos sejam representados como misturas aleatórias sobre tópicos latentes, em que cada tópico é caracterizado por uma distribuição sobre palavras. Além disso, é escolhido um vocabulário básico de "palavras" ou "termos" e, para cada documento do corpus, é formada uma contagem do número de ocorrências de cada palavra (BLEI et al., 2003).

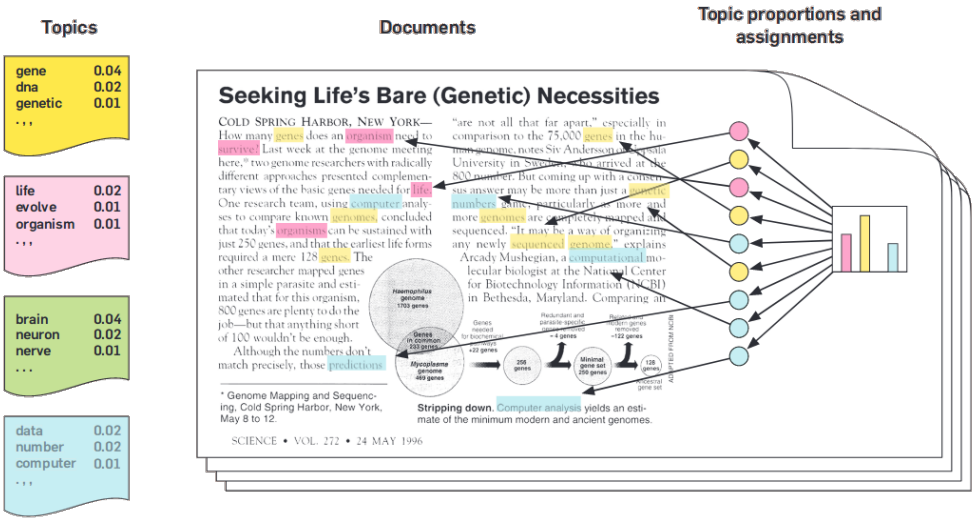

Figura 2.2. A intuição por trás do LDA. Adaptado de (BLEI, 2012)

Na Figura 2.2, Blei (2012) exemplifica a intuição por trás do modelo LDA, onde assume-se que existe um certo número de tópicos, que são distribuições por palavras, para toda a coleção. Além disso, presume-se que inicialmente cada documento é gerado a partir da escolha de uma distribuição sobre os tópicos. Na sequência, para cada palavra, é escolhido uma atribuição de tópico e definido a palavra para o tópico correspondente.

2.6. Considerações Finais

Em síntese, a fundamentação teórica baseou-se na contextualização dos partidos políticos brasileiros e sua respectiva classificação em esquerda, centro e direita. Logo após, foi apresentado uma abordagem referente ao uso de redes sociais por políticos para divulgação de trabalhos e campanhas com o intuito de alcançar o maior número de eleitores. Devido ao uso massivo da internet, as redes sociais se tornaram uma grande fonte dados conhecido como *Big Data* permitindo a extração de informações por meio de Crawlers. Assim, todo conteúdo extraído é passível de análises utilizando a Modelagem de Tópicos, permitindo a condensação das informações e a interpretação do conteúdo com base na hipótese levantada no presente trabalho.

Diante do exposto, algumas metodologias foram aplicadas, como pode ser observado como próximo capítulo.

Capítulo 3

Trabalhos Relacionados

Na literatura existem diversos trabalhos abordando a análise de redes sociais com a aplicação da modelagem de tópicos, a fim de compreender como os candidatos políticos divulgam informações e consequentemente como os usuários reagem a eles. Além disso, a modelagem de tópicos permite a identificação dos principais assuntos abordados nas atividades dos candidatos, bem como, quais fatores afetam, direta ou indiretamente, nos acontecimentos do mundo real, uma vez que, parte-se do princípio que as redes sociais é um reflexo da realidade em que estão inseridos. Neste contexto, serão descritos alguns estudos realizados por outros autores em diferentes partes do mundo que estão reportados na literatura e relacionados com o tema do presente trabalho.

Hoang et al. (2013) realizaram a análise de redes sociais a partir de um conjunto de 56 usuários amplamente seguidos do Twitter no cenário político dos EUA, previamente identificados como usuários iniciais. Em seguida, o conjunto de usuários foi expandido considerando todos os seguidores, o qual resultou em um conjunto de mais de 408 mil usuários. Para tanto, foram aplicadas a detecção de polaridade de sentimento (positivo, negativo ou neutro), a modelagem de tópicos acompanhado com o uso do algoritmo LDA para identificar as palavras-chaves inseridas no conteúdo extraído e a classificação do usuário pelo vínculo político.

De acordo com os resultados obtidos pelos autores, verificou-se que os tweets negativos e positivos são estatisticamente mais sujeitos a retweets do que os neutros, assim como, os tweets negativos também são estatisticamente mais sujeitos a retweets quando comparados com os tweets positivos. Também foram identificados 80 tópicos e realizado a distribuição de retweets associado à afiliação política, dentre os quais observou-se que, numericamente, a maioria dos retweets são de tweets de usuários de mesma afiliação política. Os autores concluíram que o sentimento e a afiliação política têm efeitos na "retweetabilidade" de tweets políticos, bem como, verificaram que usuários com diferentes afiliações políticas demonstram

emoções distintas em cada tópico.

Song et al. (2014) utilizaram a análise de redes sociais com o intuito de compreender o modo em que as questões sociais e políticas, que são discutidas no Twitter, estavam relacionadas às eleições presidenciais da Coréia em 2012. Para tanto, foi aplicado o algoritmo LDA para analisar e validar a relação entre os tópicos extraídos de tweets e eventos relacionados. Além disso, desenvolveram um termo técnico de recuperação de co-ocorrência para rastrear termos cronologicamente co-ocorrentes e, assim, compensar as limitações da LDA. Por fim, identificaram a coerência temática entre usuários identificados no envio-recebimento de menções.

Os autores verificaram que o Twitter é um meio útil para rastrear tendências atuais em tempo hábil. Em particular, descobriram que questões controversas no Twitter são geradas, propagadas e extintas de uma maneira semelhante à mídia existente, porém mais inovadora. Também observaram que a análise de rede baseada em menção indica que usuários do Twitter com disposições políticas semelhantes, tendem a se comunicar frequentemente com seus companheiros sociais através do envio ou recebimento de menções, mas a conexão entre os usuários pode mascarar o comportamento da opinião no mundo real. Dessa forma, a análise de tendências temporais por meio da modelagem de tópicos, juntamente com a análise de co-palavras por co-ocorrência de termos, mostra os diferentes aspectos das questões sociais discutidas nas redes sociais, permitindo uma interpretação das implicações das questões sociais de maneira inovadora.

Alashri et al. (2016) aplicaram a análise de redes sociais referente às eleições presidenciais dos Estados Unidos em 2016 visando compreender a conexão entre as atividades online e offline. Assim, foram coletadas 9.700 postagens no Facebook de cinco candidatos à presidência (Hillary Clinton, Donald Trump, Bernie Sanders, Ted Cruz e John Kasich) por meio de suas páginas oficiais do Facebook e 12.050.595 comentários sobre postagens. Além disso, foram aplicadas a modelagem de tópicos, análise de sentimentos e detecção de tendências usando transformadas de *wavelet* para descobrir tópicos, tendências e reações.

Os autores verificaram que os candidatos republicanos apresentaram maior probabilidade em compartilhar informações sobre eventos controversos que ocorreram durante o ciclo eleitoral, enquanto os candidatos democratas se concentram em questões de política social. Também, observaram que os comentaristas das páginas de candidatos republicanos expressam sentimentos negativos em relação às políticas públicas recentes, uma vez que raramente apoiam as decisões tomadas pelo governo Obama, enquanto os comentaristas nas páginas de candidatos democráticos apresentaram maior probabilidade de expressar apoio à continuação ou avanço das políticas. Por meio da correlação entre as tendências on-line de comentários com os sentimentos e eventos off-line verificaram que as mudanças nas tendências on-line são motivadas por três fatores: publicação popular, debates off-line e candidatos que abandonaram a corrida presidencial.

Seva et al. (2016) analisaram as atividades da rede social Twitter buscando compreender o comportamento de usuários durante as eleições governamentais na Croácia ocorridas em 2015. Ao todo foram coletados 28.049 tweets gerados por 2.838 usuários e submetidos à modelagem de tópicos utilizando o algoritmo LDA. Assim, o modelo aplicado foi capaz de gerar 30 tópicos, mas apenas 7 tópicos abrangeram a maioria dos tweets, os quais abordaram assuntos relacionados à "obrigações presidenciais", "futuro", "partidos políticos", "líderes políticos", "comentários negativos", "chances de permanecer no cargo" e "comentários na TV".

Os autores observaram que naturalmente os maiores volumes de tweets foram registrados no dia das eleições, seguido de considerável diminuição de postagens nos dias seguintes ao evento. Dessa forma, concluiu-se que a análise de redes sociais são úteis para construir uma imagem da opinião pública sobre as opções políticas incluídas, prever o resultado do evento ou identificar os principais conceitos usados pelo público, medir a popularidade dos candidatos, bem como, mostrou-se mais efetivas que as pesquisas tradicionalmente manuais.

Ainda, é possível verificar que a análise de redes sociais pode ser aplicada em diferentes temas ou público alvo, como pode ser observado no trabalho a seguir.

Alexandrov et al. (2016) realizaram o mapeamento de temas e tópicos étnicos associados ao Cáucaso (área geográfica que divide a Europa Oriental e a Ásia Ocidental) na rede social VKontakte, popular na Eurásia e similar ao Facebook. Neste trabalho foram coletados dados sobre comunidades virtuais associadas aos principais grupos étnicos (armênios, georgianos e azerbaijanos) e supra-étnicos (pan-caucasianos), bem como, a aplicação da modelagem de tópicos utilizando o algoritmo LDA para identificar as ideologias e características culturais que unem e dividem o Cáucaso virtual.

De acordo com os autores, apesar da região do Cáucaso ser considerada uma área de conflito por possuir diversas etnias, idiomas, religiões e posições políticas, observou-se que a maioria dos grupos da rede social VKontakte não possui ideologia política ou religiosa específica. Muitos usuários utilizam as comunidades virtuais para manter sua lealdade simbólica a seus grupos nacionais e étnicos, bem como, para construir novas comunidades ou movimentos virtuais pan-caucasianos, mas sem qualquer ideologia religiosa ou política específica.

A modelagem de tópicos mostrou-se fundamental para a realização dos trabalhos abordados acima. Entretanto, os autores não buscaram formas de validação dos resultados obtidos pelo algoritmo. Além disso, no melhor do nosso conhecimento, não há trabalhos que apliquem a modelagem de tópicos em redes sociais com a língua portuguesa.

Capítulo 4

Metodologia

Ainda que haja uma diversidade de partidos políticos brasileiros, acredita-se que os indivíduos nas redes sociais abordam assuntos em comum independente do posicionamento ou partido político. De modo geral, a comunidade política utiliza as redes sociais para promover sua campanha política, divulgar as atividades realizadas, gerar debates e expor suas opiniões relacionados aos acontecimentos do cotidiano. Tais assuntos podem conter informações ocultas, que podem auxiliar na compreensão a respeito do comportamento de seus usuários. Assim, a análise de redes sociais é capaz de gerar tópicos relevantes que podem estar relacionados direta ou indiretamente com eventos ocorridos durante o período de coleta de dados.

Nessas circunstâncias, pretende-se nesse trabalho verificar a **hipótese** de que:

A aplicação do algoritmo Latent Dirichlet Allocation (LDA) pode sintetizar precisamente os assuntos relevantes abordados em uma determinada classe política com diferentes ideologias partidárias.

Para verificar essa hipótese, aplicou-se um método de pesquisa com as seguintes etapas: (i) seleção dos políticos para avaliação no contexto nacional; (ii) coleta de dados das páginas do Facebook dos políticos selecionados; (iii) realização de um pré-processamento das informações; (iv) aplicação da modelagem de tópicos; e (v) análise dos dados.

A Figura 4.1 ilustra o fluxograma das atividades realizadas para atingir o objetivo desta proposta de pesquisa.

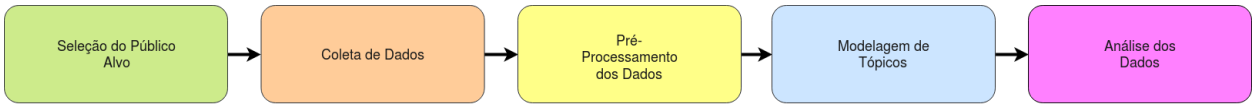

Figura 4.1. Fluxograma das atividades a serem realizadas

Nas seções seguintes serão apresentados os detalhes de cada etapa citada anteriormente.

4.1. Seleção do Público-alvo

Diante do elevado conjunto de dados existentes na política brasileira, o presente trabalho tem como público-alvo no contexto nacional os senadores da república devido à quantidade reduzida de indivíduos, viabilizando uma análise integral dos dados.

A partir de informações contidas na página[1] do Senado Federal foi levantado um total de 81 indivíduos, podendo ser verificados na Apêndice A. Dentre eles, três não possuíam uma página na rede social e um deles não estava vinculado a nenhum partido político. Sendo assim, para esta pesquisa, foram contemplados 78 indivíduos que representam 26 estados mais o Distrito Federal.

4.2. Coleta de Dados

A coleta de dados foi baseada em páginas públicas da rede social Facebook com a extração de informações datadas até fevereiro de 2018, sendo este o período referente ao término da coleta de dados em virtude da mudança da Política de Privacidade do Facebook, que dificultou a extração ilimitada de dados. Dentre as informações coletadas, o presente trabalho teve como intuito analisar os principais tópicos abordados pelo público alvo durante o período correspondente ao início de 2015 até o final de 2017, totalizando 3 anos.

O Facebook disponibiliza aos desenvolvedores a Graph API[2] como a principal forma de inserir, excluir e consultar dados via página Web. Assim, os usuários do Facebook podem explorar e compreender o funcionamento dos métodos utilizados pela rede social em detalhes por meio de um *token* de acesso.

Desta forma, as aplicações externas heterogêneas que necessitam de comunicação com a rede social requerem métodos para realizá-la. Consequentemente, bibliotecas externas são elaboradas proporcionando a interação entre a aplicação e a rede social.

Logo, o algoritmo coletor utilizado no presente trabalho armazenou as informações obtidas em um banco de dados relacional em que cada atividade contida na página foi organizada em tabelas.

Para a criação e armazenamento do banco de dados utilizou-se o PostgreSQL[3] pelo fato de ser um poderoso sistema de gerenciamento para banco de dados de objetos relacionais de código aberto.

O banco de dados foi modelado para armazenar conteúdos em duas tabelas principais,

[1] <http://www25.senado.leg.br/web/senadores/em-exercicio>
[2] <https://developers.facebook.com/docs/graph-api/>
[3] <https://www.postgresql.org>

sendo a primeira contendo informações das páginas (identificador da página no Facebook, nome do senador, estado e partido político) e a segunda com informações das publicações (identificador da mensagem no Facebook, conteúdo textual da publicação, data e hora de criação e o identificador da página no qual está relacionada).

A partir das ferramentas acima citadas, foi desenvolvido um mecanismo de coleta que desempenhou a função de um Web Crawler. Para tanto, utilizou-se a biblioteca RestFB[4], que possui a capacidade de incorporar as atividades existentes na rede social Facebook por meio de métodos implementados em Java, bem como, permitiu a extração de informações contidas na rede social.

Em seguida, foi realizado uma varredura manual com a finalidade de identificar o nome do indivíduo, o identificador de sua página, o estado ao qual o indivíduo pertencia e o partido político associado para posteriormente adicioná-los no banco de dados.

Para realizar a conexão com o Facebook foi necessário um *token* de acesso. Porém, o *token* de um usuário comum possui um tempo de conexão limitado. Com isso, foi necessário recorrer a um *token* de aplicação, que garantiu o acesso às informações com uma conexão de tempo ilimitado à API do Facebook até a data de término da coleta. Com a alteração da Política de Privacidade a quantidade de requisições a API foram restringidas para apenas 200 chamadas por hora para cada token, o que inviabilizou a continuidade da coleta de dados.

Dessa forma, o coletor foi desenvolvido para interpretar cada página, publicação e comentário como objetos que, por sua vez, possuíam atributos e métodos que realizavam os devidos tratamentos das informações coletadas e armazenavam na respectiva tabela do banco de dados. As publicações referem-se as postagens realizadas pelos políticos em suas páginas, e os comentários são as mensagens dos usuários sobre as publicações dos políticos.

Por padrão, a API do Facebook retornava 25 objetos a cada consulta juntamente com um cursor, sendo respeitado um esquema de paginação para se obter os objetos seguintes até a última informação.

Logo, com a finalidade de otimizar o tempo de coleta de dados, o coletor foi modificado para realizar a captura das informações de maneira paralela, por meio de *Threads*.

Para orientar as *Threads*, foram criados campos de controle no banco de dados com os status “novo”, “coletando” e “coletado”. Deste modo, cada *Thread* instanciada pelo coletor selecionava uma página com o status “novo”, subsequentemente alterava o status para “coletando”, realizava a coleta das publicações e no término de seu processamento atualizava o status da página para “coletado”.

A partir da coleta de dados, as informações obtidas foram submetidas à um pré processamento das palavras presentes nas publicações.

[4] <http://restfb.com/>

4.3. Pré Processamento dos Dados

O intuito principal do pré processamento dos dados para este trabalho foi reduzir ao máximo as informações irrelevantes contida nas publicações tais como *stopwords*, links, *hashtags*, entre outros.

Para isso, foi elaborado uma rotina utilizando a linguagem de programação Python[5] em conjunto com a biblioteca de processamento de linguagem natural *Natural Language Toolkit* (NLTK)[6], que possui métodos de tratamento para a língua portuguesa.

O passo inicial do pré processamento foi submeter cada mensagem contida nas publicações para a retirada de quebras de linha, links, acentuações, *hashtags* e citações. Na sequência, todas as palavras restantes foram separadas individualmente para se tornarem *tokens*, que foram utilizados nas próximas etapas.

O processamento seguinte foi a remoção das *stopwords*, que são palavras irrelevantes para o contexto deste trabalho. O NLTK tem disponível uma lista de *stopwords* direcionadas para a língua portuguesa. Dessa forma, criou-se uma rotina para remover essas palavras quando presentes nas publicações.

Por fim, foi realizado o processo de stemização ou *stemming*, que consiste em extrair o radical da palavra, ou seja, o elemento básico e significativo. Com ajuda do NLTK, que disponibiliza ao desenvolvedor um *Stemmer* para língua portuguesa, foi extraído o radical de cada palavra e armazenados em uma lista temporária para ser utilizada na modelagem de tópicos.

4.4. Modelagem de Tópicos

Basicamente, a modelagem de tópicos foi realizada com a aplicação do algoritmo LDA proposto por Blei et al. (2003), o qual foram compostas por duas etapas distintas, sendo estas: treinamento do algoritmo não supervisionado e classificação das publicações.

4.4.1. Treinamento

A primeira etapa para a modelagem de tópicos resultou na criação de condições adequadas para o treinamento do algoritmo. Nesta etapa foi utilizado a biblioteca Gensim[7] do Python que implementa o algoritmo LDA para modelagem de tópicos.

Inicialmente foi criado um dicionário a partir da lista de palavras "stemizadas" que foram obtidas no pré-processamento de informações. O dicionário, gerenciado por um módulo da própria biblioteca Gensim, realizou o mapeamento entre as palavras e seus identificadores.

[5] https://www.python.org/
[6] https://www.nltk.org/
[7] https://pypi.org/project/gensim/

A partir do dicionário, aplicou-se o conceito da BoW para mapear a frequência de ocorrência de cada palavra, resultando na criação do *corpus* que possui a finalidade de extrair características do texto para o LDA.

Em seguida foi estabelecido a quantidade de repetições e o número de tópicos proporcionando as condições de execução para o algoritmo. Baseado nas informações obtidas na literatura, foi definido uma quantidade equivalente à 50 (cinquenta) repetições, ou seja, a quantidade de vezes que o LDA irá executar sobre o *corpus*. Também foi estabelecido a quantidade de tópicos que, por meio da média dos valores de coerência, encontrou-se a quantidade ideal de 16 tópicos.

Por fim, para o treinamento do algoritmo, foram submetidos como parâmetro: o dicionário, o *corpus*, o número de repetições e a quantidade de tópicos, resultando no modelo LDA já treinado com os 16 tópicos estabelecidos.

4.4.2. Classificação

Nesta etapa, cada publicação coletada foi submetida individualmente ao modelo LDA já treinado na primeira etapa, visando classificar as publicações por relevância de tópicos.

Assim, para cada conteúdo classificado foi atribuído uma nota (*Score*) entre 0 e 1, sendo 1 correspondente à 100% de compatibilidade em relação aos tópicos previamente determinados, ou seja, será levado em consideração as associações entre publicação e tópico que obtiveram maior nota.

Por fim, todos os valores atribuídos pelo modelo de tópicos para cada publicação foram armazenados em um banco de dados para serem analisados posteriormente.

4.5. Análise de Dados

Nesta etapa, foi realizada a interpretação dos resultados obtidos na classificação do algoritmo, onde foram definidos os assuntos ou tema para cada tópico.

Além disso, adotou-se o critério de seleção das publicações que atingiram a nota de classificação igual ou maior que 0,7 para um determinado tópico, referentes ao período entre 2015 à 2017. Este critério foi estabelecido com o intuito de maximizar a compatibilidade entre o assunto abordado e o tópico.

A partir da criação dos tópicos pelo algoritmo LDA, foi necessário uma verificação manual de cada publicação a fim de analisar a taxa de acertos e erros da classificação, bem como averiguar a precisão do algoritmo.

Adicionalmente, com base nos tópicos determinados, realizou-se uma análise temporal associando os picos de publicações com os acontecimentos no cenário político naquele momento em específico.

Por fim, foi feito uma distribuição dos tópicos de acordo com a classificação ideológica dos partidos identificados como direita, centro e esquerda, com intuito de verificar a incidência de publicações abordadas pelas frentes ideológicas para cada tópico.

Capítulo 5

Resultados e Discussões

Neste capítulo serão apresentados os resultados encontrados, trazendo informações do conjunto dos dados coletados, detalhes da modelagem de tópicos e as análises realizadas para a validação da hipótese.

5.1. Dados Coletados

O conjunto de dados coletados a partir das páginas públicas na rede social Facebook de 78 Senadores da República resultou em um total de 199.495 publicações. Com o intuito de analisar o conteúdo das postagens, considerou-se apenas as publicações que continham mensagens de texto, uma vez que, o algoritmo utilizado é capaz de analisar somente conteúdos em formato de texto. Assim, foram analisadas um total de 182.558 (91,5%) publicações e o restante das informações foram descartadas por apresentar apenas algum tipo de mídia como fotos e vídeos que não estão disponíveis pela API. Em complemento, foram coletados 13.217.555 comentários, mas somente 12.510.808 (94,6%) comentários estavam vinculados às publicações que continham mensagem de texto.

Ao realizar o levantamento de dados, considerando as publicações até o ano de 2018, foi possível observar um crescimento considerável do uso da rede social pelo público alvo, como pode ser visto na Tabela 5.1.

De acordo com as informações apresentadas na Tabela 5.1, observa-se que houve um aumento gradativo na quantidade de publicações ao longo dos anos. No entanto, em 2018 a quantidade de publicações foi visivelmente inferior aos demais anos, pois trata-se de resultados parciais. Nesse período, ocorreu o escândalo[1] envolvendo o Facebook com a venda de dados para empresa Cambridge Analytica, o que resultou em uma drástica alteração na política de privacidade da rede social, alterando as restrições de acesso a API e impossibilitando a

[1] <https://www.bbc.com/news/technology-43557803>

Tabela 5.1. Total de publicações por ano

Ano	Publicações
$\leq$ 2011	5.053
2012	10.831
2013	16.985
2014	28.867
2015	34.092
2016	36.985
2017	41.507
2018*	8.238

*Dados referentes aos conteúdos publicados até o dia 26/02/2018, quando foi finalizada a coleta das publicações para o presente trabalho.

coleta integral das publicações no ano de 2018.

Com o intuito de analisar outras informações de senadores efetivamente em seu tempo de exercício, também foram coletados os comentários das publicações a partir do período eleitoral de 2014, considerando os anos de 2015, 2016 e 2017.

Nesta abordagem, assim como os critérios adotados para a coleta de publicações, foram contabilizados apenas os comentários relacionadas com as publicações coletadas, o que resultou em um total de 12.340.059 comentários. Ao classificá-los por ano, a contagem considerou apenas os comentários correspondentes ao ano da publicação, ou seja, comentários realizados em anos posteriores foram desconsiderados, como pode ser observado na Tabela 5.2.

Tabela 5.2. Total de comentários por ano

Ano	Comentários	Publicações	Média por publicação	Desvio padrão
2015	2.830.033	34.092	83,0	468,9
2016	4.883.991	36.985	132,0	729,3
2017	4.626.035	41.507	111,4	669,9

Na Tabela 5.2 é possível observar que houve um aumento significativos na quantidade de comentários no ano de 2016. Tal fenômeno pode estar associado aos inúmeros eventos políticos que ocorreram naquele ano, por exemplo:

- **Impeachment da Dilma Rousseff**: Becker et al. (2017) relatam que o processo de impeachment teve início em setembro de 2015, quando a Câmara Federal aceitou a acusação contra Dilma pelo crime de responsabilidade fiscal associada a má gestão governamental e possível ligação com o esquema de corrupção investigado pela Operação Lava Jato. No dia 17/04/2016, a Câmara Federal aprovou o processo, dando sequência para a próxima etapa no Senado Federal que, por sua vez, aprovou o impeachment da

Presidente Dilma Rousseff com 60 votos a favor e 21 votos contra no dia 31/08/2016. Esta situação resultou em diversas manifestações populares pró e contra o processo de impeachment.

- **PEC 241/55**: De acordo com Amaral (2016), no dia 26/10/2016 o Congresso Nacional aprovou a Proposta de Emenda à Constituição nº 241 (PEC 241) que posteriormente foi renomeada para PEC nº 55/2016. Esta proposta gerou polêmica, uma vez que, o projeto buscava instituir um Novo Regime Fiscal que consistia em "congelar", nos valores de 2016, as despesas primárias do Poder Executivo, Legislativo, Judiciário, Ministério Público e Tribunal de Contas da União (TCU) em um período de 20 anos.
- **Prisão do ex-presidente da Câmara dos Deputados Eduardo Cunha**: Em 19/10/2016 foi determinado a prisão preventiva de Eduardo Cunha por decisão do ex-Juiz Federal Sérgio Moro como parte da Operação Lava Jato. No processo, Cunha foi acusado por corrupção, lavagem de dinheiro e evasão de divisas na ocasião de receber propina em um contrato de exploração de petróleo em Benin (África) e utilizar contas na Suíça para lavagem de dinheiro.[2,3]
- **Luiz Inácio Lula da Silva torna-se réu pela 5ª vez na Operação Lava Jato**: No dia 19/12/2019 estava sendo divulgada em todas as fontes midiáticas a notícia de que o então, Juiz Federal Sérgio Moro havia aceito a denúncia contra o ex-presidente Lula.

  Dessa forma, Lula passou a ser réu em cinco ações penais sendo: Três ações pela Operação Lava Jato (duas ações no Paraná e uma em Brasília), Operação Janus (desdobramento da Operação Lava Jato) e Operação Zelotes.[4,5,6]

Para o presente trabalho, também foi realizado um levantamento da quantidade de publicações por partido político e seus respectivos membros, totalizando em 19 grupos, como pode ser observado na Tabela 5.3.

Deste modo, observou-se que PMDB foi o partido com maior número de membros e consequentemente com o maior número de publicações. No entanto, em termos de média de publicações por membro, houve destaque para o partido Podemos (PODE) com uma média equivalente a 3773,6 publicações. Por outro lado, a menor quantidade de publicações foi indicada ao Senador José Reguffe, que no período de análise não possuía partido definido (identificado como Sem Partido).

[2] <http://g1.globo.com/pr/parana/noticia/2016/10/juiz-federal-sergio-moro-determina-prisao-de-eduardo-cunha.html>
[3] <https://veja.abril.com.br/brasil/preso-desde-2016-em-curitiba-eduardo-cunha-deve-ser-transferido-para-o-rj/>
[4] <https://politica.estadao.com.br/blogs/fausto-macedo/moro-aceita-denuncia-e-lula-vira-reu-pela-5a-vez/>
[5] <https://noticias.uol.com.br/politica/ultimas-noticias/2016/12/19/lula-vira-reu-pela-quarta-vez-na-lava-jato.htm>
[6] <http://g1.globo.com/pr/parana/noticia/2016/12/ex-presidente-lula-e-mais-sete-viram-reus-em-processo-da-lava-jato.html>

Tabela 5.3. Total de publicações por membros dos partidos

Partido	Publicações	Membros	Média	Desvio padrão
DEM	6.551	3	2.183,6	1.041,9
PCdoB	3.515	1	3.515	0
PDT	4.617	2	2.308,5	316,1
PMDB	41.567	21	1.979,3	1.138,5
PODE	11.321	3	3.773,6	353,1
PP	13.281	6	2.213,5	1.029,8
PPS	2.444	1	2.444	0
PR	8.485	4	2.121,2	1.248,2
PRB	1.916	1	1.916	0
PROS	1.288	1	1.288	0
PSB	11.765	4	2.941,2	1.258,6
PSC	993	1	993	0
PSD	5.511	4	1.377,7	601,7
PSDB	28.305	12	2.358,7	1.067,5
PT	31.978	9	3.553,1	733,6
PTB	3.387	2	1.693,5	272,2
PTC	2.908	1	2.908	0
REDE	2.535	1	2.535	0
Sem Partido*	191	1	191	0

*Senador Reguffe não havia vínculo partidário no momento da coleta.

Em termos de publicações por candidato, o Senador Jorge Viana do PT apresentou o maior quantidade de publicações, com um total de 5.101 e a menor quantidade foi identificada para o Senador Airton Sandoval do PMDB com valor equivalente a 94 publicações.

Com relação aos comentários das publicações, também foi realizado um levantamento da quantidade de comentários por partido político e seus respectivos membros, como pode ser observado na Tabela 5.4. Nesta Tabela, verifica-se que o PT foi o partido que apresentou maior número de comentários com valor equivalente a 3.494.882 e o menor valor foi observado para o Partido Social Cristão (PSC), com um total de 14.218 comentários. Em termos de média, partido PODE destacou-se com maior valor, totalizando 626.782,3 comentários e a menor média foi verificada no partido, que consequentemente, apresentou menor quantidade de comentários.

A considerável quantidade de comentários detectados no PT, pode estar associada aos eventos citados anteriormente, o qual teve como participação os políticos desse partido. Tais comentários podem estar relacionados tanto à críticas como também em apoio ao processo do impeachment da Dilma Rousseff e ao Lula na acusação da Operação Lava Jato.

Utilizando as classificações partidárias adaptadas de (TAROUCO; MADEIRA, 2015; CODATO et al., 2018), o cenário político em que os Senadores estavam inseridos durante o período de análise, foram classificados por frente ideológica, verificando a existência de 5 partidos identificados como Esquerda com 17 Senadores, 4 partidos como Centro com

Tabela 5.4. Total de comentários por membros dos partidos

Partido	Publicações	Membros	Comentários	Média	Desvio padrão
DEM	6.551	3	1.268.789	422.929,6	643.965,9
PCdoB	3.515	1	235.020	235.020	0
PDT	4.617	2	56.751	28.375,5	11.290,3
PMDB	41.567	21	1.107.402	52.733,4	81.644,4
PODE	11.321	3	1.880.347	626.782,3	361.678,5
PP	13.281	6	440.322	73.387	120.755,2
PPS	2.444	1	500.622	500.622	0
PR	8.485	4	1.241.364	310.341	605.720,6
PRB	1.916	1	18.982	18.982	0
PROS	1.288	1	25.407	25.407	0
PSB	11.765	4	78.082	19.520,5	2.376,0
PSC	993	1	14.218	14.218	0
PSD	5.511	4	179.380	44.845	25.351,8
PSDB	28.305	12	1.720.329	143.360,7	265.829,8
PT	31.978	9	3.494.882	388.320,2	556.964,7
PTB	3.387	2	41.492	20.746	7.708,8
PTC	2.908	1	26.018	26.018	0
REDE	2.535	1	132.415	132.415	0
Sem Partido*	191	1	48.986	48.986	0

35 Senadores e 9 partidos como Direita com 25 Senadores. No entanto, as informações do Senador Reguffe não foram contabilizadas nesta classificação, haja vista que, o mesmo não possuía filiação partidária e consequentemente não foi possível o seu enquadramento em quaisquer frentes ideológicas. A partir disso, averiguou-se a quantidade de publicação e comentários por frentes ideológicas.

A Tabela 5.5 contém as informações referentes às publicações para cada frente ideológica.

Tabela 5.5. Total de publicações por frente ideológica

Frente Ideológica	Publicações	Partidos	Média	Desvio padrão
Esquerda	54.410	5	10.882	12.339,1
Centro	73.604	4	18.401	19.851,3
Direita	54.353	9	6.039,2	4.274,4

Apesar do Centro possuir a menor quantidade de partidos, foi a frente ideológica que apresentou o maior número de publicações, com total de 73.604 e consequentemente maior média, apresentando valor igual a 18.401. É possível ressaltar que nesta frente estão enquadrados alguns dos partidos com maior número de membros, podendo citar o PMDB e o Partido da Social Democracia Brasileira (PSDB), o qual refletiu diretamente nos dados

acima apresentados.

Na Tabela 5.6 foi verificado a quantidade de comentários por frente ideológica. Diferente do observado na quantidade de publicações, a Direita apresentou a maior quantidade de comentários quando comparada com as demais frentes, resultando em valor igual 5.110.912 comentários. No entanto, em relação à média, o valor máximo foi observado no Centro, com média equivalente a 838.440.

Tabela 5.6. Total de comentários por frente ideológica

Frente Ideológica	Publicações	Partidos	Comentários	Média	Desvio padrão
Esquerda	54.410	5	3.997.150	799.430	1.508.380,2
Centro	73.604	4	3.353.760	838.440	736.028,1
Direita	54.353	9	5.110.912	567.879,1	708.268,7

Apesar do PT ser o partido com o maior número de comentários, como analisado anteriormente, o total de comentários obtidos nos demais partidos enquadrados como Esquerda correspondem a aproximadamente 13% em relação a quantidade total de comentários apresentados por essa frente ideológica.

Ainda que tenha sido coletado uma grande quantidade de dados envolvendo números de publicações e comentários, as etapas seguintes do presente trabalho considerou apenas as publicações que apresentavam mensagem de texto, ou seja, os dados referentes aos comentários não foram utilizados para a modelagem de tópicos. Esta restrição se fez necessária, tendo como finalidade de direcionar os principais assuntos abordados pelo público alvo, sendo que os comentários, de modo geral, são realizados por quaisquer usuários da rede social.

5.2. Interpretação dos Tópicos

Ao realizar uma análise minuciosa do conteúdo de cada publicação, contabilizou-se um total de 12.700.145 palavras. No entanto, no pré-processamento das informações foram descartadas um total de 6.717.179 palavras que eram consideradas termos de ligação que não denotavam sentido ao texto ou *stopwords*. Assim, somente as 5.982.966 palavras restantes foram submetidas ao treinamento do algoritmo LDA.

Com o intuito de obter um valor ideal de tópicos para o LDA, foi utilizado a ferramenta *MAchine Learning for LanguagE Toolkit* (Mallet)[7] que permitiu a execução de simulações do LDA para identificar qual o melhor número de tópicos, partindo da medida da coerência. Para tanto, foram realizadas três execuções variando o número de tópicos de 0 a 48 com um incremento igual 2 para analisar os valores da coerência. Além disso, foi definido o número de passadas equivalente a 50, tendo como finalidade controlar a frequência de um

[7] <http://mallet.cs.umass.edu/>

loop sobre cada documento do *corpus*.

A partir do ajuste dos dados da Figura 5.1 por meio de uma expressão matemática (exponencial) adequada ao seu comportamento, observa-se que 16 tópicos são suficientes para representar 92% da coerência máxima (31 tópicos), sendo escolhido como valor satisfatório.

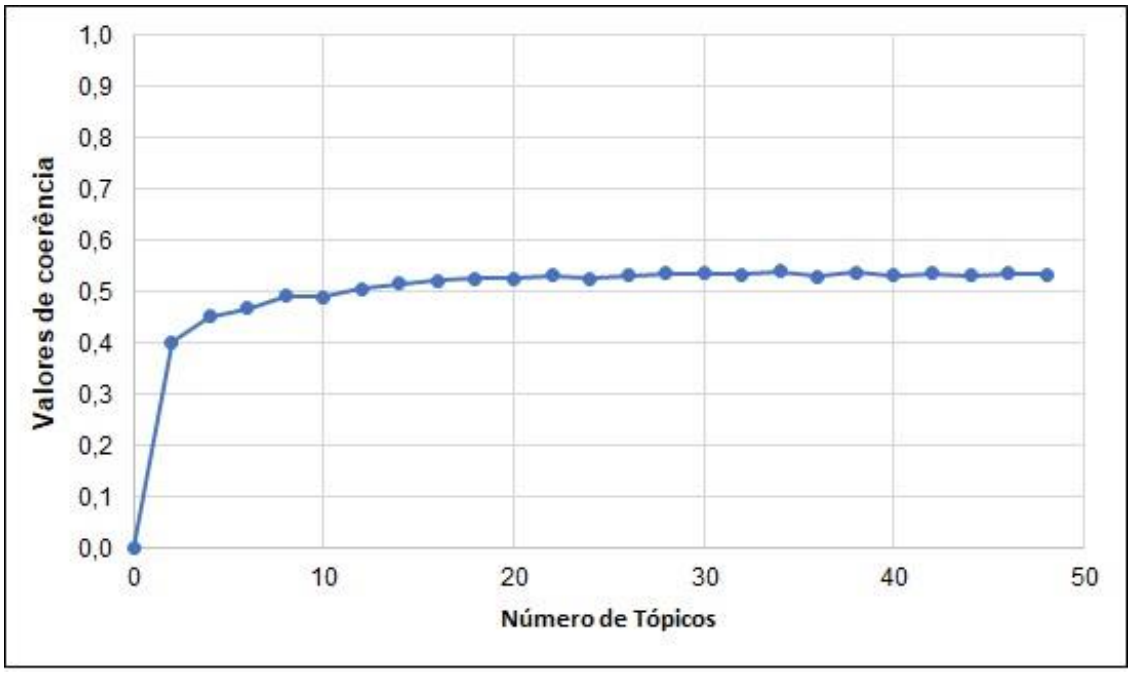

Figura 5.1. Valor médio da coerência obtidos pela ferramenta Mallet

Após a definição da quantidade de tópicos e o número de passadas, a etapa seguinte corresponde ao treinamento e execução do algoritmo LDA, onde foi possível verificar os tópicos gerados com o respectivo valor atribuído a cada palavra em termos de significância, como pode ser observado na Figura 5.2.

A partir da determinação dos principais termos representativos, pode-se identificar, por meio de associação, um assunto específico para cada tópico. Assim, cada tópico foi associado ao seu respectivo assunto, como pode ser observado na Tabela 5.7.

Nota-se que apesar do uso da coerência (veja Figura 5.2) para a determinação da quantidade de tópicos ideais, o algoritmo gerou tópicos com assuntos em comum, ou seja, os tópicos 2 e 15 apresentaram termos associados ao assunto “Diário de Bordo”, bem como, os tópicos 6 e 14 também apresentaram termos relacionados ao assunto “Desenvolvimento do País”.

Após a definição do assunto de cada tópico, todas as publicações foram submetidas a classificação individualmente, com a finalidade de direcionar o conteúdo da mensagem ao seu respectivo tópico, conforme o exemplo da Tabela 5.8.

Nesta etapa, cada conteúdo textual obteve um valor para cada assunto e sua classificação foi determinada a partir do maior valor atribuído entre os tópicos. Dessa forma, o maior valor atribuído indica que a significância das palavras da publicação eram

0	0.042*"president" + 0.030*"dilm" + 0.024*"govern" + 0.021*"vot" + 0.019*"senador" + 0.018*"polít" + 0.017*"lul" + 0.017*"impeachment" + 0.016*"brasil" + 0.014*"contr"
1	0.050*"tod" + 0.031*"trabalh" + 0.016*"melhor" + 0.016*"vam" + 0.015*"lut" + 0.015*"amig" + 0.013*"quer" + 0.012*"aqu" + 0.012*"vid" + 0.011*"junt"
2	0.031*"projet" + 0.026*"aprov" + 0.023*"sen" + 0.019*"senador" + 0.017*"trabalh" + 0.015*"servidor" + 0.014*"propost" + 0.013*"federal" + 0.012*"comissã" + 0.012*"pod"
3	0.026*"prefeit" + 0.018*"deput" + 0.015*"particip" + 0.015*"receb" + 0.012*"part" + 0.012*"senador" + 0.011*"vereador" + 0.011*"paul" + 0.010*"encontr" + 0.010*"lider"
4	0.027*"energ" + 0.018*"mat" + 0.018*"agricultur" + 0.016*"produtor" + 0.016*"wild" + 0.015*"gross" + 0.011*"jorg" + 0.010*"terr" + 0.010*"min" + 0.009*"famili"
5	0.032*"contr" + 0.022*"crim" + 0.018*"direit" + 0.012*"violênc" + 0.011*"políc" + 0.010*"justic" + 0.010*"milit" + 0.010*"segur" + 0.010*"mort" + 0.009*"lut"
6	0.018*"govern" + 0.014*"públic" + 0.011*"aument" + 0.008*"impost" + 0.008*"empres" + 0.007*"cont" + 0.007*"pag" + 0.007*"milhõ" + 0.007*"med" + 0.007*"bilhõ"
7	0.040*"pesso" + 0.027*"saúd" + 0.024*"deficient" + 0.022*"magn" + 0.020*"malt" + 0.015*"cas" + 0.014*"gal" + 0.013*"médic" + 0.012*"doenc" + 0.012*"imag"
8	0.023*"parabéns" + 0.023*"anos" + 0.022*"tod" + 0.020*"cidad" + 0.020*"hoj" + 0.012*"comemor" + 0.012*"imag" + 0.011*"fot" + 0.011*"filh" + 0.011*"sant"
9	0.020*"pod" + 0.013*"tod" + 0.010*"precis" + 0.009*"brasil" + 0.009*"outr" + 0.009*"faz" + 0.008*"govern" + 0.008*"porqu" + 0.007*"quer" + 0.007*"dev"
10	0.057*"sobr" + 0.027*"fal" + 0.025*"jornal" + 0.020*"entrev" + 0.020*"program" + 0.017*"acompanh" + 0.017*"víd" + 0.017*"viv" + 0.015*"sen" + 0.013*"assist"
11	0.036*"educ" + 0.023*"mulh" + 0.017*"escol" + 0.012*"estud" + 0.012*"brasil" + 0.012*"professor" + 0.012*"país" + 0.011*"ensin" + 0.011*"crianc" + 0.010*"públic"
12	0.022*"municípi" + 0.020*"recurs" + 0.017*"estad" + 0.014*"saúd" + 0.014*"obras" + 0.014*"govern" + 0.012*"cidad" + 0.010*"emend" + 0.010*"milhõ" + 0.010*"projet"
13	0.022*"brasil" + 0.020*"futebol" + 0.019*"esport" + 0.017*"brasileir" + 0.017*"cultur" + 0.016*"livr" + 0.015*"jog" + 0.013*"mund" + 0.013*"histór" + 0.012*"homenag"
14	0.023*"brasil" + 0.021*"país" + 0.018*"desenvolv" + 0.014*"setor" + 0.013*"empreg" + 0.013*"import" + 0.013*"econom" + 0.012*"produt" + 0.012*"econôm" + 0.012*"empres"
15	0.051*"senador" + 0.023*"ministr" + 0.018*"federal" + 0.017*"president" + 0.015*"nest" + 0.014*"jos" + 0.014*"sen" + 0.014*"nacional" + 0.013*"comissã" + 0.011*"reuniã"

Figura 5.2. Definição dos tópicos pelo algoritmo LDA

compatíveis ao tópico.

Na etapa seguinte foi realizada uma classificação manual para validar a eficiência do algoritmo, o qual será descrito detalhadamente na próxima seção.

5.3. Classificação Manual

Na classificação manual dos dados, foi realizado o agrupamento dos assuntos equivalentes resultando em 14 tópicos.

Além disso, adotou-se o critério de considerar apenas as publicações no período entre 2015 à 2017 (112.584 publicações) que obtiveram uma nota de classificação $\geq 0{,}7$, esperando uma maior confiabilidade na avaliação do algoritmo em associar o conteúdo textual para cada

Tabela 5.7. Assunto dos tópicos

Tópico	Assunto
0	Impeachment Presidente Dilma
1	Mensagens Motivacionais e Agradecimentos
2	Diário de Bordo
3	Apoio a Parceiros
4	Agricultura
5	Segurança Pública
6	Desenvolvimento do País
7	Saúde
8	Felicitações aos Municípios Aniversariantes
9	Protestos
10	Convites para a população
11	Educação e Direitos das Mulheres
12	Destinação de Verbas
13	Esporte
14	Desenvolvimento do País
15	Diário de Bordo

Tabela 5.8. Exemplo de classificação do algoritmo LDA

Mensagem original	Mensagem processada	Tópico	*Score*
Seminário sobre Os desafios da Escola Pública	seminári sobr desafi escol públic	0	0.0104
		1	0.0104
		2	0.0104
		3	0.0104
		4	0.0104
		5	0.0104
		6	0.0104
		7	0.0104
		8	0.0104
		9	0.0104
		10	0.0104
		11	0.8437
		12	0.0104
		13	0.0104
		14	0.0104
		15	0.0104

assunto identificado. Dessa forma, foram classificadas manualmente 7.694 publicações que correspondem a 6,8% do total de publicações, conforme a Tabela 5.9.

Ainda, de acordo com a Tabela 5.9, observou-se que houve destaque aos assuntos relacionados a Felicitações aos Municípios Aniversariantes, com total de 1.901 publicações e

Tabela 5.9. Quantidade de publicações com nota de classificação $\geq$ 0,7 utilizadas na classificação manual

Tópico	Assunto	Quantidade
0	Impeachment Presidente Dilma	629
1	Mensagens Motivacionais e Agradecimentos	1.103
2	Diário de Bordo	1.525
3	Apoio a Parceiros	819
4	Agricultura	7
5	Segurança Pública	24
6	Desenvolvimento do País	146
7	Saúde	37
8	Felicitações aos Municípios Aniversariantes	1.901
9	Protestos	428
10	Convites para a população	371
11	Educação e Direitos das Mulheres	157
12	Destinação de Verbas	437
13	Esporte	110

Diário de Bordo, com valor equivalente a 1.525 publicações.

Nota-se que assuntos considerados pilares para o desenvolvimento de um País, como a agricultura, segurança pública, saúde e educação, não tiveram quantidade de publicações relevantes, quando comparados aos assuntos de divulgação das atividades políticas. De modo geral, observou-se que a maioria das publicações do público alvo estava direcionada em promover suas atividades como senadores, ao invés de abordar demais assuntos relevantes ao desenvolvimento do País.

Com o intuito de verificar a precisão de avaliação do LDA, foi analisado o desempenho de classificação considerando a taxa de acerto de cada publicação em relação ao seu respectivo assunto, como pode ser observado na Tabela 5.10.

Segundo a Tabela 5.10, foi possível observar que nem todos os tópicos foram classificados corretamente. Assuntos como agricultura (100%), segurança pública (89%), saúde (97%) e esporte (98%) apresentaram alta porcentagem de erro. Os demais assuntos, exceto Impeachment da Presidente Dilma (42% de acerto), apresentaram a quantidade de acerto acima de 50%. Em termos de média geral, o algoritmo utilizado foi capaz de classificar o conteúdo textual com precisão de aproximadamente 68% de acerto em relação ao montante de 7.694 publicações.

Diante do total de 112.584 publicações referente ao período de análise, apenas 7.694 (6,8%) publicações foram avaliadas com nota $\geq$ 0,7. No entanto, somente 5.239 publicações foram classificadas corretamente. De modo geral, observou-se que o algoritmo LDA apresentou uma margem de acerto equivalente a 4,7% quando comparados com a quantidade total de 112.584 publicações, podendo notar um resultado abaixo do esperado.

Com o intuito de verificar o comportamento das publicações, também foi realizada uma análise temporal das publicações associando aos fatos ocorridos no período de coleta,

Tabela 5.10. Resultado da classificação manual

Tópico	Assunto	Quantidade	Acerto	(%)	Erro	(%)
0	Impeachment Presidente Dilma	629	265	42	364	58
1	Mensagens Motivacionais...	1.103	863	78	240	22
2	Diário de Bordo	1.525	1.081	71	444	29
3	Apoio a Parceiros	819	605	74	214	26
4	Agricultura	7	0	0	7	100
5	Segurança Pública	24	4	17	20	83
6	Desenvolvimento do País	146	87	60	59	40
7	Saúde	37	1	3	36	97
8	Felicitações aos Mun...	1.901	1.412	74	489	26
9	Protestos	428	202	47	226	53
10	Convites para a população	371	278	75	93	25
11	Educação e Direitos das Mulheres	157	92	59	65	41
12	Destinação de Verbas	437	347	79	90	21
13	Esporte	110	2	2	108	98
Total		7.694	5.239	68	2.455	32

bem como, o enquadramento das publicações por frente ideológica, o qual será abordado detalhadamente na próxima seção.

5.4. Análises Complementares

Nesta seção serão discutidas, em escala temporal, alguns assuntos identificados pelo algoritmo associados com os fatos ocorridos no meio político durante o período de coleta, bem como, os principais assunto abordados por frentes ideológicas.

Em relação a escala temporal, alguns tópicos apresentaram comportamento relevante ao longo dos anos de 2015, 2016 e 2017.

- **Impeachment da Dilma Rousseff**: De acordo com a Figura 5.3, nota-se que houve picos de publicações em dezembro de 2015, abril de 2016 e agosto de 2016. Segundo Becker et al. (2017), o processo de impeachment começou no segundo semestre de 2015, o qual o ex-Presidente da Câmara Federal do Deputados Eduardo Cunha acolheu a acusação formal sobre o crime de responsabilidade fiscal contra a ex-presidente. Em abril de 2016 ocorreu a votação na Câmara Federal dos Deputados e a aprovação por 367 votos a favor e 137 contra a destituição, em que ficou determinado o afastamento da ex-presidente Dilma por 180 dias e o encaminhamento do processo ao Senado Federal. Somente em agosto de 2016, foi finalizado o processo de impeachment após a aprovação por votação no Senado Federal com 60 votos a favor e 21 contrários, dando poder ao vice-presidente Michel Temer assumir o cargo como Presidente da República Interino.
- **Diário de Bordo**: Na Figura 5.4, observou-se que ocorreu uma grande variação de publicações em junho de 2017, a qual pode estar associado aos acontecimentos referentes

Figura 5.3. Linha temporal das publicações do Impeachment de Dilma Rousseff

à aprovação da Lei nº 13.467/2017 que previa a Reforma Trabalhista. Carvalho (2017) afirma que a Lei nº 13.467, de 13 de julho de 2017, resultou em profundas alterações no ordenamento jurídico que regula as relações trabalhistas desde a instituição da Consolidação das Leis do Trabalho (CLT) em 1943. A Reforma Trabalhista resultou em diversas debates no meio político e movimentos populares contra as alterações, pois muitos acreditavam que poderia ocorrer um aumento das desigualdades no mercado de trabalho, com alteração na jornada de trabalho, baixa remuneração de horas extras e consequentemente, a elevação dos efeitos adversos sobre a saúde e os acidentes de trabalho.

- **Apoio ao parceiros**: Ao analisar a Figura 5.5, verificou-se que houve um pico de publicações em setembro de 2016, seguido por uma queda brusca nos meses seguintes. Tal fato pode estar relacionado ao apoio dos senadores aos seus candidatos no período de eleições municipais que ocorreu em outubro de 2016.
- **Felicitações aos municípios**: De acordo com a Figura 5.6, destaca-se que houve diversas variações na quantidade de publicações com comportamentos similares ao longo dos anos, o qual estão associadas às felicitações aos municípios pertencentes aos estados que o senador representa. Dentre as publicações destaca-se os senadores da região nordeste do país.

Por fim, as publicações de cada tópico identificadas pelo algoritmo foram distribuídas de acordo com a frente ideológica do autor, como podemos verificar na Figura 5.7. Nota-se que o tópico referente ao assunto Agricultura não foi levado em consideração pelo fato de não obter nenhum acerto na classificação. Também não foram consideradas as publicações

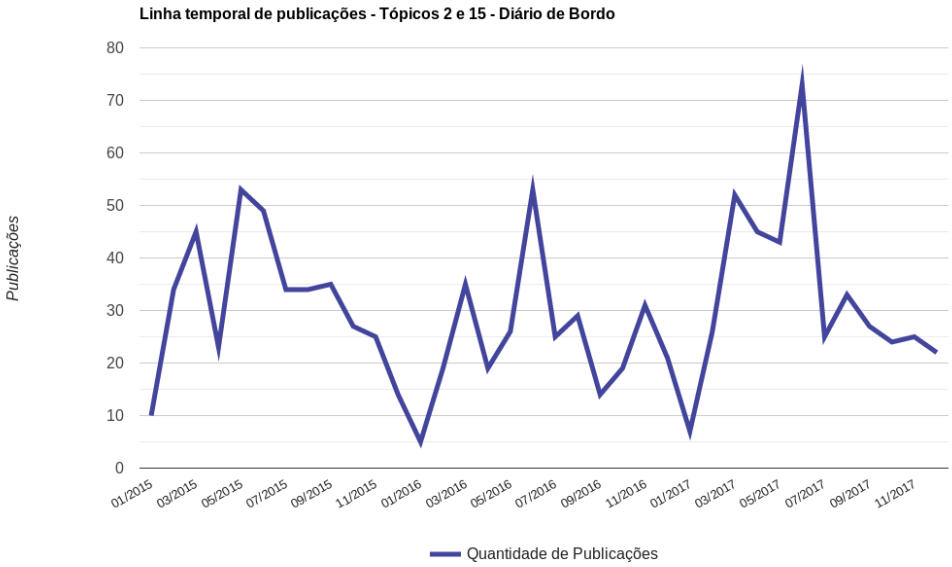

Figura 5.4. Linha Temporal das publicações do Tópico Diário de Bordo

Figura 5.5. Linha temporal das publicações do Tópico Apoio a Parceiros

do Senador Reguffe, pois o mesmo não pertencia a nenhum partido na época da coleta dos dados.

Nesta abordagem, verificou-se uma maior participação da frente ideológica Centro na maioria dos tópicos. Esse fato pode estar diretamente relacionado a quantidade de Senadores dos partidos alocados nessa frente, uma vez que, os dois partidos com maior número de membros estão classificados como Centro.

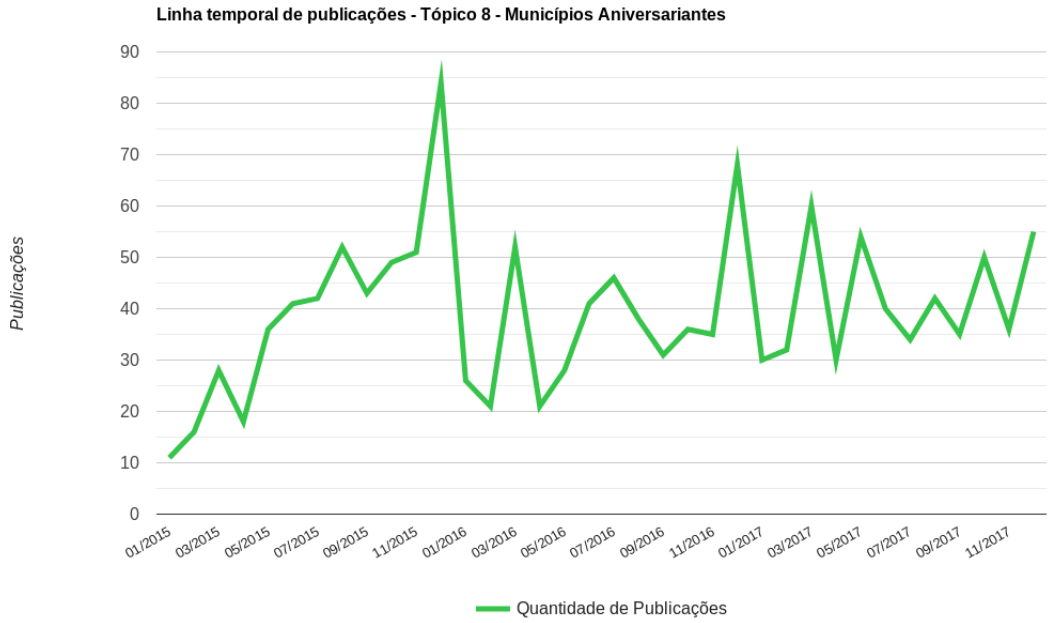

Figura 5.6. Linha temporal das publicações do Tópico Municípios Aniversariantes

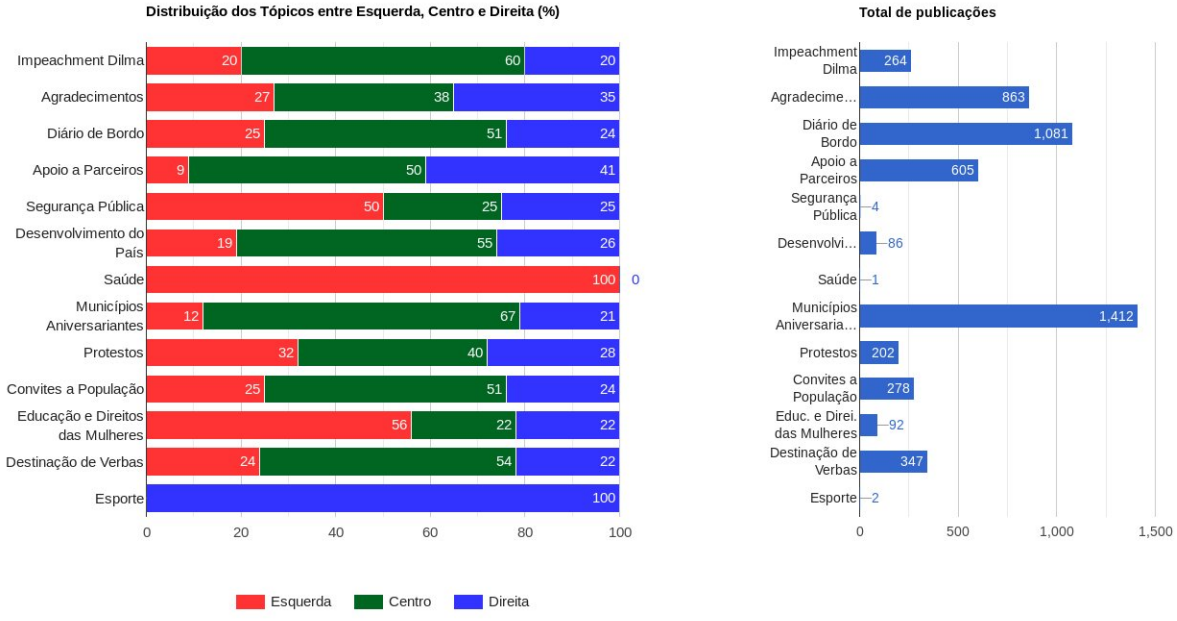

Figura 5.7. Distribuição das publicações dos tópicos por frente ideológica

Em termos de saúde, esporte e segurança pública, apesar dos valores significativos de porcentagem, não foi possível admitir quaisquer conclusões levando em consideração as frentes ideológicas, uma vez que, a quantidade de acertos foram relativamente muito baixas.

Capítulo 6

Conclusão

Para o presente trabalho, o Facebook foi escolhido como base de dados por ser a maior rede social da atualidade, bem como, foi possível evidenciar a magnitude da quantidade de informações geradas ao longo dos anos por meio da coleta de dados. Dessa forma, foi fundamental a aplicação mecanismos visando a compactação de dados com a finalidade de auxiliar na interpretação das informações.

Além disso, foi necessário o desenvolvimento do Web Crawler, uma vez que, não havia nenhuma ferramenta disponível para a realização da coleta de dados. Assim, a elaboração do coletor de dados em modo paralelo permitiu a otimização do tempo de coleta em aproximadamente duas semanas.

Tendo em vista a aplicação do algoritmo LDA em diversos trabalhos na literatura utilizando uma abordagem não supervisionada, buscou-se verificar a sua eficiência aplicando-o sobre a base de dados coletada. Em conjunto com a adoção de critério de avaliação, foi possível sintetizar as informações por meio de tópicos, permitindo a identificação dos principais assuntos abordados em publicações criadas pelo público alvo.

No presente trabalho, o algoritmo LDA foi submetido ao treinamento utilizando 182.558 publicações para a determinação de 16 tópicos previamente estabelecidos. No entanto, foram gerados dois tópicos com assuntos em comum e posteriormente foram agrupados, resultando em 14 tópicos válidos.

Ao realizar a classificação das publicações com o algoritmo LDA já treinado, direcionando o estudo para o período pós eleição de 2014, apenas 7.694 publicações receberam notas superior a 0,7. Tais publicações representam 6,8% em relação a quantidade de 112.584 publicações que corresponde aos anos de 2015, 2016 e 2017. Com esses dados, observou-se que a classificação adequada das publicações foi muito abaixo do esperado, pois representa uma pequena parcela das publicações.

Com base no montante de publicações classificadas, foi realizada uma verificação

manual com o intuito de apurar a quantidade de acertos e erros, o qual observou-se que o algoritmo apresentou 5.239 acertos (4,7%) e 2.455 erros (2,2%). Dessa forma, o algoritmo apresentou eficiência de classificação equivalente a 68%. Em termos de acerto, verificou-se um desempenho mediano do algoritmo, ao considerar a quantidade reduzida de publicações (6,8% do total).

Ao aplicar a análise de dados em escala temporal a partir das publicações classificadas corretamente, foi possível verificar o comportamento da rede social em relação a um determinado assunto encontrado pelo algoritmo LDA. Tópicos como "Impeachment da Dilma Rousseff" e "Apoio a Parceiros" apresentaram picos de publicações associados aos eventos que ocorreram no meio político. Assim, conclui-se que os fatos ocorridos no mundo real refletem diretamente nas redes sociais.

Neste contexto, para as condições estabelecidas neste trabalho, a aplicação do algoritmo LDA não foi capaz de sintetizar precisamente os assuntos relevantes em um determinado conjunto de dados. Por outro lado, as razões deste resultado negativo ainda necessitam ser melhor exploradas.

6.1. Sugestões para Trabalhos Futuros

- Explorar a eficiência do algoritmo LDA em um conjunto de dados reduzido e intervalo de tempo menor.
- Aplicar abordagens supervisionadas ou semi-supervisionadas a fim de identificar uma quantia maior de tópicos relevantes.
- Replicar o experimento utilizando o algoritmo *Latent Semantic Analysis* (LSA) proposto por Deerwester et al. (1990).

Apêndices

Apêndice

A

Tabela dos Senadores em Exercício

Tabela A.1. Senadores em exercício

Nome	Partido	Estado
Acir Gurgacz	PDT	RO
Aécio Neves	PSDB	MG
Airton Sandoval	PMDB	SP
Alvaro Dias	PODE	PR
Ana Amélia Lemos	PP	RS
Ângela Portela	PDT	RR
Antonio Anastasia	PSDB	MG
Antonio Carlos Valadares	PSB	SE
Armando Monteiro	PTB	PE
Ataídes Oliveira	PSDB	TO
Benedito de Lira	PP	AL
Cássio Cunha Lima	PSDB	PB
Cidinho Santos	PR	MT
Ciro Nogueira	PP	PI
Cristovam Buarque	PPS	DF
Dalirio Beber	PSDB	SC
Dário Berger	PMDB	SC
Davi Alcolumbre	DEM	AP
Eduardo Amorim	PSDB	SE
Eduardo Braga	PMDB	AM
Eduardo Lopes	PRB	RJ

Continua na próxima página

Tabela A.1 – Senadores em exercício

Nome	Partido	Estado
Elmano Férrer	PMDB	PI
Eunício Oliveira	PMDB	CE
Fátima Bezerra	PT	RN
Fernando Bezerra Coelho	PMDB	PE
Fernando Collor	PTC	AL
Flexa Ribeiro	PSDB	PA
Garibaldi Alves	PMDB	RN
Gladson Cameli	PP	AC
Gleisi Hoffmann	PT	PR
Hélio José	PROS	DF
Humberto Costa	PT	PE
Ivo Cassol	PP	RO
Jader Barbalho	PMDB	PA
João Alberto de Souza	PMDB	MA
João Capiberibe	PSB	AP
Jorge Viana	PT	AC
José Agripino	DEM	RN
José Maranhão	PMDB	PB
José Medeiros	PODE	MT
José Pimentel	PT	CE
José Serra	PSDB	SP
Kátia Abreu	PMDB	TO
Lasier Martins	PSD	RS
Lídice da Mata	PSB	BA
Lindbergh Farias	PT	RJ
Lúcia Vânia	PSB	GO
Magno Malta	PR	ES
Marta Suplicy	PMDB	SP
Omar Aziz	PSD	AM
Otto Alencar	PSD	BA
Paulo Bauer	PSDB	SC
Paulo Paim	PT	RS
Paulo Rocha	PT	PA
Pedro Chaves	PSC	MS

Continua na próxima página

Tabela A.1 – Senadores em exercício

Nome	Partido	Estado
Raimundo Lira	PMDB	PB
Randolfe Rodrigues	REDE	AP
Regina Sousa	PT	PI
Reguffe	Sem Partido	DF
Renan Calheiros	PMDB	AL
Ricardo Ferraço	PSDB	ES
Roberto Requião	PMDB	PR
Roberto Rocha	PSDB	MA
Romário Faria	PODE	RJ
Romero Jucá	PMDB	RR
Ronaldo Caiado	DEM	GO
Rose de Freitas	PMDB	ES
Sérgio Petecão	PSD	AC
Simone Tebet	PMDB	MS
Tasso Jereissati	PSDB	CE
Telmário Mota	PTB	RR
Valdir Raupp	PMDB	RO
Vanessa Grazziotin	PCdoB	AM
Vicentinho Alves	PR	TO
Waldemir Moka	PMDB	MS
Wellington Fagundes	PR	MT
Wilder Morais	PP	GO
Zeze Perrella	PMDB	MG

Referências

ABRAMOVA, Veronika; BERNARDINO, Jorge. Nosql databases: Mongodb vs cassandra. In: *Proceedings of the International C* Conference on Computer Science and Software Engineering*. New York, NY, USA: ACM, 2013. (C3S2E '13), p. 14–22. ISBN 978-1-4503-1976-8. Disponível em: <http://doi.acm.org/10.1145/2494444.2494447>.

Alashri, S.; Kandala, S. S.; Bajaj, V.; Ravi, R.; Smith, K. L.; Desouza, K. C. An analysis of sentiments on facebook during the 2016 u.s. presidential election. In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.: s.n.], 2016. p. 795–802.

ALEXANDROV, Daniel; GORGADZE, Alexey; MUSABIROV, Ilya. Virtual caucasus on vk social networking site. In: *Proceedings of the 8th ACM Conference on Web Science*. New York, NY, USA: ACM, 2016. (WebSci '16), p. 215–217. ISBN 978-1-4503-4208-7. Disponível em: <http://doi.acm.org/10.1145/2908131.2908205>.

AMARAL, Nelson Cardoso. Pec 241/55: a "morte" do pne (2014-2024) e o poder de diminuição dos recursos educacionais. *Revista Brasileira de Política e Administração da Educação - Periódico científico editado pela ANPAE*, v. 32, n. 3, p. 653–673, 2016. ISSN 2447-4193. Disponível em: <https://seer.ufrgs.br/rbpae/article/view/70262>.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. *Modern Information Retrieval*. 2nd. ed. USA: Addison-Wesley Publishing Company, 2008. ISBN 9780321416919.

BAQUERO, Marcello; BAQUERO, Rute Vivian Angelo; MORAIS, Jennifer Azambuja de. Socialização política e internet na construção de uma cultura política juvenil no sul do brasil. *Revista Educação e Sociedade*, v. 37, n. 137, p. 989–1008, 2016. ISSN 0101-7330.

BECKER, Camlia; CESAR, Camila; GALLAS, Débora; WEBER, Maria Helena. Manifestações. *Revista Latinoamericana de Ciências De La Comunicación*, v. 13, n. 24, p. 96–113, 2017.

BLEI, David M. Probabilistic topic models. *Commun. ACM*, ACM, New York, NY, USA, v. 55, n. 4, p. 77–84, abr. 2012. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2133806.2133826>.

BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. *J. Mach. Learn. Res.*, JMLR.org, v. 3, p. 993–1022, mar. 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944937>.

BOSSETTA, Michael. The digital architectures of social media: Comparing political campaigning on facebook, twitter, instagram, and snapchat in the 2016 u.s. election. *Journalism & Mass Communication Quarterly*, v. 95, n. 2, p. 471–496, 2018. Disponível em: <https://doi.org/10.1177/1077699018763307>.

BRAGA, Maria do Socorro Sousa; BOURDOUKAN, Adla. Partidos políticos no brasil: Organização partidária, competiç ao eleitoral e financiamento público. *Perspectivas: Revista de Ciências Sociais*, 2009. ISSN 0101-3459.

BRAGA, Sérgio; CARLOMAGNO, Márcio. Eleições como de costume? uma análise longitudinal das mudanças provocadas nas campanhas eleitorais brasileiras pelas tecnologias digitais (1998-2016). *Revista Brasileira de Ciênencia Política*, scielo, p. 7 – 62, 08 2018. ISSN 0103-3352. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-33522018000200007&nrm=iso>.

CARVALHO, Sandro Sacchet de. Uma visão geral sobre a reforma trabalhista. Instituto de Pesquisa Econômica Aplicada (Ipea), 2017.

CHEN, Min; MAO, Shiwen; LIU, Yunhao. Big data: A survey. *Mobile Netw Appl*, p. 171–209, 2014.

CODATO, Adriano; BERLATTO, Fábia; BOLOGNESI, Bruno. Tipologia dos políticos de direita no Brasil: uma classificação empírica. *Análise Social*, scielopt, p. 870 – 897, 12 2018. ISSN 0003-2573. Disponível em: <http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S0003-25732018000400002&nrm=iso>.

DEERWESTER, Scott; DUMAIS, Susan T.; FURNAS, George W.; LANDAUER, Thomas K.; HARSHMAN, Richard. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, v. 41, n. 6, p. 391–407, 1990.

ELLISON, Nicole B.; STEINFIELD, Charles; LAMPE, Cliff. Connection strategies: Social capital implications of facebook-enabled communication practices. *New Media & Society*, v. 13, n. 6, p. 873–892, 2011.

FILIMONOV, Kirill; RUSSMANN, Uta; SVENSSON, Jakob. Picturing the party: Instagram and party campaigning in the 2014 swedish elections. *Social Media + Society*, v. 2, 08 2016.

FREELON, Deen. Campaigns in control: Analyzing controlled interactivity and message discipline on facebook. *Journal of Information Technology & Politics*, Routledge, v. 14, n. 2, p. 168–181, 2017. Disponível em: <https://doi.org/10.1080/19331681.2017.1309309>.

GRIFFITHS, Thomas; STEYVERS, Mark; TENENBAUM, Joshua. Topics in semantic representation. *Psychological review*, v. 114, p. 211–44, 05 2007.

HAN, Jing; E, Haihong; LE, Guan; DU, Jian. Survey on nosql database. *Pervasive Computing and Applications (ICPCA), 2011 6th International Conference on*, 2011.

HOANG, Tuan-Anh; COHEN, William W.; LIM, Ee-Peng; PIERCE, Doug; REDLAWSK, David P. Politics, sharing and emotion in microblogs. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. New York, NY, USA: ACM, 2013. (ASONAM '13), p. 282–289. ISBN 978-1-4503-2240-9. Disponível em: <http://doi.acm.org/10.1145/2492517.2492554>.

JUNGHERR, Andreas. *Analyzing Political Communication with Digital Trace Data*. [S.l.]: Springer International Publishing, 2015.

JüRGENS, Pascal; JUNGHERR, Andreas. The use of twitter during the 2009 german national election. *German Politics*, Routledge, v. 24, n. 4, p. 469–490, 2015. Disponível em: <https://doi.org/10.1080/09644008.2015.1116522>.

KARLSEN, Rune; ENJOLRAS, Bernard. Styles of social media campaigning and influence in a hybrid political communication system: Linking candidate survey data with twitter data. *The International Journal of Press/Politics*, v. 21, n. 3, p. 338–357, 2016. Disponível em: <https://doi.org/10.1177/1940161216645335>.

KOBAYASHI, Tetsuro; IKEDA, Ken'ichi; MIYATA, Kakuko. Social capital online: Collective use of the internet and reciprocity as lubricants of democracy. *Information, Communication & Society*, v. 9, n. 5, p. 582–611, 2006.

KREISS, Daniel; MCGREGOR, SHANNON C. Technology firms shape political communication: The work of microsoft, facebook, twitter, and google with campaigns during the 2016 u.s. presidential cycle. *Political Communication*, Routledge, v. 35, n. 2, p. 155–177, 2018. Disponível em: <https://doi.org/10.1080/10584609.2017.1364814>.

LEV-ON, Azi; HALEVA-AMIR, Sharon. Normalizing or equalizing? characterizing facebook campaigning. *New Media & Society*, v. 20, 09 2016.

LEVATO, Vanina. Redes sociales, lenguaje y tecnología facebook. the 4th estate media? *Cuadernos del Centro de Estudios en Diseño y Comunicación*, n. 45, p. 65–77, 2013. ISSN 1668-5229.

LOBO, Luiz Carlos. Inteligência artificial e medicina. *Revista Brasileira de Educação Médica*, v. 41, n. 2, p. 185–193, 2017.

MADEIRA, Rafael Machado; TAROUCO, Gabriela da Silva. Esquerda e direita no brasil: Uma análise conceitual. *Revista Pós Ciências Sociais*, 2011. ISSN 1983-4527.

MARTELETO, Regina Maria. Análise de redes sociais - aplicação nos estudos de transferência da informação. *Ciência da Informação On-line version*, v. 30, n. 1, p. 71–81, 04 2001. ISSN 0100-1965. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652001000100009&nrm=iso>.

MARZ, Nathan; WARREN, James. *Big Data Principles and best pratices of scalable realtime data systems*. [S.l.]: Manning Early Access Program, 2014. v. 17.

MCAFEE, Andrew; BRYNJOLFSSON, Erik. Big data: The management revolution. *Harvard Bisuness Review*, 2012.

MORALES, Marsy Dayanna Ortiz; AGUILAR, Luis Joyanes; MARíN, Lillyana María Giraldo. Los desafíos del marketing en la era del big data. *e-Ciencias de la Información*, v. 6, n. 1, 2016.

MURTA, Felipe; ITUASSU, Arthur; CAPONE, Letícia; LEO, Luiz; ROVERE, Roberta La. Eleições e mídias sociais: Interação e participação no facebook durante a campanha para a câmara dos deputados em 2014. *Compolítica*, v. 7, n. 1, p. 47–72, jun. 2017. Disponível em: <http://compolitica.org/revista/index.php/revista/article/view/111>.

POWER, Timothy J.; Zucco Jr., Cesar. Estimating ideology of brazilian legislative parties. *Latin American Research Review*, v. 44, p. 218–246, 2009.

QIN, Zengchang; CONG, Yonghui; WAN, Tao. Topic modeling of chinese language beyond a bag-of-words. *Computer Speech and Language*, v. 40, p. 60 – 78, 2016. ISSN 0885-2308. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0885230816300626>.

QUAN-HAASE, Anabel; MCCAY-PEET, Lori. Social network analysis. *The International Encyclopedia of Communication Theory and Philosophy*, 2016. Disponível em: <http://dx.doi.org/10.1002/9781118766804.wbiect130>.

RECUERO, Raquel. Contribuições da análise de redes sociais para o estudo das redes sociais na internet: o caso da hashtag #tamojuntodilma e #calaabocadilma. *Revista Fronteiras – Estudos midiáticos*, v. 16, n. 2, p. 60–77, 2014.

SEVA, Jurica; DURIC, Bogdan Okresa; SCHATTEN, Markus. Visualizing public opinion in croatia based on available social network content. *European Quarterly of Political Attitudes and Mentalities*, v. 5, n. 1, p. 22–35, 2016. ISSN 2285-4916.

Song, M.; Kim, M. C.; Jeong, Y. K. Analyzing the political landscape of 2012 korean presidential election in twitter. *IEEE Intelligent Systems*, v. 29, n. 2, p. 18–26, Mar 2014. ISSN 1941-1294.

SOUZA, Bruno Á.; ALMEIDA, Thais G.; MENEZES, Alice A.; FIGUEIREDO, Carlos M. S.; NAKAMURA, Fabíola G.; NAKAMURA, Eduardo F. Uma abordagem para detecção de tópicos relevantes em redes sociais online. In: *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. Porto Alegre, RS, Brasil: SBC, 2017. ISSN 2595-6094. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/3264>.

SREEJA, S.R.; CHAUDHARI, Sangita. Review of web crawlers. *International Journal of Knowledge and Web Intelligence*, v. 5, n. 1, 2014.

STIER, Sebastian; BLEIER, Arnim; LIETZ, Haiko; STROHMAIER, Markus. Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter. *Political Communication*, Routledge, v. 35, n. 1, p. 50–74, 2018. Disponível em: <https://doi.org/10.1080/10584609.2017.1334728>.

STRAUCH, Christof. Nosql databases. *Selected Topics on Software-Technology Ultra-Large Scale Sites*, 2011.

TAROUCO, Gabriela da Silva; MADEIRA, Rafael Machado. Os partidos brasileiros segundo seus estudiosos: análise de um expert survey. *Civitas: Revista de Ciências Sociais*, v. 15, 2015.

VALENZUELA, Sebastián; PARK, Namsu; KEE, Kerk F. Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation. *Journal of Computer-Mediated Communication*, v. 14, p. 875–901, 2009.

VIEIRA, Aiane de Oliveira. As estratégias persuasivas dos presidenciáveis na corrida eleitoral de 2014 nas redes sociais. Universidade Federal de Goiás, 2016. Faculdade de Ciências Sociais - FCS (RG). Disponível em: <http://repositorio.bc.ufg.br/tede/handle/tede/6508>.

WILSON, Robert E.; GOSLING, Samuel D.; GRAHAM, Lindsay T. A review of facebook research in the social sciences. *Perspectives on Psychological Science*, v. 7, n. 3, p. 203–220, 2012.