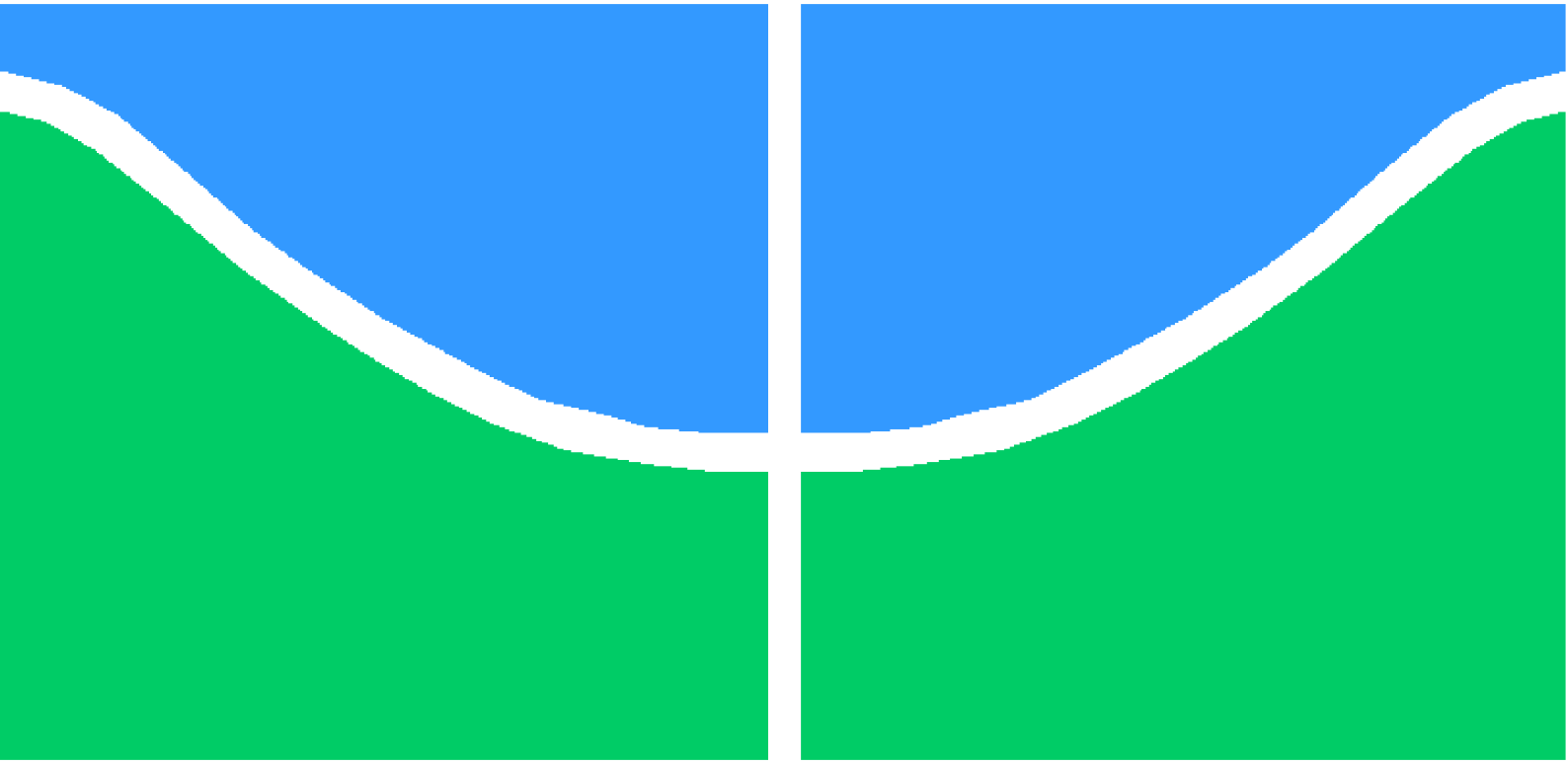


Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Engenharia de Software

Percepções sobre corrupção durante as eleições presidenciais no Brasil em 2018: uma análise baseada no Twitter.

Autor: Marcelo Cristiano Araujo Silva

Orientador: Prof. Dr. Wander Cleber M. Pereira da Silva

Brasília, DF

2019

Marcelo Cristiano Araujo Silva

Percepções sobre corrupção durante as eleições presidenciais no Brasil em 2018: uma análise baseada no Twitter.

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Orientador: Prof. Dr. Wander Cleber M. Pereira da Silva

Brasília, DF

2019

Marcelo Cristiano Araujo Silva

Percepções sobre corrupção durante as eleições presidenciais no Brasil em 2018: uma análise baseada no Twitter. / Marcelo Cristiano Araujo Silva. – Brasília, DF, 2019-

49 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Wander Cleber M. Pereira da Silva

Trabalho de Conclusão de Curso – Universidade de Brasília – UnB
Faculdade UnB Gama – FGA , 2019.

1. Twitter. 2. Corrupção. I. Prof. Dr. Wander Cleber M. Pereira da Silva. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Percepções sobre corrupção durante as eleições presidenciais no Brasil em 2018: uma análise baseada no Twitter.

CDU 02:141:005.6

Marcelo Cristiano Araujo Silva

Percepções sobre corrupção durante as eleições presidenciais no Brasil em 2018: uma análise baseada no Twitter.

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Trabalho aprovado. Brasília, DF, 12 de julho de 2019 – Data da aprovação do trabalho:

Prof. Dr. Wander Cleber M. Pereira da Silva
Orientador

Prof. Dr. Glauco Vitor Pedrosa
Convidado 1

Prof. Dr. John Lenon Cardoso Gardenghi
Convidado 2

Brasília, DF
2019

Dedico esse trabalho a minha mãe, Maria Míriam, e ao meu pai, Marcelo Cristiano (in memorian), que nunca deixaram de me apoiar.

Agradecimentos

À universidade, pela confiança em ter me acolhido como um de seus alunos, permitindo assim a realização desse trabalho.

Ao professor Dr. Wander Cleber, que me orientou durante a jornada da finalização de minha graduação.

À minha família, Maria Miriam e Macrysla Yohanna, que foram meu lar sempre que voltei para casa.

Ao meu pai Marcelo Cristiano, que não está mais aqui, mas que se orgulharia dessa conquista.

Aos meus amigos, que fazem com que os dias mais difíceis tornem-se menos penosos.

À Ludimila Bela, por nunca ter deixado de acreditar que eu poderia ir mais longe.

"In god we trust.
All others must bring data."
(Edwin R. Fisher, 1978)

Resumo

Compreender a corrupção e suas nuances, e como elas afetam o cidadão, pode ser um bom caminho para o combate eficaz e eficiente da corrupção. Neste contexto, as redes sociais são uma fonte valiosa para analisar a percepção de um grupo de pessoas. Utilizando *tweets* do primeiro turno eleitoral do ano de 2018, colhidos por meio da API da plataforma *Twitter*, esse trabalho apresenta uma análise dos comentários publicados relacionados a corrupção. Para análise dessas falas, foi utilizada a ferramenta IRaMuTeQ, obtendo-se assim análises estatísticas, análises de similitude, classificação hierárquica descendente, nuvem de palavras, entre outros resultados. Ao fim das análises foi possível observar diferentes percepções sobre à corrupção no conjunto de dados referentes a 3395 *tweets*. Como conclusão, notou-se duas grandes percepções: um grupo que apontava de maneira extremada e radical a corrupção do governo passado, e outro grupo com comentários anticorrupção mais moderados e que criticavam os dois lados do espectro político.

Palavras-chave: Twitter. Corrupção. Iramuteq.

Abstract

Understanding corruption and its nuances, and how they affect the citizen, can be a good way to an effective and efficient combat of corruption. In this context, social networks are a valuable source for analyzing the perception of a group of peoople. Using tweets from the first electoral turn of the year 2018, collected through Twitter's API platform, this study presents an analysis of the published comments related to corruption. For the analysis of these speeches, the IRaMuTeQ tool was used, obtaining statistical analyzes, similarity analyzes, descending hierarchical classification, word cloud among other results. At the end of the analyzes it was possible to observe different perceptions about corruption in the dataset referring to the 3395 *tweets*. As conclusion, two great insights were noted: a group that pointed out the corruption of the past government in an extreme and radical way. and another group with more moderate anti-corruption comments that criticized both side of the political spectrum.

Key-words: Twitter. Corruption. Iramuteq.

Lista de ilustrações

Lista de tabelas

Lista de quadros

Sumário

1 Introdução

1.1 Contextualização

A corrupção, outrora entendida principalmente como uma falha de um indivíduo, evolui para um problema maior, um problema sistêmico, onde deixa sequelas na própria sociedade (VAZ; VELASCO, 2018).

A corrupção existe em várias formas, modelos e tamanhos, o que pode tornar sua definição complexa. A *Transparency International* traz a corrupção como "o abuso do poder confiado para ganho privado". Esta organização traz com essa definição três categorias de corrupção: grande, pequena e política. As grandes são cometidas pelo alto escalão do governo, que usa os seus poderes para se beneficiar em custa da sociedade. As pequenas geralmente são cometidas por funcionários públicos de baixo ou médio escalão quando interagem com o cidadão comum que busca acesso a algum bem público. Por fim, a corrupção política é onde a máquina pública é usada para manutenção do poder e riqueza de quem a comete.

Tradando do Brasil, a corrupção não é novidade. Ainda são frescos na memória do brasileiro exemplos como: a CPI da corrupção, o "esquema PC Farias", o Mensalão e a operação Lava-jato. Eugênio Aragão, ex-ministro da justiça, chega a dizer que a corrupção pequena "[...] na verdade, serve como uma graxa na engrenagem da máquina [...]". Essa banalização traz consigo um significado mais profundo, que diz respeito a uma corrupção que é intrínseca ao Brasil, e que, na verdade, sem ela o país não funcionaria direito.

Dentro do debate e discussão sobre política, as redes sociais surgem como verdadeiros espaços para essa conversa no século XXI. Pesquisas do *Twitter* contam que cerca de 70% de seus usuários usam a plataforma para se inteirar sobre política (TWITTER, 2018b). Também foi estudado que, durante as eleições presidenciais de 2018, a corrupção foi o tema mais discutido (TWITTER, 2018a). Dentro desse cenário, emerge o questionamento sobre as percepções do cidadão em relação à corrupção. O que é dito dentro de um contexto de pouca moderação em que seus usuários têm bastante liberdade para se expressarem de forma livre?

1.2 Justificativa

Analisar os dados sobre a opinião de uma população em relação à corrupção traz à tona os possíveis sentimentos e percepções que os cidadãos tem em relação ao seu país. Expõem a transparência que o governo tem em relação ao seu povo. Revela os níveis de

liberdade da imprensa na região estudada.

Do lado do governo, este estudo revela possíveis pontos de atenção e melhoria e utiliza o que a opinião pública tem a agregar. Do lado do cidadão, surge a possibilidade de encontrar-se em meio a tantas outras falas e entender onde sua percepção se encaixa no meio de tantas outras.

1.3 Objeto de Estudo

Comentários sobre corrupção, postadas durante o processo eleitoral brasileiro de 2018 na plataforma de *microblogging Twitter*.

1.4 Questão de Pesquisa

Quais as percepções sobre à corrupção, pela população brasileira durante as eleições presidenciais no Brasil, publicadas no *Twitter*?

1.5 Objetivos

1.5.1 Objetivo Geral

Analisar os *tweets* relacionados a corrupção.

1.5.2 Objetivos Específicos

Como objetivos específicos, foram definidos:

- Desenvolver um programa para coleta de falas sobre corrupção no *Twitter*;
- Tratar os dados para que possam ser utilizados;
- Realizar a análise dos dados usando a ferramenta Iramuteq.

2 Referencial Teórico

2.1 Percepção e Corrupção

Antes de se chegar na percepção, primeiro vem a sensação, que nada mais é do que um estímulo a algum sentido. Com a sensação captada, o cérebro se encarrega da percepção, passando pelas etapas de organização, identificação e interpretação da sensação (SCHACTER; GILBERT; WEGNER, 2009).

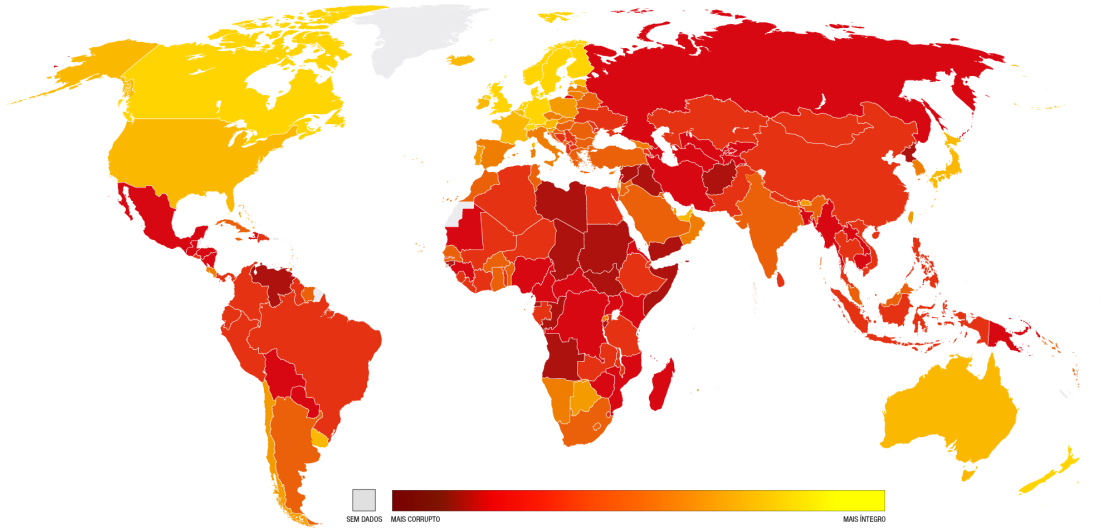
Por exemplo, quando se escuta uma música, de uma maneira bem simplificada, a sensação registrada é a de vibração no tímpano que após passar por processos biológicos, transforma-se em um impulso nervoso. Assim que o impulso chega ao cérebro, o sinal é organizado, identificado e interpretado e assim é possível perceber a música.

Com esse ponto de partida biológico, vem o paralelo social. Dadas as sensações de notícias, escândalos, abusos do poder confiado, quais são as percepções sobre a corrupção?

A *Transparency International* trabalha com uma classificação sobre a percepção da corrupção ao redor do mundo. A figura 1 mostra de forma gráfica os níveis de percepção da corrupção nos 180 países e territórios estudos pela organização.

Esse estudo é realizado anualmente pela *Transparency* desde 2012, ano que o Brasil tinha a nota 43 em uma escala de 0 a 100, sendo 0 muito corrupto e 100 muito íntegro. No relatório de 2019 foi registrada a pior nota do país desde 2012, com a pontuação passando a ser 35. Essa queda foi registrada de maneira consecutiva desde o ano de 2016. Observando

Figura 1 – Índice de Percepção da Corrupção



Fonte: adaptada de Transparency International (2019)

esse declínio ano após ano, é notável que mesmo com todo o esforço contra a corrupção ainda não se chegou à raiz do problema (TRANSPARENCY INTERNATIONAL, 2019).

Aprofundando-se na definição e nas categorias de corrupções entregadas pela *Transparency International*, temos:

- *Grandes Corrupções:* quando alguém do alto escalão do governo abusa do poder para beneficiar poucos à custa de muitos. Por exemplo, um presidente de um país roubar os cofres públicos e depositar o dinheiro em uma conta privada. São corrupções que raramente são punidas.
- *Pequenas Corrupções:* são as corrupções que acontecem quando um funcionário do governo de escalão mais baixo comete ao interagir com o cidadão comum. Conhecida também como a corrupção do dia a dia. Um exemplo seria o pagamento de propina para um fiscal acelerar o andamento de um alvará.
- *Corrupções Políticas:* embora as grandes e as pequenas corrupções também tenham base na esfera política, a corrupção política é aquela que usa o próprio sistema para o abuso do poder. Usar a máquina pública para manter o poder, o status e a riqueza. Exemplo desse tipo de corrupção seria o mau uso do poder confiado para repressão política de oponentes do governo.

2.2 Eleições Brasileiras

A lei nº 9.504/97 (BRASIL, 2018) rege as eleições brasileiras, sendo ela responsável por definir aspectos como data da eleição, quais cargos que concorreram simultaneamente, limite de doações, resolução para empates e regras para um segundo turno.

É estabelecido que as eleições brasileiras acontecem no primeiro domingo de outubro do ano eleitoral, e em caso de segundo turno, ele acontecerá no último domingo de outubro (BRASIL, 2018). As eleições acontecem nessa data, mas o calendário do processo eleitoral brasileiro começa antes desse período. No ano de 2018, por exemplo, ela teve início no dia 28 de novembro de 2017, com os testes públicos de segurança do sistema eletrônico de votação, e teve término no dia 31 de dezembro de 2018, data final para o ministério público aplicar penalidades e sanções relativas às doações irregulares de campanha (TRIBUNAL SUPERIOR ELEITORAL, 2017). Os cargos de 2018 ano foram de deputado federal, deputado estadual, senador, governador e presidente.

2.2.1 Principais Partidos e seus Candidatos no Ano Eleitoral de 2018

2.2.1.1 Partido dos Trabalhadores (PT)

Inicialmente o PT lançou como candidato à presidência o ex-presidente da república Luiz Inácio Lula da Silva tendo como vice o ex-prefeito da cidade de São Paulo, Fernando Haddad. A coligação do ex-presidente foi alvo de diversos pedidos para o barramento de sua candidatura assim que foi registrada no Tribunal Superior Eleitoral (TSE), sendo a sua maioria por conta da Lei da Ficha Limpa (EXAME, 2018).

Com o indeferimento do registro de Luiz Inácio por seis votos a um, o Partido dos Trabalhadores teve o prazo de 10 dias para anunciar um novo candidato à presidência (EXAME, 2018). Foi indicado, então, o ex-prefeito Haddad para a presidência, e a jornalista Manuela D'ávila, filiada ao Partido Comunista do Brasil (PCdoB), como vice.

2.2.1.2 Partido Social Liberal (PSL)

O PSL lançou como candidato a presidente o até então deputado federal Jair Messias Bolsonaro, e como vice, o militar Antônio Hamilton Martins Mourão. Bolsonaro é capitão da reserva e foi eleito como deputado federal por sete mandatos pelo Rio de Janeiro.

O PSL é considerado um partido pequeno, e para as eleições de 2018, formou apenas uma aliança, o que fez o candidato Bolsonaro ter um tempo curto de televisão (NEXO JORNAL, 2018). Isso levou o candidato a investir pesado nas campanhas pelas redes sociais, meio pelo qual focou durante todo trajeto eleitoral (NEXO JORNAL, 2018).

2.2.1.3 Partido Democrático Trabalhista (PDT)

Como candidato à presidência pelo PDT foi lançado o ex-prefeito, ministro e governador de estado, Ciro Gomes. Ciro, como um candidato de centro-esquerda, conseguiu alcançar o terceiro lugar no primeiro turno eleitoral, com pouco mais de 13 milhões de votos que representam aproximadamente 12,47% dos votos válidos (ESTADÃO, 2018).

2.2.1.4 AVANTE

Cabo Daciolo foi o representante do AVANTE nas eleições de 2018. Daciolo ganhou notoriedade nas redes sociais após sugerir que Ciro Gomes era um dos mandantes do URSAL, União das Repúblicas Socialistas Latino-Americanas, a partir disso vários memes foram criados na rede em relação ao assunto (G1, 2018).

O Cabo declarou como gasto de campanha R$808, e com esse dinheiro chegou a sexto lugar no primeiro turno eleitoral. Henrique Meirelles, candidato do MDB, teve

a campanha mais cara de 2018 com 53 milhões de reais e ficou atrás do candidato do AVANTE (G1, 2018).

2.2.1.5 Partido da Social Democracia Brasileira (PSDB)

O ex-governador do estado de São Paulo, Geraldo Alckmin, se candidatou para presidência pelo PSDB. O PSDB é um dos maiores partidos brasileiros, e até então era o partido que disputava o poder contra o PT. Em 2018 o resultado nas urnas não foi como dos antigos anos eleitorais, ficando em quarto lugar na corrida presidencial (ESTADÃO, 2018).

2.2.2 Resultado Eleitoral

Caso um candidato não consiga maioria absoluta no primeiro turno eleitoral, é convocado um segundo turno onde somente os dois candidatos mais votados participam, tentando-se assim atingir uma maioria de votos válidos.

No primeiro turno das eleições de 2018, Jair Bolsonaro conquistou 46,03% dos votos válidos, enquanto o segundo lugar, Fernando Haddad, conseguiu 29,28% do total, e em terceiro lugar ficou o candidato do PDT, Ciro Gomes, com 12,47% dos votos, convocando-se assim um segundo turno entre os dois primeiros candidatos (GAZETA DO POVO, 2018).

Dentro da apuração do segundo turno foram contabilizados 55,13% dos votos válidos para o até então presidenciável do PSL e 44,87% para o do PT, tornando assim Jair Bolsonaro presidente do Brasil com cerca de 58 milhões de votos (GAZETA DO POVO, 2018).

2.3 *Twitter*

O *Twitter* é uma rede social que funciona como uma plataforma *online* para *microblogging*. Cada usuário tem sua própria linha do tempo e cada postagem nessa linha é conhecida como um *tweet*, que são textos limitados a até 240 caracteres, fora algumas línguas como a japonesa, chinesa e a coreana que tem o limite de 140 caracteres (TWITTER, 2017). Cada *tweet* pode ser compartilhado por outros usuários em suas próprias linhas do tempo, esse processo é chamado de *retweet*.

> O Twitter é como o seu bar favorito funcionando dia e noite: a hora que você aparecer encontrará alguns frequentadores habituais e mais outras pessoas relacionadas a eles. Você poderá ficar para um dedo de prosa durante um intervalo no trabalho ou passar horas interagindo e trocando idéias[sic] (SPYER et al., 2009, p.8).

A empresa foi fundada em 2006 por Jack Dorsey, Noah Glass, Christopher Isaac e Evan Williams. Ela teve um crescimento acelerado ao longo dos seus primeiros anos, sendo que em 2012 o *Twitter* já tinha mais de 140 milhões de usuários ativos, que juntos somavam mais de 340 milhões de *tweets* por dia (TWITTER, 2012). Em 2018, o número de usuários ativos ao redor do mundo chegou a 335 milhões (VARIETY, 2018).

Com 27,7 milhões de usuários em 2016, o Brasil já era a segunda maior população de pessoas no *Twitter*, ficando atrás apenas dos Estados Unidos (EMARKETER, 2016).

2.3.1 O *Twitter* e a Política

O *Twitter* tem ganhado força no debate político. SPYER et al. (2009) citam cinco maneiras de como o poder dos ativistas políticos são ampliados:

- Aumenta a distribuição rápida de informação;
- Cria consciência de grupo;
- Ajuda na coordenação de protestos;
- Aumenta a segurança;
- Fortalece vínculos com a sociedade civil.

Falando especificamente de política, em média 70% dos brasileiros utilizam a rede social para se informar sobre política (TWITTER, 2018b). Em pesquisa, 79% dos votantes indecisos falaram que as propostas postadas nas contas dos candidatos poderiam influenciar sua decisão de voto (TWITTER, 2018b).

Nas eleições de 2018 os presidenciáveis Jair Bolsonaro e Fernando Haddad foram os que tiveram mais comentários na rede, justamente os candidatos que chegaram ao segundo turno. Os mesmos foram seguidos por Ciro Gomes e Geraldo Alckmin (TWITTER, 2018a). Como Haddad só se tornou candidato oficial do PT no dia 11 de setembro de 2018, as pesquisas só começaram a contabilizar dados a seu respeito a partir dessa data (TWITTER, 2018a).

2.4 Trabalhos Relacionados

Para poder encontrar artigos científicos com relevância para o estudo, foi elaborada uma revisão sistemática de literatura descrita no apêndice A. Com a revisão, foi possível encontrar 14 artigos que compreendiam *Twitter* e política ou corrupção.

Com todos os artigos lidos, foi feito um estudo para saber quais são as principais categorias de análises feitas com dados do *Twitter*, dentro da comunidade cientifica. Esse

estudo foi sumarizado no Quadro 1. Para a classificação dos artigos foram utilizados quatro tipos de métodos analíticos, definidos por Grover et al. (2017), sendo eles:

Quadro 1 – Metodologias Aplicadas em Trabalhos Relacionados (fonte: Autor)

Artigo	**Metodologias Utilizadas**
Guimaraes, Wang e Weikum (2017)	Modelagem de Tópicos usando *Latent Dirichlet Allocation*
Calderon et al. (2015)	Análise de Sentimentos usando Dicionário Léxico
Ahmed e Skoric (2014)	Análise Descritiva e Espaço-temporal
Vallina-Rodriguez et al. (2012)	Análise Descritiva e Espaço-temporal
Nugroho, Doewes et al. (2017)	Análise de Sentimentos usando *Multinomial Naive Bayes* e *Recursive Feature Elimination*
Fink et al. (2013)	Análise de Sentimentos usando *Support Vector Machine - Linear Kernel*
Carvalho et al. (2016)	Modelagem de Tópicos usando *Latent Dirichlet Allocation* e *Latent Semantic Indexing*
Rodrigues, Rao e Chiplunkar (2017)	Análise de Sentimentos usando Dicionário Léxico
Paiva, Garcia e Alcântara (2017)	Análise Descritiva
Nechai e Goncharov (2017)	Modelagem de Tópicos usando *Latent Dirichlet Allocation*
França, Goya e Penteado (2018)	Análise de Sentimentos usando *Extreme Gradient Boosting Tree*
Grover et al. (2017)	Modelagem de Tópicos e Análise de Sentimentos
Teixeira et al. (2018)	Análise Descritiva
Takikawa e Nagayoshi (2017)	Modelagem de Tópicos usando *Latent Dirichlet Allocation* e Extração de Comunidades usando o Método de Louvain

- Análise descritiva;
- Análise de conteúdo;
- Análise de rede;
- Análise espaço-temporal.

Análise descritiva é o tipo que se concentra nas estatísticas, como número de *tweets*, quantidade de usuários, dados sobre *hashtags* e palavras mais usadas. Mesmo que não fosse a abordagem principal, a maior parte dos artigos estudados na revisão de literatura usaram esse tipo de análise para poder descrever seus dados.

Análise de conteúdo é onde consegue-se retirar informação do texto. Métodos que envolvem Processamento de Linguagem Natural (PLN) são utilizados. Para esse tipo de

análise, foi aprofundado um nível no Quadro 1. No lugar do método, foi descrito o que foi feito com o dado e qual algoritmo foi utilizado para extração da informação. As três principais abordagens utilizadas dentro dessa análise foram: modelagem de tópicos, análise de sentimentos e a extração de comunidades. Para a modelagem de tópicos, foram utilizados os algoritmos *Latent Dirichlet Allocation* (LDA) e *Latent Semantic Indexing* (LDI). Para análise de sentimentos as principais abordagens foram, dicionário léxico, *Multinomial Naive Bayes* (MNB), *Support Vector Machine - Recursive Feature Elimination* (SVM - RFE), *Support Vector Machine - Linear Kernel* (SVM - LK) e *Extreme Gradient Boosting Tree* (EGBT). Para extração de comunidades foi utilizado o método de Louvain.

Análise de rede é onde se estuda o tamanho de redes de relacionamento, a detecção de grupos e o fluxo da informação, por exemplo. Nechai e Goncharov (2017) usaram esse tipo de abordagem para mostrar graficamente como grupos contra e pró-governo se interagem.

Análise espaço-temporal compreende o estudo da informação durante um certo período. Análise de tendências durante um período, análise geoespacial e evolução de tópicos são exemplos desse tipo de abordagem. A espaço-temporal foi outro tipo de análise bastante utilizada por vários dos artigos estudados, mesmo não sendo o principal objeto de análise desses artigos.

Em vários casos, os métodos se misturam durante a análise, por isso, para simplificar o Quadro 1, foi descrito somente os principais métodos utilizados.

2.4.1 Dados

2.4.1.1 Organização

Organizar e selecionar são as etapas iniciais da análise de conteúdo, também chamada de pré-análise (TEIXEIRA et al., 2018). Para Teixeira et al. (2018), um bom começo para encontrar o tema a ser tratado é descobrir quais assuntos estão sendo discutidos na mídia nacional. Para eles, tópicos como política e corrupção podem gerar desentendimento entre as pessoas, tornando-se assim tópicos para discussões polarizadas.

2.4.1.2 Coleta

O *Twitter* fornece uma Interface de programação de aplicações (API) para filtragem de *tweets* em tempo real. Alguns softwares como o *STACKS*, são *open-source* e ajudam na coleta de dados (CALDERON et al., 2015). Esses softwares encapsulam a API em um programa que simplifica o trabalho de quem precisa fazer a coleta.

Teixeira et al. (2018) comentam que quando se está fazendo a coleta de dados sobre um tópico específico, essa coleta deve ser mantida enquanto a relevância do tópico ainda estiver em alta. Eles utilizaram duas técnicas para medir essa relevância: primeiramente

observar se o volume de dados coletados está diminuindo e em seguida verificar se o interesse de buscas pelo tópico no *Google Trends* estava em declínio.

2.4.1.3 Seleção

Sabendo da quantidade massiva de *tweets* gerados, Teixeira et al. (2018) escolheram o método dos *retweets* como forma de seleção. Para os autores, o *retweet* é uma boa métrica, pois, quem o fez, considerou o conteúdo de importância o suficiente para colocá-lo em sua própria *timeline*.

2.4.1.4 Análise de Dados

Takikawa e Nagayoshi (2017) usaram o método de Louvain para fazer a identificação de comunidades dentro do seu conjunto de dados. Louvain é um método que, por heurísticas, consegue extrair rapidamente comunidades de grandes redes (TAKIKAWA; NAGAYOSHI, 2017).

A descoberta de tópicos dentro das bases de dados apresenta-se como um assunto de interesse para vários dos pesquisadores (AHMED; SKORIC, 2014). Para fazer a extração desses tópicos, alguns autores usaram o modelo *Latent Dirichlet Allocation* (TAKIKAWA; NAGAYOSHI, 2017; GUIMARAES; WANG; WEIKUM, 2017). Uma outra equipe (CALDERON et al., 2015) utilizou a ferramenta *IN-SPIRE*^{TM} *Visual Document Analysis*, que junta documentos que tenham tópicos parecidos e determina quais são os principais tópicos ou temas de cada coleção. Nugroho, Doewes et al. (2017) usaram *Multinomial Naive Bayes* (MNB) e *Support Vector Machine - Recursive Feature Elimination* (SVM-RFE) para realizar a classificação do texto. Nugroho, Doewes et al. (2017) fizeram um *dataset* com uma classificação manual de 96 mil *tweets* em positivo, negativo e neutro.

Antes de começar o processo de extração, é realizado um pré-processamento, onde são removidas as *stop words*[1].

A análise de sentimentos foi outra técnica utilizada (CALDERON et al., 2015). A análise de sentimentos em português pode ser um desafio, principalmente com as ferramentas *open-source* que entregam bons resultados para textos na língua inglesa (CALDERON et al., 2015). Por essa razão, uma ferramenta utilizada foi o *SentiStrength*[2], que é gratuito para uso não comercial e tem a flexibilidade para trabalhar com dicionários que não estejam na língua inglesa.

Vallina-Rodriguez et al. (2012) analisaram as *hashtags* mais populares relacionadas às eleições, classificaram de acordo com seu espectro político e por fim fizeram a contagem

[1] As palavras que mais aparecem em um texto, normalmente serão palavras comuns como 'o', 'ou' e 'e'. Como elas sozinhas não trazem muito significado, normalmente são removidas antes de o processamento de dados ser feito (LESKOVEC; RAJARAMAN; ULLMAN, 2014)

[2] http://sentistrength.wlv.ac.uk/

de *tweets* para cada *hashtag*. Partindo disso, várias análises foram feitas como histórico de menção de cada *hashtag*. Contagem do espectro mais mencionado, mapa com os lugares que têm mais *hashtags*.

Fink et al. (2013) usaram o *Lucene text indexing API* para selecionar somente os *tweets* que tinha haver com o assunto que estavam querendo. Para fazer a classificação de positivo, negativo e neutro, pagaram o serviço *Amazon Mechanical Turk*.

3 Metodologia

3.1 Universo da Pesquisa

Compreende-se como universo da pesquisa os *tweets* em português, feitos na rede social *Twitter*, postados no primeiro turno eleitoral, que aconteceu no dia 7 de outubro de 2018, e que apresentaram alguma das seguintes palavras:

- Corrupção;
- Corrupto;
- Corrupta.

3.2 Materiais e Equipamentos

Para a coleta e armazenagem dos dados foi utilizado um computador pessoal com as seguintes configurações:

- Intel Core i7 3610QM @ 2.30GHz Ivy Bridge;
- 8,00GB DDR3 de memória RAM;
- Linux Mint 19.1 Tessa Cinnamon.

Com essas configurações, foi possível realizar todos os procedimentos computacionais necessários do início ao fim do trabalho.

Para realizar as principais análises de dados, foi utilizado o *software* de análises textuais IRaMuTeQ - *Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires* (Interface R para Análises Multidimensionais de Textos e Questionários), na versão 0.7 *alpha* 2.

3.3 Procedimento de Coleta de Dados

A coleta de dados foi feita por meio da API "Busca de Tweets: Arquivo Completo" da *Twitter Developer Platform*[1]. O modo *sandbox* dessa API permite com que se faça 50 requisições e para cada requisição pode ser retornado até 100 *tweets*. O número de requisições feitas é zerado a cada mês.

[1] https://developer.twitter.com/

Para que seja encontrado somente o conteúdo que o usuário está procurando, a API permite o uso de alguns filtros como: palavras-chave, período, *hashtags* e até mesmo parâmetros geoespaciais. Para o modo de *sandbox* existe o limite de 256 caracteres para o tamanho do filtro.

Nesse estudo, foi escolhida a abordagem de busca por palavras-chave, utilizando as palavras definidas no universo da pesquisa. Outra abordagem possível seria colher os *tweets* de pessoas influentes do processo eleitoral, que nesse caso poderiam ser as contas do *Twitter* dos candidatos, dos partidos ou da grande mídia. Essa abordagem não foi seguida, pois o estudo não visa analisar os dados específicos de contas pontuais.

A API não faz distinção de letras maiúsculas ou minúsculas nas frases, mas faz a correspondência exata de cada caractere contido no UTF-8, sendo assim, faz distinção de letras acentuadas. Por exemplo, procurar pela palavra "Corrupção", trará resultados para "corrupção", mas não trará para "Corrupçao". Por isso a lista do universo de pesquisa foi expandida para todas as variações de acentuação, que poderiam ocorrer na digitação, quando o usuário fosse publicar um *tweet*.

Para utilizar a interface do *Twitter*, foi elaborado um *software*[2] na linguagem de programação *JavaScript*. Para auxiliar a coleta, foi utilizada a biblioteca *npm twitter*[3], que encapsula a API em funções que já seguem as regras de negócio necessárias da plataforma.

Para realizar o armazenamento dos dados, foram estudadas algumas opções como Postgres, MySQL, MongoDB e RethinkDB. A opção escolhida foi a do MongoDB, por ser um banco de dados *open-source*, e principalmente por trabalhar nativamente com o formato de dados JSON. A API do *Twitter* retorna seus dados e metadados no formato JSON, o que facilita a inserção dos dados

Com tudo organizado, a coleta e armazenagem foram feitas duas vezes, devido ao limite de requisições da *sandbox*. Cada uma das coletas retornou cerca de 5000 *tweets*. De modo a diminuir a repetição de textos, foram eliminados os *retweets* do montante total, chegando a quantidade final de 3567 textos.

3.4 Análise e Tratamento de Dados

As análises textuais foram feitas por meio do IRaMuTeQ. Esse *software* é uma ferramenta utilizada para análise de texto, feita nas linguagens de programação R e Python. Dentre as formas de análise, o *software* conta com estatísticas textuais, análise de especificidades, classificação hierárquica descendente (CHD), análise de similitude e nuvem de palavras. Desses métodos não será usada a análise de especificidades, pela complexidade

[2] https://github.com/marceloabk/TCC1
[3] https://github.com/desmondmorris/node-twitter

de extração de variáveis sociodemográficas dos *tweets*, e a nuvem de palavras. Está última será implementada fora do IRaMuTeQ para que a customização da mesma seja maior.

Antes de realizar a CHD, por padrão, é feita uma seleção das formas plenas: verbos, substantivos, advérbios e adjetivos (RATINAUD; MARCHAND, 2012). Essas formas fazem oposição as suplementares: preposições, pronomes, advérbios de frequência, alguns tipos de verbos, entre outras classes gramaticais. Feita a seleção, as formas de cada texto passam pelo processo de lematização, sendo esse um processo que transforma as flexões de uma palavra em sua forma base (CAMARGO; JUSTO, 2013), por exemplo: as palavras "corrupção" e "corrupções" transformam-se em "corrupção".

Aplicadas as técnicas iniciais, começa o processo da CHD. O processo começa com a construção de uma matriz binária que cruza as unidades de texto com as formas plenas lematizadas (RATINAUD; MARCHAND, 2012; CHARTIER; MEUNIER, 2011). O método segue com uma série de bipartições que são feitas nessa matriz utilizando a análise fatorial de correspondência (AFC). Para realizar cada uma das bipartições são feitos três processos:

- O processo é iniciando aplicando a AFC na matriz binária. Para todas partições possíveis são calculados os valores de inércia[4] até ser achado a partição onde esse valor é maximizado.
- Em seguida é feita uma permutação de cada unidade da tabela e o valor de inércia entre as classes é recalculado. Caso o valor tenha aumentado a permutação é mantida. O ciclo de permutações continua até que o valor de inércia não cresça.
- As formas que são específicas de uma classe, confirmadas pelo teste de chi-quadrado, são retiradas das outras.

A análise de similitude é um tipo de análise feita a partir da teoria dos grafos que também usa como entrada a matriz binária de formas e unidades de texto (SALVIATI, 2017). A análise de coocorrência é realizada na matriz entregando a frequência de ocorrência entre dois termos. Com o resultado da coocorrência é construído um grafo com todas as ligações possíveis para cada um dos termos (MARCHAND; RATINAUD, 2012). O último processamento realizado é feito nos pontos do grafo que formam ciclos, para cada ciclo é realizada uma busca pela aresta com menor peso e quando encontrada a mesma é removida (MARCHAND; RATINAUD, 2012).

Para que os textos pudessem ser utilizados no *software* IRaMuTeQ, algumas alterações que eram demandadas foram realizadas e outras foram feitas para que os resultados

[4] Para a análise de correspondência entende-se como inércia o percentual de variância e corresponde a soma ponderada das distâncias de pontos de um conjunto em relação ao seu centroide. Seu cálculo é dado pelo qui-quadrado do conjunto dividido pelo número de formas (GONÇALVES; SANTOS, 2009).

obtidos fossem melhorados. Como essas alterações eram de cunho textual, foi utilizada a técnica de expressão regular (em inglês *regex*) que é uma forma para encontrar padrões dentro de um texto. Essa técnica foi utilizada para:

- **Remover sinais proibidos:** existe uma lista de sinais proibidos e uma de sinais permitidos no IRaMuTeQ, sendo os proibidos: aspas, apóstrofo, cifrão, porcentagem, asterisco, reticências, travessão, entre outros. E os permitidos: dois-pontos, vírgula, ponto de exclamação, ponto de interrogação e travessão. Todas as ocorrências nos textos que estivesse na lista de sinais permitidos foram mantidas enquanto o resto foi descartado.
- **Remover formas de baixo valor para análise textual:** menções a usuários feitas pelo sinal de "@", onomatopeias para risadas como "kkkk" e "hahaha", e URLs foram removidas pois apresentam pouca ou nenhuma informação para análise textual.
- **Transformação de sinais proibidos:** *hashtags* e hifens fazem parte da lista de sinais proibidos, então, para que se realize a análise de palavras que são separadas por hífen como também a análise de *hashtags* esses sinais foram substituídos por sinais de *underscore*. Por exemplo: "vice-presidente" transforma-se em "vice_presidente" e "#Eleições2018" transforma-se em "_Eleições2018".
- **Junção de palavras que apresentam sentido único:** algumas expressões como lava jato e *fake news* apresentam um sentido único, então essas palavras foram conectadas por meio do *underscore* e tornaram-se "lava_jato" e "*fake_news*".
- **Correção ortográfica:** Cada vez que uma CHD era feita no IRaMuTeQ gerava-se uma lista com palavras que não eram reconhecidas pelo *software*. Essa lista era composta por nomes próprios, erros de português e abreviações de palavras. Eram geradas *regexes* para correção dos erros de português observados e também para trazer à forma normal as abreviações encontradas. Os textos corrigidos eram utilizados para uma nova CHD, e esse processo foi realizado repetidamente até ser gerada uma lista das palavras não reconhecidas que não apresentasse erros ortográficos ou abreviações.

Para nuvem de palavras houve o mesmo processo de eliminação das URLs, das menções a usuários e correção ortográfica. Juntos a esses processos, foi incluído passos para remoção de pontuação e de *stopwords*. Terminada essa etapa, foi criado um contador de palavras e cada *tweet* recebia um peso que era a quantidade de vezes que ele foi *retweetado*. Logo, caso um *tweet* tivesse as palavras "haddad", "bolsonaro", "agora" e "haddad" novamente e esse mesmo *tweet* tivesse 15 *retweets*, o contador de palavras teria a seguinte contagem:

- haddad: 30;
- bolsonaro: 15;
- agora: 15;

4 Resultados

4.1 Classificação Hierárquica Descendente (CHD)

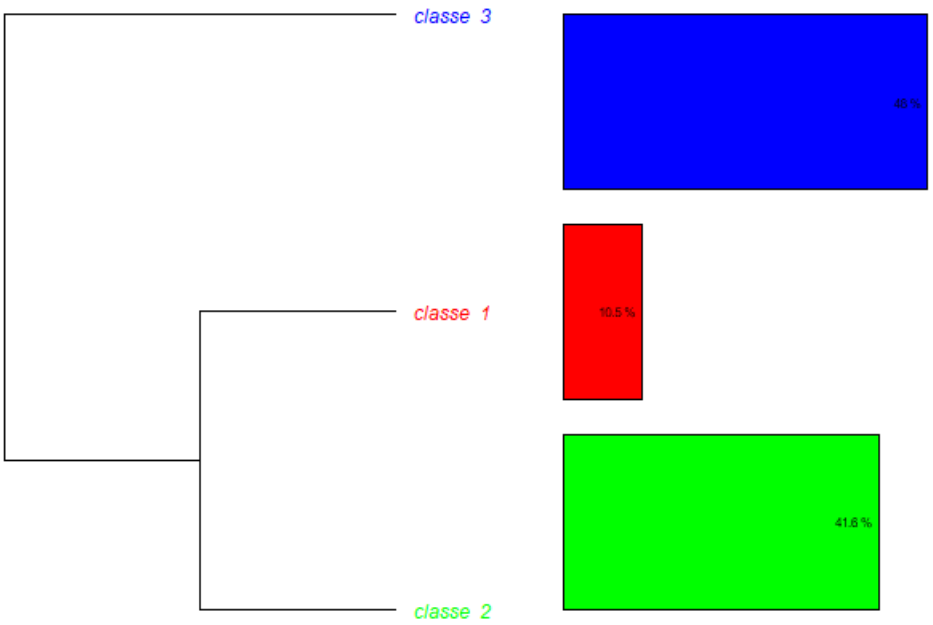
O método de CHD, desenvolvido por Reinert em 1990, visa analisar segmentos de texto com base na frequência de suas formas, utilizando o teste qui-quadrado para cruzar os textos e as palavras (JUSTO; CAMARGO, 2014). Dessa maneira, são formadas classes que têm vocabulários semelhantes, separadas de classes com vocabulários opostos.

Foram analisados 3567 *tweets* pelo CHD. Dentro desses *tweets*, foram identificadas 94079 ocorrências de palavras, formas ou vocabulários, sendo que esse conjunto apresenta 8528 palavras distintas e 2835 aparecem apenas uma vez no texto. Dos 3567 segmentos inicialmente computados, foram classificados com sucesso 3395 (95.18%). Os textos analisados foram divididos em três classes:

- Classe 1: 355/3395 (10.46%);
- Classe 2: 1412/3395 (41.59%)
- Classe 3: 1628/3395 (47.95%)

Vale ressaltar que na figura 2 existem dois subgrupos, um formado unicamente pela classe 3 e um segundo formado pelas classes 1 e 2. A classe 3 será chamada de Anticorrupção — comentários moderados, sendo uma classe que não é possível classificar em algum dos lados do espectro político, existem comentários que atacam e defendem ambos os lados. A classe 2 receberá o nome de Anticorrupção — comentários extremados, é uma classe que deixa clara a predileção pelo candidato de direita, com comentários extremados e discurso de ódio em relação ao candidato de esquerda e seus votantes. Por último, a classe 1 será a Anticorrupção — sistema, essa, no que lhe concerne, têm em suas falas comemorações por muitos candidatos ditos corruptos não serem reeleitos, com vários comentários alegando corrupção e fraude nas urnas eletrônicas. Tendo um nível compartilhado com a classe 2, é possível notar na classe 1 um viés político de direita.

Figura 2 – Dendrograma a partir da CHD.



Fonte: do Autor.

A figura 3 mostra as palavras que são comuns entre si dentro de cada classe, bem como a diferença de vocabulário entre as outras classes.

Por exemplo, a seguir algumas das frases que ficaram em destaque na classificação do IRaMuTeQ para cada uma das classes.

Classe 1: Anticorrupção - sistema:

> "Querido, isso é a fraude combinada com o o sistema. Gleise e Aécio, os dois mais odiados! só falta Renan Calheiros se reeleger. Aí não precisa polícia! é fraude garantida. Essas urnas são o caminho da corrupção garantida"

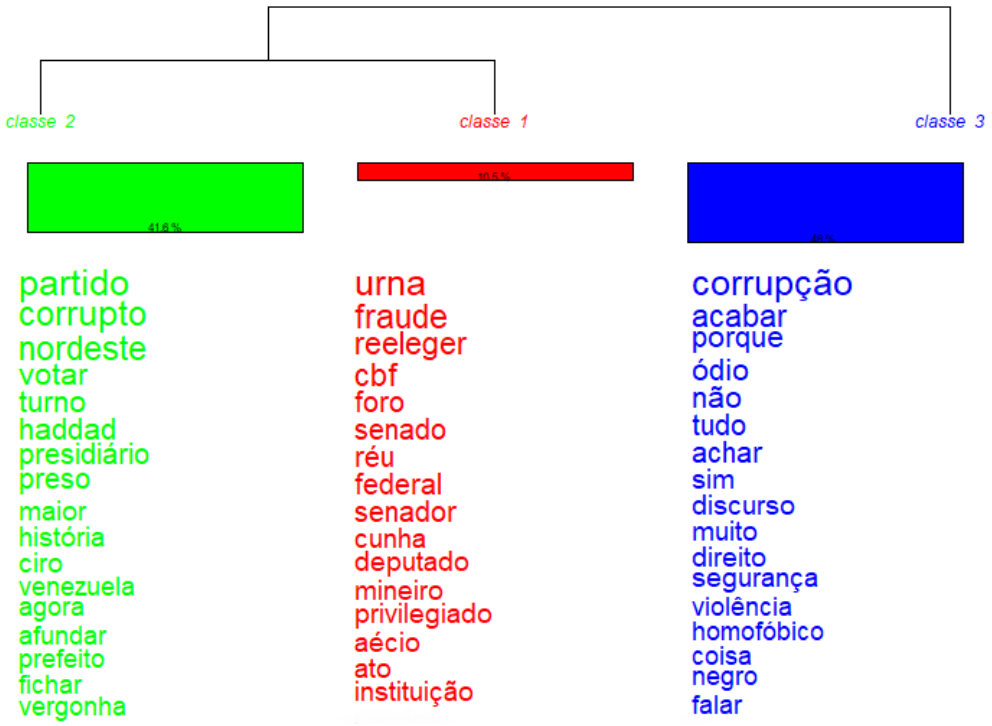
> "Varios réus e investigados por corrupção não foram reeleitos! Parabéns sociedade brasileira! Mandou o recado pelas urnas. Chega de corrupção!"

> "Por conta das urnas fraudadas teremos segundo turno para presidente da república. Nem as eleições escapa da corrupção. É lamentável. #bolsonaro17 #bolsonaropresidente #brasilcombolsonaro"

Classe 2: Anticorrupção - comentários extremados:

> "O nordeste é um câncer, vivem das migalhas do pt e mesmo assim não conseguem sair da miséria. Merecem se ferrarem, o partido mais corrupto

Figura 3 – Dendrograma com palavras-chave a partir da CHD.



Fonte: do Autor.

e nojento da história desse país disputando o maior cargo do executivo é inadmissível. Mas mesmo assim venceremos no segundo turno. #17neles"

"Nordeste votando no pt e ciro e afundando o brasil, parabéns por votar em partido de corrupto e presidiário, por isso não vai pra feente"

"Sua débil mental dos infernos ! Haddad foi o pior prefeito da história de são paulo ! E o pt é o partido mais corrupto do brasil envolvido no maior escândalo de corrupção do mundo ! Aff, é mta burrice mesmo pqp !!!!!"

Classe 3: Anticorrupção - comentários moderados:

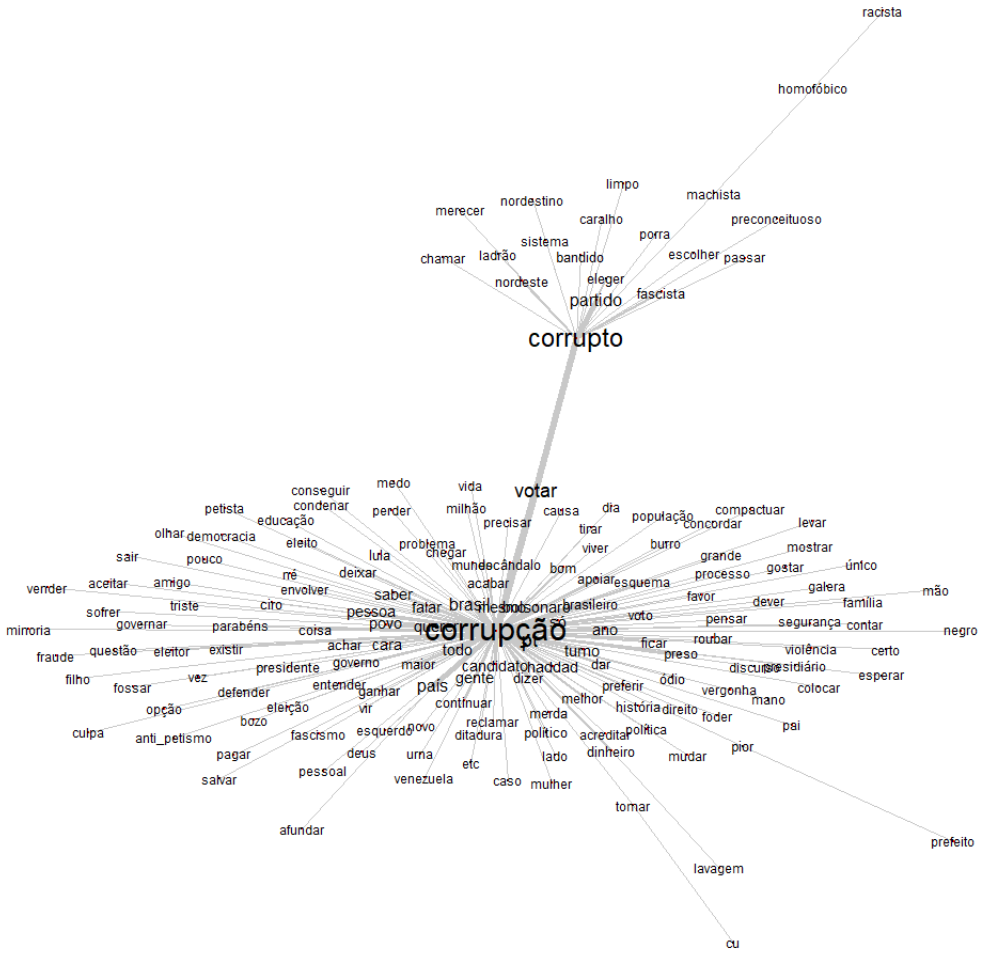
"Realmente mano tem coisa que ele falou que eu tambem não concordo, porem na minha opiniao ele é o unico honesto e que vai acabar com a corrupção do brasil e acho nada a ver a pessoa massacrar o cara sendo que o candidato dela pode ser muito pior mas só querer falar mal do outro"

"O pt não é culpado pela homofobia, racismo, etc, mas ele é responsável pelo crescimento do bolsonaro ele é sim, até porque tem dois tipos de eleitores do bolsonaro, o racista, homofóbico e fascista e o que não é assim, mas está preocupado com a corrupção"

"Numa situação de fascismo, se você não é contra, você é a favor. Corrupção tem investigação, tem punição... Está tendo mesmo não entendendo porque você coloca Haddad como corrupto! O outro foi racista, machista, homofóbico, defendeu tortura, fuzilamento de rivais, e deu nada!"

4.2 Análise de Similitude

Figura 4 – Análise de Similitude.



Fonte: do Autor.

A figura 4 mostra a análise de similitude do texto. A análise inicial contava com palavras que tinham uma frequência de no mínimo três aparições, utilizando assim 1743 palavras. Esse fato gerava um gráfico bastante poluído onde a visualização das informações era prejudicada. Dessa maneira, foi testado de forma experimental uma quantidade de palavras que não prejudicasse o entendimento do gráfico, chegando ao número de 134 formas.

4.3 Nuvem de Palavras - Palavras mais utilizadas

Figura 5 – Nuvem de Palavras - Palavras mais utilizadas[1]

Fonte: do Autor.

A figura 5 mostra a nuvem de palavras criada a partir das palavras mais utilizadas durante o período em que foi feita a coleta de dados.

4.4 Dados Estatísticos

Dentro do volume de dados foi possível extrair algumas métricas estatísticas sobre as informações dos *tweets*.

4.4.1 Usuários mais citados

A figura 6 mostra os usuários mais citados da base dados. É possível perceber um destaque de pessoas famosas, jornais e o próprio Jair Bolsonaro, presidenciável na época.

[1] Imagem interativa em: <https://wordart.com/g0y7ntm09j2d/twitter_wc>

Figura 6 – Usuários mais citados

Fonte: do Autor.

4.4.2 *Hashtags* mais compartilhadas

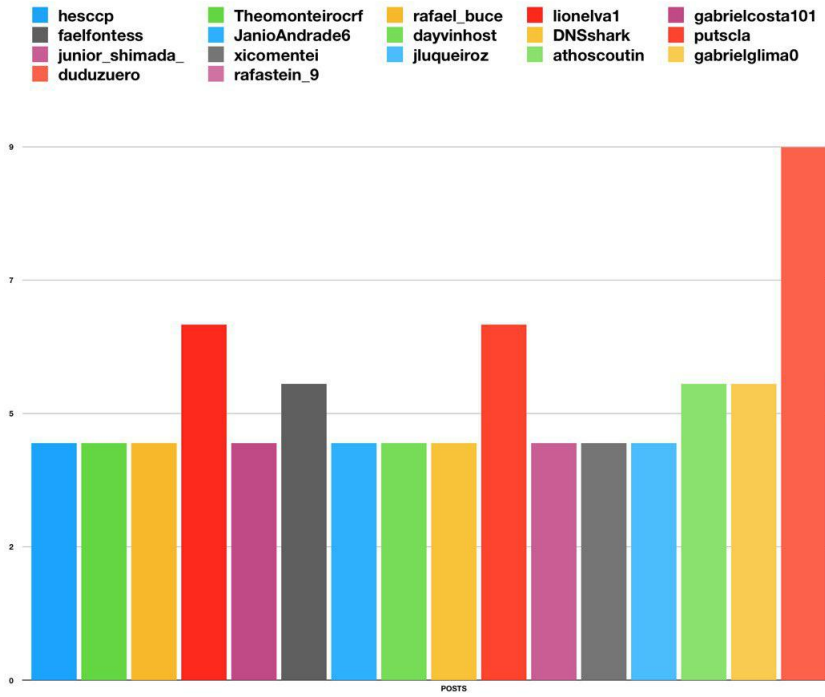
Figura 7 – *Hashtags* mais compartilhadas

Fonte: do Autor.

A figura 7 mostra uma nuvem de palavras com as *hashtags* mais compartilhadas no período. *Hashtags* idênticas, mas com diferenças de grafia foram consideras como uma só, por exemplo: #elenão, #EleNão e #Elenao foram contabilizadas como se fossem uma só *hashtag*.

4.4.3 Quantidade de *Posts* por Usuário

Figura 8 – *Posts* por usuário



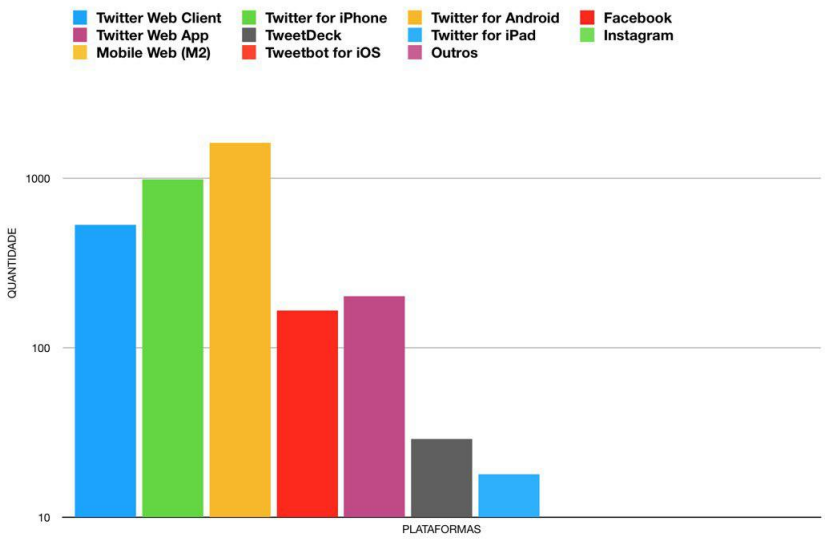
Fonte: do Autor.

A figura 8 mostra os usuários que mais postaram, considerando o período em que os dados foram coletados. Por conta da grande quantidade de usuários com poucas postagens, foram considerados usuários com pelo menos 4 *tweets*.

4.4.4 Plataformas mais utilizadas

As plataformas mais utilizadas, como mostra a figura 9, são respectivamente o *Twitter* para Android, o *Twitter* para iPhone e em terceiro lugar o cliente *web* do *Twitter*.

Figura 9 – Plataformas mais utilizadas



Fonte: do Autor.

5 Discussão dos resultados

Com a CHD foi possível destacar, entre todos os *tweets*, três categorias. Essas categorias foram chamadas de classes e cada uma delas foi nomeada de acordo com o conteúdo considerado destaque. A primeira classe foi chamada de Anticorrupção — sistema. Entre as frases mais representativas, foi possível perceber uma indignação pela possibilidade de fraudes nas urnas. Outras frases dessa mesma classe comemoram a não eleição de réus e investigados de corrupção.

Dentro da visão de corrupção no TSE, é possível pensar em pontos de melhoria. A opinião pública alega fraude e corrupção. Pensar em soluções para eliminar essas suspeitas ou eliminar a corrupção em si, caso exista, pode ser um foco do TSE. Soluções como voto impresso são as mais pedidas pelo cidadão, existe a necessidade de avaliação dessa medida, mas existem opções, como tornar o *software* das urnas aberto ou livre. Software livre pode ser uma boa solução porque quando algum problema é encontrado, rapidamente é solucionado pela quantidade de pessoas que estão os observando, além da redução de custos da produção.

A segunda classe recebeu o nome de Anticorrupção — comentários extremados. Essa classe considera o Partido dos Trabalhadores corrupto por definição. Assim repudiam o partido, e culpam o Nordeste por levar o PT para o segundo turno eleitoral. Em meio a esses comentários a maior parte é de discurso de ódio contra os votantes do PT e em outros é possível notar xenofobia.

Para terceira classe, Anticorrupção — comentários moderados, as percepções são mistas. Os comentários a favor do candidato Bolsonaro tratam sobre um fim da corrupção imposta pelo PT. Os comentários a favor do Haddad falam sobre o desprezo do outro candidato em relação às minorias, e falam sobre como escolher Bolsonaro não muda o fato da corrupção no governo, que para isso existem as investigações.

Como resultado do gráfico de similitude, é possível notar como é montada a conectividade entre as palavras. Como uma das palavras-chave para busca de *tweets* foi corrupção, é simples entender a lógica do motivo de ser uma das palavras centrais. Logo, sua análise fica por conta de perceber quais palavras apresentam um peso maior para aparecerem ligadas a mesma. Algumas ligações ressaltam aos olhos, como “machista-homofóbico-racista”, que foram adjetivações dadas ao candidato Bolsonaro por pessoas contrarias a sua candidatura. Outra cadeia de palavras notável é a “pior-prefeito-história” fazendo menção a reportagem sobre candidato Haddad que comentava sobre ele ter sido o pior prefeito da história de São Paulo.

Justo e Camargo (2014) comentam como a nuvem de palavras forma uma análise

mais simples sobre o conteúdo do texto em si. É uma análise da frequência de palavras, organizando com uma fonte maior, palavras com maior frequência. Também é dito que mesmo simples, é uma análise graficamente interessante. É possível notar quais as palavras são mais utilizadas no texto, quais os autores dão maior destaque.

As métricas estatísticas sobre os dados mostram informações interessantes como as *hashtags* mais utilizadas e como são bem divididas no espectro político. Inclusive uma das *hashtags* mais utilizadas, a #FicaTemer, apresenta um teor cômico por referenciar-se ao presidente com níveis de reprovação recordes, o que mostra também uma insatisfação com as opções para representante do mais alto cargo do poder executivo.

O outro dado estatístico, plataformas mais utilizadas, pode revelar um pouco dos padrões socioeconômicos dos usuários desse volume de dados. É notório como as informações postadas por Android e iPhone são predominantes em relação as outras, ainda sendo possível observar um volume considerável de dados postados por meio do iPad.

6 Conclusões

A tendência de plataformas sociais ganharem espaço no discurso político só aumenta. É possível que dentro de alguns anos chegue ao mesmo patamar que as grandes mídias como a TV, as revistas e os jornais, no sentido de discussão e no poder de influência sobre seus usuários.

Mesmo conjecturando uma maior influência futura, já agora durante as eleições de 2018, grande parte das informações sobre as eleições foram disponibilizadas via redes sociais, fazendo delas um fator decisivo para todo o processo. Esses fatos culminaram na escolha no *Twitter* como uma das ferramentas de comunicação oficial da presidência e também a plataforma na qual os estudos desse trabalho foram feitos, o que ocorreu também nos Estados Unidos, onde o *Twitter* também é a ferramenta de comunicação do presidente.

A corrupção no Brasil é um tema atual e atuante. Notando-se cada escândalo que surge diariamente dentro do governo, repetindo o que se tem visto em todos os governos passados, estudar os aspectos dessa corrupção enraizada e como os cidadãos reagem a eles, é essencial. Faz notar como a opinião pública está reagindo aos acontecimentos e possíveis focos de onde podem ser aplicadas melhorias.

Entender como o cidadão percebe a corrupção ajuda a entender o país, ajuda e entender qual foco deve ser seguido. A corrupção em si, é um fator que atrapalha o desenvolvimento econômico e social da região onde ocorre. Conseguir estudá-la já mostra ao menos níveis maiores de liberdade da imprensa local.

Foram colhidos pouco mais de 10 mil *tweets*, e com a eliminação de *retweets*, 3567 foram usados para as análises. Com a quantidade de textos publicados, isso somente olhando os que falam sobre corruptos ou corrupções, já foi possível perceber como a corrupção em si é um tema falado, discutido, sentido e como chama uma atenção significativa do brasileiro.

Analisando os *tweets* colhidos pelo método de CHD, emergiram três classes bem definidas onde cada um dos textos se encaixaria. Todas compartilharam um fervor contra a corrupção, mas em níveis diferentes. Um apontava o sistema eleitoral como foco da corrupção, outro apontava de maneira extremada e radical a corrupção do governo passado, e por último eram os comentários anticorrupção mais moderados e que criticavam os dois lados do espectro político.

Ao fazer uma análise nos resultados coletados, percebe-se que geralmente, o *Twitter* é utilizado fortemente como uma ferramenta para expressar-se de maneira calorosa e

emocional. São poucos os argumentos utilizados nas discussões, e abundantes as adjetivações passionais. Até quando são encontrados comentários baseados em fontes e argumentos fortes, são perdidos no mar de *tweets* que não apresentam essas características.

Dada a procura por percepções sobre corrupção utilizando *tweets* postados durante o processo eleitoral brasileiro de 2018, foram encontradas percepções como: classe política extremamente corrupta, um salvador da pátria para muitos, o sistema eleitoral corrupto sendo o próprio Tribunal Superior Eleitoral responsável por fraudes, uma faixa de pessoas entrega a eleição de corruptos quase exclusivamente a uma região brasileira. Uma última percepção apresenta a corrupção como parte da estrutura do país, logo não faz diferença trocar o governo. Essas foram as principais percepções encontradas nos *tweets* estudados do dia 7.

Para fazer uma análise mais profunda das percepções, seria interessante avaliar a possibilidade de assinar a API do *Twitter* para coletar uma maior quantidade de dados. Com um maior volume de dados existe a chance de obter resultados mais precisos.

Referências

AHMED, S.; SKORIC, M. M. My name is khan: the use of twitter in the campaign for 2013 pakistan general election. In: IEEE. *2014 47th Hawaii International Conference on System Sciences (HICSS)*. [S.l.], 2014. p. 2242–2251. Citado 2 vezes nas páginas 21 e 23.

BRASIL. Lei nº 9.504, de 30 de setembro de 1997. ago 2018. Estabelece normas para as eleições. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l9504.htm>. Acesso em: 30 out. 2018. Citado na página 17.

CALDERON, N. A. et al. Mixed-initiative social media analytics at the world bank: Observations of citizen sentiment in twitter data to explore. In: IEEE. *2015 IEEE International Conference on Big Data (Big Data)*. [S.l.], 2015. p. 1678–1687. Citado 3 vezes nas páginas 21, 22 e 23.

CAMARGO, B. V.; JUSTO, A. M. Tutorial para uso do software de análise textual iramuteq. *Florianopolis-SC: Universidade Federal de Santa Catarina*, 2013. Citado na página 26.

CARVALHO, C. de S. et al. The people have spoken: conflicting brazilian protests on twitter. In: IEEE. *System Sciences (HICSS), 2016 49th Hawaii International Conference on*. [S.l.], 2016. p. 1986–1995. Citado na página 21.

CHARTIER, J.-F.; MEUNIER, J.-G. Text mining methods for social representation analysis in large corpora. *Papers on Social Representations*, v. 20, n. 2, p. 37–1, 2011. Citado na página 26.

EMARKETER. *Twitter's User Base to Grow by Double Digits This Year*. 2016. Disponível em: <https://www.emarketer.com/Article/Twitters-User-Base-Grow-by-Double-Digits-This-Year/1014243>. Acesso em: 02 nov. 2018. Citado na página 20.

ESTADÃO. *Apuração 1º turno*. 2018. Disponível em: <https://politica.estadao.com.br/eleicoes/2018/cobertura-votacao-apuracao/primeiro-turno>. Acesso em: 11 jan. 2019. Citado 2 vezes nas páginas 18 e 19.

EXAME. Com 6 votos contrários e 1 a favor, candidatura de lula é rejeitada no tse. ago 2018. Disponível em: <https://exame.abril.com.br/brasil/tse-rejeita-a-candidatura-de-lula/>. Acesso em: 30-10-2018. Citado na página 18.

FINK, C. et al. Twitter, public opinion, and the 2011 nigerian presidential election. In: IEEE. *Social Computing (SocialCom), 2013 International Conference on*. [S.l.], 2013. p. 311–320. Citado 2 vezes nas páginas 21 e 23.

FRANÇA, F. O. de; GOYA, D. H.; PENTEADO, C. L. de C. User profiling of the twitter social network during the impeachment of brazilian president. *Social Network Analysis and Mining*, Springer, v. 8, n. 1, p. 5, 2018. Citado na página 21.

G1. *Cabo Daciolo, do Patriota, fica em 6º na corrida presidencial*. 2018. Disponível em: <https://g1.globo.com/politica/eleicoes/2018/noticia/2018/10/08/cabo-daciolo-do-patriota-fica-em-6o-na-corrida-presidencial.ghtml>. Acesso em: 11 jan. 2019. Citado 2 vezes nas páginas 18 e 19.

GAZETA DO POVO. *Histórico do total de votos para presidente*. 2018. Disponível em: <https://especiais.gazetadopovo.com.br/eleicoes/2018/resultados/historico-votos-validos-presidente/>. Acesso em: 02 nov. 2018. Citado na página 19.

GONÇALVES, M.; SANTOS, S. d. Aplicação da análise de correspondência à avaliação institucional da felilcam. *IV EPCT-Encontro de produção cientifica e tecnológica, Campo Mourão*, 2009. Citado na página 26.

GROVER, P. et al. The untold story of usa presidential elections in 2016-insights from twitter analytics. In: SPRINGER. *Conference on e-Business, e-Services and e-Society*. [S.l.], 2017. p. 339–350. Citado 2 vezes nas páginas 20 e 21.

GUIMARAES, A.; WANG, L.; WEIKUM, G. Us and them: Adversarial politics on twitter. In: IEEE. *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. [S.l.], 2017. p. 872–877. Citado 2 vezes nas páginas 21 e 23.

JUSTO, A. M.; CAMARGO, B. V. *Estudos qualitativos e o uso de softwares para análises lexicais* [1]. 2014. Citado 2 vezes nas páginas 29 e 37.

KITCHENHAM, B. et al. Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, Elsevier, v. 51, n. 1, p. 7–15, 2009. Citado na página 46.

LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. *Mining of massive datasets*. [S.l.]: Cambridge university press, 2014. Citado na página 23.

MARCHAND, P.; RATINAUD, P. L'analyse de similitude appliquée aux corpus textuels: les primaires socialistes pour l'élection présidentielle française (septembre-octobre 2011). *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. JADT*, v. 2012, p. 687–699, 2012. Citado na página 26.

NECHAI, V.; GONCHAROV, D. Russian anti-corruption protests: How russian twitter sees it? In: SPRINGER. *International Conference on Digital Transformation and Global Society*. [S.l.], 2017. p. 270–281. Citado 2 vezes nas páginas 21 e 22.

NEXO JORNAL. *Por que Bolsonaro cresce mesmo sem palanque e tempo de TV*. 2018. Disponível em: <https://www.nexojornal.com.br/expresso/2018/09/23/Por-que-Bolsonaro-cresce-mesmo-sem-palanque-e-tempo-de-TV>. Acesso em: 15 dez. 2018. Citado na página 18.

NUGROHO, A. S.; DOEWES, A. et al. Twitter sentiment analysis of dki jakarta's gubernatorial election 2017 with predictive and descriptive approaches. In: IEEE. *Computer, Control, Informatics and its Applications (IC3INA), 2017 International Conference on*. [S.l.], 2017. p. 89–94. Citado 2 vezes nas páginas 21 e 23.

PAIVA, A. L. de; GARCIA, A. S.; ALCÂNTARA, V. de C. Disputas discursivas sobre corrupção no brasil: Uma análise discursivo-crítica no twitter. *RAC-Revista de Administração Contemporânea*, Associação Nacional de Pós-Graduação e Pesquisa em Administração, v. 21, n. 5, 2017. Citado na página 21.

RATINAUD, P.; MARCHAND, P. Application de la méthode alceste à de "gros" corpus et stabilité des "mondes lexicaux": analyse du "cablegate" avec iramuteq. *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles*, p. 835–844, 2012. Citado na página 26.

RODRIGUES, A. P.; RAO, A.; CHIPLUNKAR, N. N. Sentiment analysis of real time twitter data using big data approach. In: IEEE. *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*. [S.l.], 2017. p. 1–6. Citado na página 21.

SALVIATI, M. E. *Manual do Aplicativo Iramuteq*. [S.l.]: Planaltina, 2017. Citado na página 26.

SCHACTER, D. L.; GILBERT, D. T.; WEGNER, D. M. *Introducing psychology*. [S.l.]: Macmillan, 2009. Citado na página 16.

SPYER, J. et al. Tudo o que você precisa saber sobre twitter (você já aprendeu em uma mesa de bar). *Um guia prático para pessoas e organizações. Talk:(talk2. com. br). Disponível em http://www. scribd. com/doc/18384369/Manual-Twitter-Melhorresolucao-0-MB. Acesso em*, v. 11, n. 08, 2009. Citado 2 vezes nas páginas 19 e 20.

TAKIKAWA, H.; NAGAYOSHI, K. Political polarization in social media: Analysis of the "twitter political field" in japan. In: IEEE. *Big Data (Big Data), 2017 IEEE International Conference on*. [S.l.], 2017. p. 3143–3150. Citado 2 vezes nas páginas 21 e 23.

TEIXEIRA, C. R. G. et al. Humor, support and criticism: a taxonomy for discourse analysis about political crisis on twitter. In: ACM. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. [S.l.], 2018. p. 68. Citado 3 vezes nas páginas 21, 22 e 23.

TRANSPARENCY INTERNATIONAL. *ÍNDICE DE PERCEPÇÃO DA CORRUPÇÃO 2018*. 2019. Disponível em: <https://ipc2018.transparenciainternacional.org.br/?gclid=CjwKCAjw6vvoBRBtEiwAZq-T1fbJV-lK4gkkrvvgcaknmbCyiiEG0B8Q7S_CnQRGE5WBgmiVtcwUfxoCixIQAvD_BwE>. Acesso em: 06 jun. 2019. Citado 2 vezes nas páginas 16 e 17.

TRIBUNAL SUPERIOR ELEITORAL. Calendário eleitoral (eleições 2018). dez 2017. RESOLUÇÃO Nº 23.555, DE 18 DE DEZEMBRO DE 2017. Disponível em: <http://www.tse.jus.br/legislacao-tse/res/2017/RES235552017.html>. Acesso em: 29-10-2018. Citado na página 17.

TWITTER. *Twitter turns six*. 2012. Disponível em: <https://blog.twitter.com/official/en_us/a/2012/twitter-turns-six.html>. Acesso em: 01 nov. 2018. Citado na página 20.

TWITTER. *Tweeting Made Easier*. 2017. Disponível em: <https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html>. Acesso em: 01 nov. 2018. Citado na página 19.

TWITTER. *Como foram as #Eleições2018 no Twitter*. 2018. Disponível em: <https://blog.twitter.com/official/pt_br/topics/company/2018/como-foram-as-eleicoes-2018-no-twitter.html>. Acesso em: 02 nov. 2018. Citado 2 vezes nas páginas 14 e 20.

TWITTER. *Twitter e as #Eleições2018 no Brasil*. 2018. Disponível em: <https://blog.twitter.com/official/pt_br/topics/company/2018/twitter-e-as-eleicoes-2018-no-brasil.html>. Acesso em: 02 nov. 2018. Citado 2 vezes nas páginas 14 e 20.

VALLINA-RODRIGUEZ, N. et al. Los twindignados: The rise of the indignados movement on twitter. In: IEEE. *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. [S.l.], 2012. p. 496–501. Citado 2 vezes nas páginas 21 e 23.

VARIETY. *Twitter Posts Strong Q2 Earnings, but Monthly Users Drop by 1 Million Amid Cleanup Effort*. 2018. Disponível em: <https://variety.com/2018/digital/news/twitter-q2-2018-earnings-monthly-users-drop-1202887959/>. Acesso em: 01 nov. 2018. Citado na página 20.

VAZ, P.; VELASCO, F. Corrupção: Problema e Questão. *Compolítica*, v. 7, n. 2, p. 63–86, 2018. ISSN 2236-4781. Citado na página 14.

Apêndices

APÊNDICE A – Revisão Sistemática

Para realizar uma síntese dos artigos que envolvam política, corrupção e *Twitter*, foi escolhido o método da revisão sistemática da literatura. Utilizando essa abordagem, é possível sintetizar muito do que há de relevante na comunidade científica tratando-se de um assunto específico (KITCHENHAM et al., 2009).

Para realizar a revisão, foi desenvolvido um método baseado no mesmo utilizado e desenvolvido por Kitchenham et al. (2009), para fazer estudos na área de Engenharia de Software, consistindo em:

- definir pergunta(s) guia(s);
- criar critérios para aceitação e exclusão de artigos;
- elaborar estratégia de pesquisa.

A.1 Pergunta guia

Define-se uma pergunta guia para manter o foco da pesquisa durante todo o desenvolvimento da revisão. Ao final do método a pergunta é respondida com os resultados obtidos. A pergunta relativa a esse estudo, foi:

- Política e corrupção, como estão sendo abordadas no *Twitter*?

Essa pergunta é relevante para descobrir quais são os métodos e as técnicas que estão sendo utilizadas para trabalhar com esses assuntos.

A.2 Critérios de aceitação e exclusão

Para que um artigo fosse escolhido para revisão, seu texto teria que falar sobre política ou corrupção, e ser contextualizado dentro da rede social *Twitter*. Satisfazendo esses critérios, o artigo iria para uma lista de artigos selecionados.

Como critério de exclusão, foi procurado se texto apresentava algum tipo de análise de conteúdo. Esse critério foi importante para averiguar se o texto não trazia somente aspectos sociais, psicológicos ou democráticos a respeito do assunto proposto.

A.3 Estratégia de pesquisa

Para realizar a busca dos artigos, foi elaborada uma *string* de busca e ela foi utilizada dentro de algumas bases eletrônicas, sendo elas:

- *ACM Digital Library*;
- *IEEE Xplore Digital Library*;
- *Springer Link*;
- *SciELO*.

Com essa pesquisa, foi encontrado vários artigos relativos ao tema da *string* de busca, apresentado na Tabela 1.

Tabela 1 – Artigos selecionados em cada etapa (fonte: Autor)

Base	**Total**	**Total após filtro por título**	**Total após filtro por resumo**
ACM Digital Library	1	1	1
IEEE Xplore Digital Library	13	11	10
Springer Link	13	7	3
SciELO	1	1	1
Total	**28**	**20**	**15**

Para chegar em um resultado que fosse adequado e contivesse resultados melhores, foram testadas algumas *strings*. Os melhores resultados foram alcançados com a seguinte busca:

- *string:* corruption AND (politics OR political);
- restrição: o título deve conter a palavra *Twitter*.

Com a *string*, foi possível selecionar o conteúdo acadêmico de maior relevância para esse estudo. Os artigos selecionados com essa busca, já teriam maiores evidências do que era buscado para responder a pergunta guia.

A.4 Seleção dos artigos

Antes de começar a seleção, existia um total de 28 artigos. Cada artigo foi colocado em uma tabela e separado por suas respectivas bases junto com seus metadados.

Com os artigos separados, foi aplicada a seleção de artigos pelo título. Para filtragem por título, foi analisado todos os artigos selecionados para saber se não existia nenhum artigo que fugia muito ao tema. Caso tivesse pouco haver com o assunto, o artigo era excluído da revisão.

Os artigos que passaram pelo primeiro filtro, foram para o filtro de resumo onde foi possível saber com mais exatidão o tema do texto que seria lido, podendo assim ter uma seleção dos assuntos mais pertinentes. Os artigos que passaram por ambos os filtros foram para uma lista de artigos para leitura completa. Assim foi possível ter um conteúdo para leitura bem conciso.

A.5 Artigos selecionados

Quadro 2 – Artigos Selecionados e Filtrados (fonte: Autor)

Artigo	**Filtro**
Humor, support and criticism: a taxonomy for discourse analysis about political crisis on Twitter	-
Political polarization in social media: Analysis of the “Twitter political field” in Japan	-
Cyberactivism through Social Media: Twitter, YouTube, and the Mexican Political Movement "I’m Number 132"	Não traz análise de conteúdo
Us and Them: Adversarial Politics on Twitter	-
Mixed-initiative social media analytics at the World Bank: Observations of citizen sentiment in Twitter data to explore "trust"of political actors and state institutions and its relationship to social protest	-
My Name Is Khan: The Use of Twitter in the Campaign for 2013 Pakistan General Election	-
Visibilized thematics via Twitter by the Government of Chile: Two cases of corruption	-
Los Twindignados: The Rise of the Indignados Movement on Twitter	-
Twitter sentiment analysis of DKI Jakarta’s gubernatorial election 2017 with predictive and descriptive approaches	-
Twitter, Public Opinion, and the 2011 Nigerian Presidential Election	-
The People Have Spoken: Conflicting Brazilian Protests on Twitter	-
Understanding role of Twitter in addressing social causes	Fuga ao tema
Credibility in Context: An Analysis of Feature Distributions in Twitter	Fuga ao tema
Sentiment Analysis of Real Time Twitter Data Using Big Data Approach	-
Disputas Discursivas sobre Corrupção no Brasil: Uma Análise Discursivo-Crítica no Twitter	-
Russian Anti-corruption Protests: How Russian Twitter Sees It?	-
The Pragmatics of Political Messages in Twitter Communication	Fuga ao tema
User profiling of the Twitter Social Network during the impeachment of Brazilian President	-
Spanish General Elections, Microdiscourses Around #20D and Social Mobilisation on Twitter: Reality or Appearance?	Não traz análise de conteúdo
The facebook and Twitter revolutions: Active participation in the 21st century	Análise psicológica
The President on Twitter: A Characterization Study of @realDonaldTrump	Fuga ao tema
TweetCric: A Twitter-Based Accountability Mechanism for Cricket	Fuga ao tema
A Twitter Analysis of an Integrated E-Activism Campaign: #FeesMustFall - A South African Case Study	Não aborda eleições ou corrupção
Lifting Elephants: Twitter and Blogging in Global Perspective	Não traz análise de conteúdo
Complex contagions and the diffusion of popular Twitter hashtags in Nigeria	Fuga ao tema
The Untold Story of USA Presidential Elections in 2016 - Insights from Twitter Analytics	-
Automatic targeted-domain spatiotemporal event detection in twitter	Fuga ao tema
Spanish Twitter Data Used as a Source of Information About Consumer Food Choice	Fuga ao tema