UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

RÉGIS EBELING

Um Framework para análise de comportamento de grupos baseado na polarização política aplicado ao contexto da COVID-19

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Orientador: Profa. Dra. Karin Becker

Porto Alegre
2021

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Ebeling, Régis

Um Framework para análise de comportamento de grupos baseado na polarização política aplicado ao contexto da COVID-19 / Régis Ebeling. – Porto Alegre: PPGC da UFRGS, 2021.

106 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2021. Orientador: Karin Becker.

1. Polarização política. 2. Comportamento de grupos. 3. Framework de análise. 4. Redes sociais. I. Becker, Karin. II. Título.

“Todas as vitórias ocultam uma abdicação”

— SIMONE DE BEAUVOIR

AGRADECIMENTOS

Primeiramente agradeço à UFRGS e ao Instituto de Informática pela oportunidade de realizar os cursos de graduação e mestrado, é um prazer imenso ter no currículo estas importantes instituições.

Agradeço à minha orientadora, Dra. Karin Becker, que com sua orientação, paciência, motivação e oportunidades ao longo do programa de pós-graduação deixou o caminho muito mais rico, factível e agradável. Agradeço também aos demais integrantes do nosso grupo de trabalho: o colega Carlos Córdova e o professor Dr. Jéferson Nobre, nossa parceria resultou em ótimos artigos. Por último, do núcleo acadêmico, agradeço ao colega Jean Flesch, com quem troquei muitas conversas tranquilizadoras e conselhos sobre o andamento dos nossos semestres.

Agradeço aos meus pais pelo suporte e apoio ao longo do curso.

Por último, agradeço à Mirella Aguiar, que no início do mestrado era namorada e ao término virou minha esposa. Todo o suporte, incentivo, a vibração, o entendimento das ausências e principalmente a aturação das minhas mudanças de humor foram fundamentais para superar este período.

RESUMO

O aumento de discussões envolvendo de forma direta ou indireta a polarização política tem se notabilizado após uma recente onda de eleições de candidatos ligados à direita pelo mundo. Entender a influência da polarização política em grupos polarizados é importante para o estudo das dinâmicas sociais, criação de campanhas aos diversos públicos e planejamento de políticas públicas. Neste trabalho, propomos um framework de análise da influência da polarização política no comportamento de grupos, assim como dois casos de estudo que mostram a generalização da análise em grupos polarizados em diferentes assuntos. O framework desenvolvido agrega múltiplas dimensões para análise: a) técnica para inferir automaticamente a polarização política dos usuários; b) análise da rede social dos grupos e detecção de comunidades; c) modelagem de tópicos para identificar assuntos comentados pelos grupos; d) análise das propriedades linguísticas para identificação de aspectos psicológicos; e) análise das fontes de informação disseminadas pelos grupos e f) análise da demografia. A aplicação do framework em dois casos de estudo revelou padrões de comportamento em comum nas ideologias em ambos os casos: os grupos são alinhados conforme ideologia, defendem seu lado político, criticam as partes contrárias, e demonstram um comportamento de câmara de eco. A aplicação do framework nos estudos de caso mostra a capacidade do mesmo de confirmar os posicionamentos políticos nos grupos e observar os efeitos dessa polarização no comportamento dos usuários.

Palavras-chave: Polarização política. comportamento de grupos. framework de análise. redes sociais.

A framework to analyze group behavior based on political polarization applied to the COVID-19 context

ABSTRACT

The increase in discussions directly or indirectly involving political polarization has been highlighted after a recent wave of candidates linked to the right-wing around the world. Understanding the influence of political polarization on polarized groups is essential for studying social dynamics, creating campaigns for different audiences, and planning public policies. In this work, we propose a framework for analyzing the influence of political polarization on group behavior and two case studies that show the generalization of the analysis in polarized groups in different subjects. The developed framework adds multiple dimensions for analysis: a) technique to automatically infer the political polarization of users; b) analysis of the groups' social network and detection of communities; c) topic modeling to identify issues discussed by the groups; d) analysis of linguistic properties to identify psychological aspects; e) analysis of the sources of information disseminated by the groups and f) analysis of demography. The framework's application in two case studies revealed patterns of behavior common to ideologies in both cases: groups are aligned according to ideology, defend their political side, criticize opposing parties, and demonstrate "echo chamber" behavior. The application of the framework in the case studies shows its ability to confirm political positions in groups and observe the effects of this polarization on user behavior.

Keywords: political polarization, group behavior, framework fnalysis, social networks.

LISTA DE ABREVIATURAS E SIGLAS

LDA	Latent Dirichlet Allocation
BERT	Bidirectional Encoder Representations from Transformers
LIWC	Linguistic Inquiry and Word Count
UMAP	Uniform Manifold Approximation and Projection
HDBScan	Hierarchical Density-Based Spatial Clustering of Applications with Noise
TF-IDF	Term Frequency - Inverse Document Frequency
PLSA	Probabilistic Latent Semantic Analysis
CV	Coherence Value
iGPS	GPS Ideológico
IPP	Índice de Polarização Política

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O Brasil vive um cenário político cada vez mais polarizado, com fatores que acabaram impulsionando ainda mais as tensões na última década. Desde as altamente competitivas eleições presidenciais de 2014, que resultaram em um impeachment dois anos depois da presidente Dilma Roussef, os eleitores fortaleceram suas convicções políticas em relação à esquerda ou à direita. As eleições presidenciais de 2018 - vencidas por Jair Bolsonaro - dividiram ainda mais a população, uma vez que os eleitores escolheram lados baseados principalmente em posturas "anti", principalmente contra o Partido Trabalhista (PT), que governava o país desde 2002. Muito da postura de rejeição ao PT foi baseada nas acusações de diversos escândalos de corrupção que afetaram a economia e acabaram levando Luís Inácio Lula da Silva, presidente por dois mandatos, à prisão. Jair Bolsonaro surgiu como um candidato de oposição ao PT, representando valores tradicionais como família, Deus, patriotismo e conservadorismo, e abordando temas em sua campanha como restauração da ordem, combate à corrupção e impulsão da economia. Por outro lado, ele expressou muitas opiniões polêmicas ao longo do tempo sobre temas como mudança climática, proteção do meio ambiente, gênero e movimentos LGBT, racismo e ciência, representando para os eleitores de esquerda uma figura inversa aos seus valores.

Neste cenário polarizado, com forte apoio e duras críticas ao atual governo, instalou-se a pandemia mundial da COVID-19 (SARS-Cov-2) no início de 2020, com consequências à saúde e à economia. Esta dissertação concentra-se sobre dois eventos que encadearam duas das mais fortes discussões ocorridas no contexto da pandemia: o isolamento social e a vacinação.

Em um primeiro momento, à luz do total desconhecimento da doença e possíveis tratamentos, instaurou-se um dilema em muitos países entre vidas e economia quando a ciência enfatizou o distanciamento social como a mais efetiva forma de combater o contágio do vírus. Em março de 2020, a resposta inicial do governo brasileiro, representada pelo então Ministro da Saúde Luiz Mandetta, era centrada no isolamento social, amparada pelas evidências científicas disponíveis à época e por experiências recentes de outros países. Essa orientação foi seguida pela maioria dos governadores, que tiveram que lidar diretamente com os aspectos práticos de leitos hospitalares disponíveis e recursos para atendimento médico, considerando o número crescente de casos. Por outro lado, o presidente Bolsonaro questionou sistematicamente o impacto do isolamento social na economia, dizendo que seria mais prejudicial à população do que o próprio vírus. Além

disso, ele defendeu medicamentos com eficácia questionável, como (hidroxi)cloroquina, bem como os benefícios da imunidade de rebanho por meio de infecção não controlada. Este confronto levou à demissão de dois Ministros da Saúde com formação médica e a suas substituições por um ministro interino (e após efetivado) com formação militar. O embate de ideias entre presidente e Ministros da Saúde foi representado por dois grandes movimentos: pró e anti-distanciamento social, comumente sendo referidos pelas mídias tradicionais e redes sociais como "*Quarenteners*"e "*Cloroquiners*".

Devido à urgência da pandemia, esforços foram empreendidos para desenvolver vacinas COVID-19, aprová-las e disponibilizá-las no mais curto espaço de tempo possível. Esses esforços tiveram que seguir os mesmos requisitos legais de qualidade, segurança e eficácia farmacêutica que outros medicamentos. Os programas de imunização da COVID começaram na Europa no final de 2020. Apesar de ser reconhecida como uma das medidas de saúde pública de maior sucesso, a vacinação é percebida como insegura e desnecessária por um número crescente de pessoas (HORNSEY M. J., 2018). No caso do COVID, esse medo foi ampliado, e as vacinas SARS-CoV-2 têm sido alvo de todos os tipos de notícias falsas e desinformação (CATALAN-MATAMOROS; ELíAS, 2020). No mundo todo os movimentos anti-vacinação têm implicado a redução das taxas de aceitação da vacina e o aumento de surtos de doenças evitáveis pela vacina. No entanto, estudos indicam que as atitudes anti-vacinação estão mais relacionadas ao pensamento conspiratório, do que preconceito político ou crenças religiosas (HORNSEY M. J., 2018). A hesitação vacinal e a desinformação apresentam obstáculos substanciais para alcançar a cobertura e a imunização da comunidade no contexto da COVID-19 (BURKI, 2020).

O Brasil tem uma história de sucesso na erradicação de doenças devido aos programas de imunização em larga escala. Um estudo mostra que os brasileiros apresentam uma das maiores taxas de aceitação da vacinação COVID-19 entre 19 países pesquisados (cerca de 85%) (LAZARUS et al., 2020). No entanto, muitos argumentam que o comportamento de Bolsonaro em relação à vacinação da COVID-19 retardou a definição de um programa nacional de imunização. Por outro lado, João Dória, governador de São Paulo, traçou ainda em 2020 um programa estadual de imunização para começar no final de janeiro de 2021. O programa paulista é baseado na Coronavac, vacina desenvolvida pelo Instituto Butantan em parceria com o laboratório chinês Sinovac. Visto que Bolsonaro e Dória são ambos possíveis candidatos às eleições presidenciais de 2022, a vacinação contra a COVID tem sido discutida sob um forte viés político. Dois movimentos pró e anti-vacina são formados nesta discussão: *Pro-vaxxers* e *Anti-vaxxers*, sendo que neste

último nota-se uma oposição específica à vacina "de origem chinesa", grupo que denominamos neste trabalho de *Anti-sinovaxxers*.

Esforços de pesquisa significativos abordaram o discurso relacionado à COVID nas redes sociais. Trabalhos focados em modelagem de tópicos e modelos de difusão são pesquisados em (ORDUN; PURUSHOTHAM; RAFF, 2020; LYU et al., 2020; CURIEL; RAMíREZ, 2020). Em relação à influência da polarização política, Jiang et al. (2020) examinam diferenças geográficas nos discursos online, Sha et al. (2020a) analisam narrativas de acordo com a tomada de decisão governamental e Rao et al. (2020) investigam sua relação com o comportamento anticientífico. Em relação à influência da polarização política no comportamento da população, trabalhos como (MAKRIDIS; ROTHWELL, 2020; BRUIN; SAW; GOLDMAN, 2020) demonstram através de pesquisas que a população dos Estados Unidos foi influenciada em relação à COVID-19. A desinformação sobre a COVID é abordada em trabalhos como (CINELLI et al., 2020; Furini et al., 2020; BURKI, 2020), e a influência ideológica é enfatizada em (HAVEY, 2020a). O sentimento de brasileiros e norte americanos em relação à COVID é comparado em (GARCIA; BERTON, 2021). Ao nosso conhecimento, não há trabalhos publicados que consigam relacionar ideologia política com comportamento anti-vacinal (e.g., (HORNSEY M. J., 2018; CZARNEK GABRIELA; SZWED, 2020)).

Os trabalhos relacionados à COVID-19 também trazem à tona o papel da polarização política influenciando os contextos analisados. Usando informações disponíveis em redes sociais, há estudos que identificam automaticamente a orientação política dos usuários baseado em dados de seu perfil e/ou *posts* (BARBERá et al., 2015; GARIMELLA; WEBER, 2017; CONOVER et al., 2011; PREOŢIUC-PIETRO et al., 2017), ou que analisam a influência da polarização política em um dado contexto com métodos computacionais específicos, tais como modelagem de tópicos (RAO et al., 2020; MAKRIDIS; ROTHWELL, 2020; HAVEY, 2020a) ou análise de aspectos psicológicos com base no estilo de escrita (DEMSZKY et al., 2019; PENNYCOOK et al., 2020). Nestes trabalhos a análise da polarização política restringe-se à dimensão da polarização estudada.

Outros trabalhos propõem frameworks de análise da polarização política multidimensionais. Os frameworks descritos em (BRAMSON et al., 2016; LELKES, 2016) são baseados em questionários especificamente construídos para este propósito, e portanto sua aplicação é sujeita ao custo e latência de coletas destes dados. O framework proposto em (STIEGLITZ; DANG-XUAN, 2013) é voltado a redes sociais, e inclui técnicas computacionais para explorar 3 dimensões: análise de tópicos, análise de estrutura

de uma rede e análise de sentimento.

O objetivo deste trabalho é analisar como a polarização política afeta o comportamento de grupos brasileiros formados no Twitter com posturas opostas em relação à distância social e à vacinação, ambas relacionadas ao enfrentamento à COVID-19. Para a tarefa, propomos uma estrutura de análise multidimensional com as respectivas técnicas computacionais focadas nas seguintes questões de pesquisa:

- Q1: Os grupos são politicamente polarizados?
- Q2: Esses grupos têm preocupações diferentes?
- Q3: A polarização política de grupos afeta sua estrutura de rede social?
- Q4: Os grupos manifestam aspectos psicológicos diferentes?
- Q5: As fontes de informação são de origem diferente?
- Q6: Esses grupos têm dados demográficos semelhantes?

O framework foi aplicado em dois estudos de caso utilizando dados do Twitter. Para cada caso foram coletados usuários que utilizaram hashtags específicas para caracterizar algum posicionamento, formando grupos com posicionamentos distintos, além de um grupo sem posicionamento aparente para controle. O primeiro estudo de caso é referente ao isolamento social (EBELING et al., 2020; EBELING et al., 2020; EBELING et al., 2021), onde coletamos dados de 3 grupos (*Quarenteners* - pró isolamento, *Cloroquiners* - contra isolamento, e *Neutros*) entre 22 de março e 07 de abril de 2020. O segundo estudo de caso é sobre a vacinação, com coleta de *tweets* de 4 grupos (*Pro-vaxxers* - pró vacinação, *Anti-vaxxers* - anti-vacinação, *Anti-sinovaxxers* - anti-vacinação com Coronavac, e *Neutros*), considerando o período de 1º de janeiro de 2020 a 1º de abril de 2021. Os resultados mostram que os grupos são alinhados conforme ideologia, defendem os elementos polarizados de seus lados (e.g. políticos, ações, campanhas), criticam os elementos contrários, e demonstram um comportamento de câmara de eco. Embora o framework tenha sido proposto para a análise de assuntos relacionados à COVID, ele pode ser empregado para entender o impacto da polarização política nas posturas que atualmente dividem ainda mais a população global, como mudanças climáticas, privatizações, violência de gênero e racial, entre outras.

As principais contribuições deste trabalho são:

- análise de dois estudos de casos com grupos de posicionamentos opostos no Twitter quanto ao isolamento social e vacinação no contexto da COVID-19, e a influência da polarização política nestes posicionamentos. Estas análises envolvem múltiplas dimensões que formam uma visão mais completa dos estudos de caso relacionados à COVID-19 se comparadas aos trabalhos relacionados (e.g. (RAO et al., 2020; MAKRIDIS; ROTHWELL, 2020; HAVEY, 2020a));

- um *framework* consistindo em uma estrutura de análise multidimensional que abrange as preocupações expressas, quantificação da polarização política, estrutura da rede social e aspectos psicológicos, demografia e fontes de informação para analisar o comportamento de grupos baseados na polarização, agregando mais dimensões de análise que *framework* anteriores (BRAMSON et al., 2016; LELKES, 2016; STIEGLITZ; DANG-XUAN, 2013);

- um método para modelagem de tópicos que combina de forma complementar LDA (BLEI; NG; JORDAN, 2003) e BERTopic (GROOTENDORST, 2020), para compreender os argumentos que sustentam as posturas pró/contra a distância social. Enquanto o primeiro permite identificar preocupações em granularidade maior por co-ocorrência de palavras, o último permite identificar os argumentos representativos usados para expressar cada postura utilizando similaridade de sentenças. Os trabalhos relacionados (RAO et al., 2020; MAKRIDIS; ROTHWELL, 2020; HAVEY, 2020a) restringem-se ao uso do LDA;

- construção de um índice de polarização política combinando a técnica de Garimella e Weber (2017) para calcular a orientação mais à direita ou esquerda conforme proporções de usuários de direita e esquerda seguidos por um usuário no Twitter, e a técnica de Barberá et al. (2015) que fornece uma lista referencial de pessoas seguidas, utilizada no GPS Ideológico da Folha de São Paulo (iGPS)[1], do qual extraímos os políticos de esquerda e direita referência para a polarização;

- uso das técnicas de análise de redes e detecção de comunidades (COSTA et al., 2007) para compreender características sociais de cada grupo estudado, e a influência de políticos nas mesmas.

Esta dissertação está organizada como segue. O Capítulo 2 apresenta a fundamentação teórica, resumindo as técnicas utilizadas para desenvolver o *framework* de análise.

[1] http://temas.folha.uol.com.br/gps-ideologico/

O Capítulo 3 sumariza os trabalhos relacionados. O Capítulo 4 apresenta uma visão geral do *framework*, e detalha as técnicas propostas para determinar os grupos, caracterizar os usuários conforme sua polarização política, analisar suas preocupações e demografia, examinar os efeitos da polarização na estrutura das redes sociais, identificar os aspectos psicológicos e as fontes de informações. As ameaças à validade do framework são apresentadas no Capítulo 5. Os estudos de caso discutindo o isolamento social e a vacinação no contexto da COVID-19 são desenvolvidos nos Capítulos 6 e 7, respectivamente. Por fim, conclusões e trabalhos futuros são apresentados no Capítulo 8.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados de forma resumida os principais conceitos e técnicas utilizadas neste trabalho.

2.1 Modelagem de Tópicos

Para compreender os assuntos contidos em grandes volumes da dados, diversas técnicas computacionais podem ser aplicadas em coleções de documentos. A área de Modelagem de Tópicos trata destas técnicas computacionais que buscam agrupar documentos de um corpus por sua semelhança considerando algum critério. Basicamente os modelos de tópicos consideram que cada documento é uma mistura de tópicos, e cada tópico é uma coleção de palavras.

Entre os métodos de modelagem de tópicos clássicos encontrados na literatura (KHERWA; BANSAL, 2020), podem ser citados:

- Latent Semantic Analysis (LSA) (DEERWESTER et al., 1990): criação de objetos semânticos através de uma representação vetorial de textos em baixa dimensionalidade;

- Non-Negative Matrix factorization (NNMF) (LEE; SEUNG, 1999): fatoração não negativa de uma matriz representando a coleção de documentos x termos;

- Probabilistic Latent Semantic Analysis (PLSA) (HOFMANN, 1999): baseada no LSA, diferencia o uso de palavras pelo contexto;

- Latent Dirichlet Allocation (LDA) (BLEI; NG; JORDAN, 2003): baseado na distribuição e probabilidades das palavras e documentos entre tópicos.

Mais recentemente, modelos de tópicos baseados em aprendizado profundo vêm sendo propostos e amplamente utilizados. Estes modelos de representações distribuídas utilizam redes neurais para obter vetores de palavras (MIKOLOV et al., 2013), sentenças e documentos (LE; MIKOLOV, 2014b), e tópicos (ANGELOV, 2020).

O restante desta seção detalha os métodos utilizados neste trabalho: LDA e BERTopic.

2.1.1 LDA

Uma das técnicas mais populares para ajustar um modelo de tópico de um *corpus* é o LDA (BLEI; NG; JORDAN, 2003). Essa técnica não supervisionada trata cada documento do corpus como uma mistura de tópicos, em que cada tópico tem probabilidade de estar relacionado ao documento. Por sua vez, cada tópico é composto por uma lista de palavras (termos), com a respectiva probabilidade de estarem relacionados com o tópico. O LDA resulta em tópicos nos quais os termos são mais propensos a ocorrerem juntos em documentos. Os tópicos podem ter palavras sobrepostas. A entrada do LDA é um corpus, e a descoberta do número de tópicos é um parâmetro k. A saída é um conjunto de k tópicos, consistindo em termos e suas respectivas probabilidades *beta* de pertencerem ao tópico e uma probabilidade *gamma* que relaciona cada tópico e um documento do corpus. Ações típicas de pré-processamento sobre o corpus original podem melhorar os resultados (DENNY; SPIRLING, 2018), como normalização, remoção de *stopwords* e caracteres/termos especiais, lematização, etc.

Existem alguns desafios relacionados à implantação do LDA na prática para geração de um conjunto coerente de tópicos. Primeiro, é preciso atribuir um significado a cada tópico resultante, uma tarefa subjetiva. Outro problema é o parâmetro k, visto que um k grande pode resultar em tópicos redundantes, enquanto um k menor pode não ser suficiente para agrupar documentos de acordo com uma interpretação semântica significativa.

Para avaliação de resultados do LDA existem algumas métricas baseadas na distribuição de tópicos (e.g. W-Uniform, W-Vacuous e D-BGround (ALSUMAIT et al., 2009)) e em método bayesiano (MIMNO; BLEI, 2011), porém é difícil sua interpretabilidade por humanos. Outras métricas têm sido propostas para avaliar a qualidade de tópico obtidos com o LDA, como (MANNING; RAGHAVAN; SCHÜTZE, 2010), Normalized Mutual Information (NMI) (ESTÉVEZ et al., 2009) e Coherence Value (CV ou C-Valor) (RÖDER; BOTH; HINNEBURG, 2015). Tais métricas se baseiam na coerência dos termos de cada tópico para medir índices de interpretabilidade de cada.

CV é uma métrica unificadora, ampla e completa, que combina quatro dimensões para definir a coerência. Em uma extensa avaliação envolvendo sete métricas de coerência e diversos benchmarks (RÖDER; BOTH; HINNEBURG, 2015), foi a métrica que apresentou a melhor relação com a interpretabilidade dos resultados. Esta métrica se constitui agrupando:

- a) a segmentação usada para dividir um conjunto de palavras em subconjuntos, e estes sendo comparados uns com os outros;
- b) conjunto de medidas de confirmação que marca a concordância de um dado par de comparação;
- c) conjunto de métodos utilizados para estimar as probabilidades das palavras para as medidas de confirmação, as quais podem ser calculadas de maneiras diferentes;
- d) conjunto de métodos de agregação em uma única métrica dos métodos de medidas de confirmação das probabilidades.

A métrica CV foi utilizada em diferentes trabalhos como uma referência para encontrar o número adequado de tópicos em LDA, por exemplo (VARGAS-CALDERÓN et al., 2019; PUERARI et al., 2020; WALTER; BECKER, 2018).

Neste trabalho empregamos o LDA como uma das técnicas de modelagem de tópicos e utilizamos a métrica de coerência CV para avaliar a melhor medida em diversos valores do parâmetro k.

2.1.2 Top2Vec e BERTopic

As representações distribuídas de palavras e documentos como *embeddings* ganharam popularidade devido à sua capacidade de capturar semântica. Um *embedding* é um espaço de dimensão relativamente baixa no qual se podem traduzir vetores de dimensões elevadas, como palavras ou documentos. Word2Vec (MIKOLOV et al., 2013) e Doc2Vec (LE; MIKOLOV, 2014a) são técnicas clássicas não supervisionadas para extrair *embeddings* que representam palavras e documentos, respectivamente. Técnicas mais recentes permitem a descoberta de *embeddings* contextuais para representar modelos de linguagem, como BERT (DEVLIN et al., 2019), o estado-da-arte em modelos de representações de linguagem.

Top2Vec (ANGELOV, 2020) é uma abordagem alternativa para modelagem de tópicos, que aproveita um conjunto de documentos e os *embeddings* semânticos de palavras para encontrar vetores de tópicos. Top2Vec é uma estrutura que engloba algoritmos para buscar automaticamente tópicos densos em uma coleção de documentos, assumindo que documentos semanticamente semelhantes formam tópicos dentro da coleção de entrada.

A primeira etapa para usar o Top2Vec é converter todos os documentos no corpus

em vetores semânticos usando algum modelo de *embedding* para tornar os documentos semanticamente semelhantes no espaço vetorial. A etapa seguinte visa reduzir a dimensionalidade dos vetores do documento, uma vez que os vetores em espaços de alta dimensão tendem a ser muito esparsos. Top2Vec adota UMAP (NARAYAN; BERGER; CHO, 2020), uma técnica de redução de dimensões rápida e escalável que preserva a estrutura global dos dados. A etapa final é agrupar documentos semanticamente semelhantes procurando por áreas densas no espaço vetorial usando um algoritmo de agrupamento baseado em densidade, no caso, o HDBSCAN (MCINNES; HEALY, 2017). O HDBSCAN lida com clusters de ruído e de densidade variável e, portanto, atribui um rótulo a cada cluster denso de vetores de documentos e um rótulo de ruído aos vetores de documentos que não estão em um cluster denso. As áreas densas de vetores de documentos são usadas para calcular os vetores de tópicos e os documentos de ruído são descartados.

A vantagem do Top2Vec em relação ao LDA é que não requer a definição do número de tópicos a serem descobertos nem ações de pré-processamento sobre o corpus de entrada. Como desvantagem, essa técnica pode levar a um número excessivo de tópicos que dificultam muito a interpretação dos resultados.

BERTopic (GROOTENDORST, 2020) é uma extensão do Top2Vec, que fornece suporte mais amplo para modelos de *embedding*, incluindo os *embeddings* BERT de última geração. Engloba também uma etapa adicional na construção dos tópicos usando uma abordagem TF-IDF para caracterizar as palavras mais representativas e distintas. Por último, a ferramenta fornece um recurso de visualização com os tópicos encontrados na coleção em um espaço 2D representando a distância semântica dos mesmos. Cada tópico é representado com o tamanho refletindo a quantidade de documentos contidos nele, além de ser possível visualizar os termos mais significantes para a classificação de documentos no tópico.

Neste trabalho, utilizamos o BERTopic como técnica complementar ao LDA na modelagem de tópicos, procurando localizar e entender os argumentos centrais expressos por grupos de usuários nos diferentes tópicos.

2.2 Análise de Redes

2.2.1 Métricas Topológicas

Para identificar se a polarização política afeta a estrutura da rede social de cada grupo, aplicamos técnicas de análise de redes sobre a estrutura da rede social. A análise de redes consiste em estudar as propriedades e características de redes (ou grafos), as quais são compostas por elos chamados arestas que conectam um conjunto de nós (HANSEN et al., 2020). Grafos são muito usados para representar redes sociais na internet. Neste caso, os nós normalmente representam entidades sociais (geralmente pessoas) conectadas por arestas que representam relacionamentos estáticos (por exemplo, amizade, seguidor, assinante) ou dinâmicos (e.g. responder, mencionar). No Twitter, por exemplo, os usuários e seus relacionamentos de seguidores/seguidos podem ser abstraídos como um conjunto de nós (usuários), arestas de entrada (usuários seguidores), e arestas de saída (usuários seguindo) em que a origem da aresta é o seguidor e o destino é o usuário seguido.

As métricas topológicas que descrevem uma rede apresentam várias propriedades que podem fornecer *insights* sobre a natureza e o comportamento dos usuários em uma rede social. Algumas métricas são (COSTA et al., 2007):

- Números de nodos e arestas: descrevem o quão grande é uma rede;
- Grau de entrada e saída dos nodos: no contexto de redes sociais como o Twitter, representam quantos usuários são seguidores de um nodo e quantos usuários são seguidos pelo nodo, respectivamente;
- Grau médio: média dos graus de todos os nodos, representando o quão conectados os nodos estão na rede;
- Caminho mais curto médio: representa o quão próximos, em média, os nodos estão uns dos outros na rede;
- Diâmetro da rede: comprimento do menor caminho mais longo entre quaisquer dois nodos na rede, fornecendo uma intuição de como pode ser difícil chegar a um nodo a partir de qualquer outro na rede;
- Coeficiente de clusterização: representa a probabilidade de encontrar subgrupos de nodos altamente conectados na rede;

- Centralidade de proximidade: calculada individualmente para cada nodo, é baseada no caminho mais curto entre um determinado nodo e todos os demais nodos da rede. Os nodos com uma pontuação de proximidade alta têm as distâncias mais curtas em relação a todos os outros nodos. A centralidade de proximidade é uma forma de detectar nós que podem espalhar informações de forma eficiente pela rede;

- Centralidade de intermediação de um nodo: baseada no número de caminhos mais curtos entre todos os pares de nodos que passam por ele. É uma medida que revela a importância de um nodo para possibilitar a comunicação entre outros pares de nodos, visto que faz parte do caminho mais curto entre eles. Em outras palavras, nodos com alta intermediação podem ter uma influência considerável dentro de uma rede em virtude de seu controle sobre a informação que passa entre nodos diferentes.

Neste trabalho utilizamos as métricas topológicas mencionadas para caracterizar as redes de grupos de forma geral e comunidades de interesse detectadas.

2.2.2 Detecção de Comunidades

A detecção de comunidades é uma técnica frequentemente usada na análise de redes sociais (BAZZAN, 2020; BEDI; SHARMA, 2016; CONOVER et al., 2011). A detecção de comunidade visa encontrar grupos de nodos (comunidades) que estão altamente conectados entre si, mas fracamente conectados com nodos de outras comunidades (FORTUNATO, 2010). Na análise de redes sociais, essa técnica permite identificar usuários que compartilham os mesmos padrões sociais dentro da rede, e até mesmo em esferas sociais como a política, estudadas neste trabalho.

Existem vários algoritmos para a tarefa de detecção de comunidade. Um deles é o Método Louvain (BLONDEL et al., 2008), que tem como foco a otimização da modularidade da rede. Neste trabalho, usamos este método, o qual está disponível na ferramenta Gephi[1].

[1] https://gephi.org/

2.3 LIWC e Aspectos Psicológicos

Aspectos psicológicos podem ser identificados pelas de palavras empregadas, pois revelam estados emocionais e biológicos, estilos de pensamento e outros traços de personalidade (TAUSCZIK; PENNEBAKER, 2010). Uma ferramenta amplamente utilizada nesta tarefa de tradução de palavras em categorias é o LIWC (PENNEBAKER J. W.; BOOTH, 2001), que engloba a funcionalidade de análise e um dicionário léxico bastante completo.

O LIWC foi concebido a partir de estudos e observações que indicaram que a escrita de um indivíduo traduzia aspectos psicológicos, principalmente nos processos de adoecimento ou recuperação. Tausczik e Pennebaker (2010) destacam duas categorias extensas de palavras com diferentes propriedades psicométricas e psicológicas: conteúdo e estilo. De uma perspectiva psicológica, palavras de conteúdo (por exemplo, substantivos, verbos regulares e muitos adjetivos e advérbios) transmitem o que as pessoas dizem. Em contraste, palavras de estilo (por exemplo, pronomes, preposições, artigos, conjunções, verbos auxiliares) refletem como as pessoas se comunicam.

A construção do dicionário LIWC teve por objetivo viabilizar a automação da contagem de palavras nas respectivas categorias psicológicas para os textos de entrada da ferramenta, possibilitando que as mesmas sejam combinadas para caracterizar um aspecto psicológico de um indivíduo ou grupo. Os resultados empíricos usando LIWC resumidos em (TAUSCZIK; PENNEBAKER, 2010) demonstram sua capacidade de detectar significado em uma ampla variedade de configurações experimentais, incluindo foco atencional, emoções, relações sociais, estilos de pensamento e diferenças individuais.

O dicionário LIWC divide as palavras em categorias e subcategorias de acordo com 4 dimensões linguísticas:

- Processos linguísticos: engloba categorias de pronomes, tempos verbais, artigos, preposições, entre outros;
- Processos psicológicos: engloba categorias de ações e sentidos como processos cognitivos, afetivos, sociais, entre outros;
- Preocupações pessoais: engloba categorias em forma de tópicos de preocupações tais como trabalho, lazer, dinheiro, etc;
- Linguagem informal: engloba categorias de preenchimento de sentenças, palavras remetendo concordância e abreviações.

Cada categoria, por sua vez, pode estender-se em até três níveis de subcategorias. Por exemplo, as subcategorias "Preposições" e "Conjunções" fazem parte da categoria de "Palavras de função total", que por sua vez integra a dimensão de "Processos Linguísticos".

Neste trabalho, utilizamos o LIWC para avaliar se grupos polarizados apresentam aspectos psicológicos característicos.

3 TRABALHOS RELACIONADOS

Este capítulo apresenta os principais trabalhos relacionados à proposta do framework, assim como cada dimensão.

3.1 Análise de fenômenos sociais no Twitter

O Twitter é uma rede social que disponibiliza um ambiente propício para coleta de dados, análise de diferentes cenários e com diferentes propósitos. Os dados disponíveis através da rede social podem caracterizar atributos do usuário (foto de perfil, localização informada, listas de seguidores e de quem seguem), atributos de um tweet (conteúdo textual do tweet, menções) ou comportamentos de postagem (curtidas, respostas ou *retweets*). Devido ao tamanho historicamente limitado de um tweet, hashtags passaram a ser usadas para resumir o assunto postado e as opiniões e posicionamentos do usuário (MACHADO et al., 2018). Este grande volume de dados tem permitido estudar fenômenos sociais em larga escala, tais como igualdade de gênero (ELSHERIEF; BELDING; NGUYEN, 2017), igualdade racial (CHOUDHURY et al., 2016), reações emocionais a eventos violentos (HARB; EBELING; BECKER, 2020), e comportamentos anti-vacinação (TOMENY; VARGO; EL-TOUKHY, 2017; COSSARD et al., 2020).

A COVID-19 tornou-se uma atrativa área de pesquisa nas redes sociais, principalmente utilizando dados do Twitter. Muitos trabalhos investigam conversas online em termos de tópicos, difusão de informação e mudança de tópicos ao longo do tempo (ORDUN; PURUSHOTHAM; RAFF, 2020). Outros trabalhos analisam os sentimentos em tópicos relacionados à COVID-19, comparando o cenário brasileiro com outros países, como os Estados Unidos (GARCIA; BERTON, 2021). Um estudo longitudinal (SHA et al., 2020b) relaciona as narrativas do Twitter a ações de governadores e presidentes. Em relação à vacinação, estudos revelaram que embora os grupos anti-vacina não mudem sua visão sob nenhum argumento, os grupos pró-vacina têm sido reativos e reticentes por supôr que a ciência falaria por si própria (BURKI, 2020). As percepções sobre as dúvidas e a disseminação de desinformação da vacinação foram abordadas por diversos trabalhos na literatura (LYU et al., 2020; CURIEL; RAMíREZ, 2020; CATALAN-MATAMOROS; ELíAS, 2020; CINELLI et al., 2020; KALIYAR; GOSWAMI; NARANG, 2021).

Muitos trabalhos apontam a influência da polarização política no contexto da COVID-19. Nos Estados Unidos da América, estudos (MAKRIDIS; ROTHWELL, 2020;

BRUIN; SAW; GOLDMAN, 2020; MILOSH et al., 2020; BARRIOS; HOCHBERG, 2020) revelam que a afiliação partidária é frequentemente o indicador individual mais forte de comportamento e atitudes sobre COVID-19, ainda mais poderoso do que as taxas de infecção locais ou demográficas características. Outro estudo sobre dados de mobilidade (GROSSMAN et al., 2020) revelou que o partidarismo político influencia as decisões dos cidadãos de se envolverem voluntariamente no distanciamento físico em resposta às comunicações do governador do condado. Rao et al. (2020) examinam o alinhamento ideológico dos usuários ao longo das dimensões de moderação, política e ciência, concluindo que a moderação é a principal influência no comportamento da ciência. Jiang et al. (2020) examinam as diferenças geográficas no discurso online da COVID-19, relacionando a polarização ao domínio político de cada estado dos EUA. Havey (2020a) relaciona tópicos de desinformação sobre a COVID-19 com ideologia política, concluindo que os liberais são mais propensos a acreditar e espalhar desinformação. Eles concluíram que esses grupos apresentam aspectos psicológicos semelhantes, mas os liberais tendem a formar um grupo mais coeso e socialmente conectado, que não está interessado nem aberto a outros pontos de vista. A influência da polarização política também foi observada no Brasil (AJZENMAN; CAVALCANTI; MATA, 2020; SOARES F. B., 2021).

A polarização política é um fenômeno social de grande impacto que vem sendo cada vez mais estudada ao longo dos anos. Cada vez mais políticos e indivíduos politizados têm recorrido a redes sociais para propagar suas convicções entre seus grupos de apoio (EFFING; HILLEGERSBERG; HUIBERS, 2011; CONWAY; KENSKI; WANG, 2015). Para estudar a polarização política nas redes sociais, encontramos três categorias de trabalhos: identificação automática da orientação política de usuários baseada em seus dados e/ou postagens, uso de métodos computacionais para analisar a influência da polarização política em um dado contexto, e proposta de métodos/processos visando a geração de *insights* sobre um caso de análise. Estes trabalhos são detalhados nas próximas seções.

3.2 Identificação automática da orientação política

Diferentes trabalhos contribuem com técnicas para identificar automaticamente a orientação política de usuários de redes sociais, principalmente no Twitter. Conover et al. (2011) e Preoţiuc-Pietro et al. (2017) utilizam em seus estudos atributos extraídos dos *tweets* para inferir a orientação política do usuário em classes (e.g. conservadores/liberais). Já Barberá et al. (2015) e Garimella e Weber (2017) propõem métodos para calcular índices

de orientação política a partir de uma lista de usuários com orientação política conhecida de esquerda ou direita.

3.2.1 Features extraídas dos *tweets*

O trabalho de Conover et al. (2011) busca identificar a orientação política de grupos a partir da construção de redes de comunicação no Twitter. São buscados *tweets* com hashtags políticas de um conjunto de postagens coletadas antes da eleição ao Congresso dos Estados Unidos de 2010, e então são construídas redes de menções e *retweets*. A execução de algoritmos de clusterização revela que ambas as redes mostram a divisão em dois conjuntos de usuários, sendo que a rede de *retweets* revela uma separação mais acentuada. Utilizando uma técnica de análise qualitativa de conteúdo, anotadores classificaram uma amostra de usuários de cada *cluster* em cada rede baseados em seus últimos *tweets*, atribuindo rótulos de esquerda, direita, e indecisos. Como resultado da anotação, cada *cluster* é retratado como uma composição de percentuais de orientações políticas, sendo que a rede de *retweets* apresenta 80.1% de um *cluster* com usuários de esquerda, e no outro *cluster* 93.4% com usuários de direita.

Outro trabalho que busca inferir a orientação política de usuários através de características extraídas dos *tweets* é apresentado em (PREOŢIUC-PIETRO et al., 2017). Os experimentos utilizam um dataset construído através de questionário com usuários auto declarados em 7 graus de orientação política (muito liberais a muito conservadores) e outro dataset com orientação binária (liberais ou conservadores). Para construir os modelos preditivos de orientação política, os autores propõem features extraídas dos *tweets* dos usuários, a saber: a) os percentuais de uso dos atributos unigramas (*bag of words*); b) categorias LIWC de palavras usadas; c) *clusters* pré-definidos construídos com Word2Vec; d) sentimentos/emoções; e e) termos políticos. Foram construídos dois modelos de predição, um utilizando regressão logística para inferir qual a classe de orientação, e outro de regressão linear para detectar o nível de engajamento (desconsiderando a orientação). Com treino e teste em diferentes configurações dos dois datasets mencionados acima, os modelos apresentaram bom desempenho.

3.2.2 Features extraídas de dados dos usuários

Considerando atributos coletados do usuário, Barberá et al. (2015) e Garimella e Weber (2017) utilizam a lista de usuários seguidos no Twitter para inferir um índice de polarização política do usuário a partir de uma lista de usuários com orientação política conhecida de esquerda ou direita.

Barberá et al. (2015) utilizam a lógica de modelos de espaços latentes aplicados em redes sociais, onde as características de um indivíduo podem ser consideradas similares às suas conexões. São calculadas as posições dos usuários em um espaço latente ideológico, baseado nas conexões de usuários com uma lista de perfis com ideologia abertamente conhecida, resultando em uma estimativa de posicionamento de indivíduos em um espectro político ideológico.

A técnica proposta em (BARBERá et al., 2015) é utilizada em ferramentas para mensuração da orientação política de usuários envolvidos em política como o GPS Ideológico da Folha de São Paulo (iGPS), calculando a polarização política baseada na estrutura de perfis seguidos dos usuários. A Figura 3.1 mostra a posição de presidentes brasileiros no iGPS, como Lula, Dilma Roussef, Fernando Henrique Cardoso (FHC) e Jair Bolsonaro, além de políticos influentes que apareceram em nosso estudo. Na figura, os políticos com viés de esquerda situam-se na na parte vermelha à esquerda, enquanto os políticos com viés de direita encontram-se na parte verde à direita. Na parte central da figura estão os políticos moderados (neutros), as quais têm seguidores de ambos os lados do espectro e/ou mais moderados.

Figura 3.1: iGPS com Presidentes e influencers políticos

O trabalho proposto em (GARIMELLA; WEBER, 2017), apesar de utilizar o mesmo recurso da listas de usuários seguidos, difere em não necessitar calcular um espaço inteiro para obter um índice de polarização. Sua abordagem calcula dentro da lista de usuários seguidos quantos estão em uma lista de perfis reconhecidos de esquerda e quantos em outra de direita, obtendo uma razão representando o índice de polarização. Quanto mais perfis de um lado forem seguidos, mais próximo de um extremo este índice será localizado.

Neste trabalho, adotamos uma técnica de cálculo da polarização política baseada no trabalho de Garimella e Weber (2017), utilizando listas de políticos de esquerda e direita do iGPS, calculadas com a técnica de Barberá et al. (2015).

3.3 Análise de cenários influenciados por polarização política

Considerando os trabalhos que analisam a influência política em algum contexto, observamos a diversidade de técnicas computacionais para derivar *insights* sobre estudos de caso. Entre os métodos mais usados estão a modelagem de tópicos para compreensão dos assuntos de interesse (JIANG et al., 2020; HAVEY, 2020b), a caracterização de aspectos psicológicos subjacentes a grupos de interesse (DEMSZKY et al., 2019; SLATCHER et al., 2007; PENNYCOOK et al., 2020), ou a caracterização de propriedades demográficas (BOXELL; GENTZKOW; SHAPIRO, 2017). Estes trabalhos são resumidos no restante desta seção.

3.3.1 Análise de Tópicos

A análise de tópicos sumarizando os assuntos envolvidos em discussões que envolvem polarização política é um grande gerador de *insights* sobre o estudo da polarização política, mostrando os principais temas em debate. Um estudo (JIANG et al., 2020) aborda o engajamento de tópicos online relacionados a discussões sobre a COVID-19 em diferentes locais dos Estados Unidos, utilizando clusterização multidimensional de séries temporais para construção dos tópicos. Jiang et al. (2020) concluem que a polarização política apresentada no local geográfico afeta a condução de uma discussão, mostrando correlação com os sentimentos ligados às medidas governamentais. Havey (2020b) analisa o engajamento de usuários segundo sua polarização política em tópicos pré definidos sobre desinformações da COVID-19, e encontra evidências de que usuários conservadores (direita) espalham e acreditam mais neste tipo de assunto.

De uma forma geral, trabalhos envolvendo modelagem de tópicos utilizam abordagens tradicionais como PLSA e LDA, porém nos últimos anos há uma exploração de extensões e novos modelos para melhores resultados de tópicos com textos curtos (LIKHITHA; HARISH; KUMAR, 2019; QIANG et al., 2020). Os experimentos discutidos em (BIANCHI; TERRAGNI; HOVY, 2021) mostram bons resultados na coerência de

tópicos ao agregar *embeddings* contextuais BERT em técnicas de modelagem, mostrando a capacidade das abordagens desta categoria.

Os estudos previamente mencionados falham em analisar somente o contexto geral do tópico, deixando de explorar as ideias principais das discussões. Propomos uma abordagem agregativa, combinando uma técnica de clusterização geral e outra para identificar argumentos mais propagados dentro de um tópico.

3.3.2 Análise de aspectos psicológicos

Dicionários léxicos como o LIWC são utilizados para caracterizar aspectos psicológicos básicos, como estados emocionais e biológicos, estilos de pensamentos, honestidade, diferenças individuais (TAUSCZIK; PENNEBAKER, 2010). Os traços são identificados a partir do uso de palavras de estilo. Tausczik e Pennebaker (2010) resumem vários estudos que relacionam aspectos psicológicos com categorias específicas, tais como o uso de palavras de concordância e da 1ª pessoa do plural que remetem a aspectos de união e coesão em grupos.

No contexto político, Slatcher et al. (2007) utilizam o LIWC para obter os percentuais de uso de determinadas classes e subclasse para construir pontuações para os aspectos de honestidade, feminilidade, depressão, envelhecimento, presidencialidade e complexidade cognitiva. A análise destes aspectos é realizada a partir dos discursos de candidatos presidenciais e vices das eleições presidenciais dos Estados Unidos em 2004, encontrando diferenças entre eles.

O estudo proposto em (DEMSZKY et al., 2019) explora o uso de palavras para caracterizar o posicionamento de pessoas partidárias em relação a tiroteios em massa nos Estados Unidos. O trabalho utiliza palavras emocionais, visto que o afeto influi no compartilhamento de *tweets* políticos, bem como de pronomes pessoais, que refletem a personalização de uma experiência. Demszky et al. (2019) concluem que republicanos concentram-se em comentários sobre o atirador e fatos específicos dos eventos, enquanto democratas nas vítimas e mudanças nas políticas.

Além dos trabalhos envolvendo o contexto político, aspectos psicológicos são explorados em outros fenômenos sociais como violência racial (CHOUDHURY et al., 2016). Choudhury et al. (2016) analisam o movimento Black Lives Matter no Twitter e utilizam as classes e subclasses do LIWC para construir medidas de aspectos de expressão afetiva, estilo linguístico, comportamento, interação interpessoal e estado psicológico

de usuários. São propostas no estudo três questões: mudanças nas características temporais do movimento, como estas medidas são manifestadas geograficamente, e como as medidas nas mídias sociais se relacionam com protestos físicos. Os autores apontam mudanças no engajamento e na linguagem ao longo do tempo, prevalência de negatividade e referência a perdas de vidas em estados com altas taxas de mortes de negros pela polícia, e que a mudança em medidas como o afeto são preditivas de protestos futuros.

Aspectos psicológicos podem também ser extraídos de questionários direcionados, como os resultados relatados em (PENNYCOOK et al., 2020). Este trabalho adotou um conceito de cognição como a interação entre capacidade de compreender probabilidades e números, conhecimentos básicos de ciência, cognição reflexiva, e ceticismo à informações absurdas, e realizou baterias de questões para estes temas. Pennycook et al. (2020) concluem que este conceito de cognição está inversamente relacionado com a ideologia.

Assim como os trabalhos mencionados nessa seção, exploramos a complexidade cognitiva, emoções, e o aspecto de coesão dos grupos, e ainda complementamos a análise com o aspecto envolvendo preocupações pessoais. Mensuramos estas dimensões por meio do dicionário léxico LIWC para ser possível a extração destes aspectos por meio de textos.

3.3.3 Análise da Demografia

A demografia tem sido muito explorada para revelar comportamentos políticos. Por exemplo, Boxell, Gentzkow e Shapiro (2017) avaliam a evolução da polarização política em faixas de idade, mostrando que é maior em pessoas mais velhas (65 ano ou mais) comparando com a faixa entre 18 e 39 anos. Os autores utilizam dados históricos de questionários do American National Election Studies (ANES) para este estudo, atrelando a coleta de dados a pesquisas nacionais de censos, visto que estas tarefas são custosas e não permitem escalar a análise se efetuadas fora de um contexto governamental.

Há trabalhos que analisam a demografia dos usuários em outros contextos a partir de dados de redes sociais, os quais propõem a inferência automática a partir de dados contidos no perfil, tal como gênero e idade a partir das fotos (ELSHERIEF; BELDING; NGUYEN, 2017), gênero a partir de nome de usuário (MUELLER; STUMME, 2016; KNOWLES; CARROLL; DREDZE, 2016), ou localização informada no perfil para compensar o baixo volume de *tweets* geo-referenciados (SAKAKI; OKAZAKI; MATSUO, 2010; LOTAN et al., 2011; HARB; EBELING; BECKER, 2020). Estes trabalhos assu-

mem que o viés de inferências errôneas são diluídos nas vantagens de uma análise de um grande volume de dados. Neste trabalho mensuramos gênero e idade por meio das fotos de perfil, tal como em (ELSHERIEF; BELDING; NGUYEN, 2017; VIKATOS et al., 2017; CHAKRABORTY et al., 2017).

3.4 Frameworks de polarização política

Trabalhos que propõem frameworks e processos ressaltam a importância da análise de polarização política de grupos em um dado contexto através de múltiplas dimensões. Há trabalhos que buscam definir as dimensões para analisar os meios de manifestação de polarização de grupos de pessoas a partir de questionários especificamente projetados. O estudo realizado em (LELKES, 2016) define quatro tipos de manifestações distintas de polarização política (consistência ideológica, divergência ideológica, polarização percebida e polarização afetiva), e foi aplicado sobre dados históricos de questionários do American National Election Studies (ANES), para calcular a evolução das quatro medidas ao longo do tempo. Já o trabalho desenvolvido em (BRAMSON et al., 2016) define nove formas distintas de manifestação da polarização política (propagação, dispersão, cobertura, regionalização, fragmentação da comunidade, distinção de grupos, divergência de grupo, consenso de grupo e paridade de tamanho), propondo medidas para seus cálculos e apresentando famílias de modelos computacionais que podem ser utilizados para analisar os nove tipos de manifestações. Ambos os frameworks se limitam em tamanho e representatividade das amostras, visto sua dependência específica a dados coletados em questionários. Assim, sua aplicação deve considerar o custo e latência da coleta destes dados.

Um framework voltado à análise em dados de redes sociais na visão de instituições políticas, mas podendo ser utilizado para uma visão geral, é proposto por Stieglitz e Dang-Xuan (2013), envolvendo tanto a coleta de dados quanto sua análise. O estudo é planejado visando atender dois propósitos: gerenciamento de reputação de um candidato e planejamento de sua estratégia de campanha/atuação política. O framework é construído focando na coleta de mensagens do Twitter (*tweets*), Facebook (mural) e weblogs (posts) como fontes principais de dados para a análise. A etapa de análise dos dados define três tipos de técnicas a aplicar dependendo do propósito: modelagem de tópicos, análise de sentimentos e análise de estrutura e conexões sociais. Os autores definem um leque de opções para analisar o comportamento em algum contexto nas redes sociais, porém seu

propósito é a utilização individual de uma das dimensões de análise, potencialmente ocasionando uma falha de entendimento sobre o contexto estudado.

3.5 Considerações finais

A Tabela 3.1 lista os trabalhos relacionados com seu objetivo, a fonte de dados utilizada, técnicas empregadas, e as dimensões de análise abordadas, comparando com o framework proposto no presente trabalho. Nossa proposta agrega múltiplas dimensões, incluindo o cálculo de um índice de polarização política combinando as técnicas propostas por Garimella e Weber (2017) e Barberá et al. (2015). Propomos também a análise das dimensões de aspectos psicológicos, demografia, e uma nova abordagem para análise de tópicos. Uma nova dimensão de análise proposta é a exploração da estrutura das redes sociais dos grupos, procurando comunidades destes grafos e obtendo métricas e dinâmicas dos sub grafos onde se encontram políticos de direita e esquerda. Agregamos também no framework proposto neste trabalho uma dimensão referente ao estudo da disseminação de fontes de informação entre os grupos, seja utilizando endereços de redes sociais, seja compartilhando notícias de portais de mídias reconhecidos.

Tabela 3.1: Trabalhos Relacionados

Trabalho	**Objetivo do Trabalho**	**Fonte dos Dados**	**Técnicas e Métodos**	**Dimensão Relacionada**
Preotiuc-Pietro et al. 2017	Classificação da orientação política	Twitter Questionário	Análise de Sentimentos Embeddings	Polarização Política
Garimella and Weber 2017	Classificação da orientação política Análise da polarização histórica	Twitter	Análise de rede Cálculo estatístico	Polarização Política
Conover et al. 2011	Classificação da orientação política	Twitter	Análise de rede	Polarização Política
Barbera et al. 2015	Classificação da orientação política	Twitter	Análise de rede	Polarização Política
Demszky et al. 2019	Análise de Aspectos Psicológicos	Twitter	Análise léxica	Aspectos Psicológicos
Slatcher et al. 2007	Análise de Aspectos Psicológicos	Textos de Discursos	Análise léxica	Aspectos Psicológicos
Pennycook et al. 2020	Análise de Aspectos Psicológicos	Questionário	Análise estatística	Aspectos Psicológicos
Boxell et al. 2017	Análise da Demografia na Internet	Questionário	Análise Estatística	Demografia
Jiang et al. 2020	Análise de Engajamento à COVID	Twitter	Clusterização Multidimensional de séries Temporais	Preocupações
Havey 2020	Análise de Desinformação	Twitter	Análise estatística	Preocupações
Lelkes 2016	Framework	Questionário	Análise estatística	Aspectos Psicológicos e Sociológicos
Bramson et al. 2016	Framework	Questionário	Análise estatística	Estrutura Social
Stieglitz and Dang-Xuan 2013	Framework	Rede Social	Análise de Tópicos Análise de Sentimento Análise de Rede	Preocupações Sentimento Estrutura Social
Este Trabalho	Framework	Twitter	Análise de rede Análise estatística Análise léxica Clusterização Inferência demográfica	Polarização Política Estrutura Social Preocupações Aspectos Psicológicos Fontes de Informação Demografia

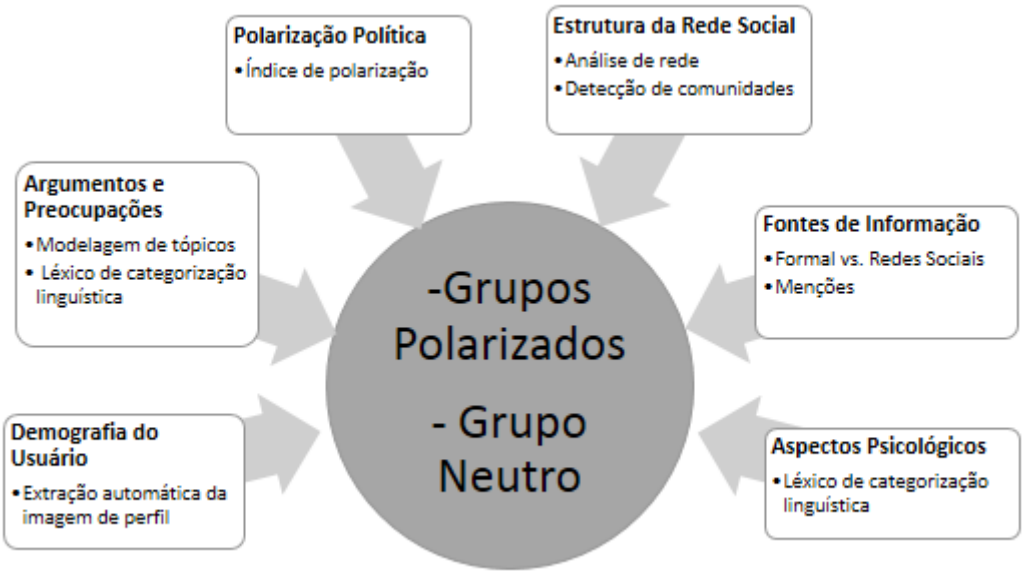
4 FRAMEWORK DE ANÁLISE

Para caracterizar a influência da polarização política no comportamento de grupos elaboramos questões de pesquisa para auxiliar no processo de análise dos cenários a serem estudados. Cada questão de pesquisa guia a incorporação de uma dimensão de análise, e a combinação dessas múltiplas dimensões resulta no framework proposto na Figura 4.1. As propostas para responder as questões de pesquisa são as seguintes:

- Q1) *Índice de Polarização Política*: Para identificar o grau de polarização para esquerda ou direita de usuários de um grupo representando um posicionamento específico, propomos a construção de um índice de polarização a partir de políticos de orientação mais extrema à esquerda ou direita que estes usuários seguem. O iGPS é usado para identificar estes políticos;
- Q2) *Argumentos e Preocupações*: Para analisar os assuntos mais comentados nos grupos e seus principais argumentos, propomos uma abordagem de modelagem de tópicos em duas etapas: começamos utilizando a técnica LDA para formar os tópicos gerais de assuntos, e após identificamos com BERTopic as sentenças similares dentro dos tópicos, mostrando assim os argumentos mais utilizados. Esta abordagem em dois níveis de granularidade provê um nível de interpretabilidade nos argumentos, uma vez que estão inseridos em um contexto específico (tópico) previamente identificado;
- Q3) *Estrutura da Rede Social*: Para analisar se os grupos apresentam estruturas sociais influenciadas por polarização politica, extraímos métricas topológicas das redes dos grupos representando posicionamentos específicos, detecção de comunidades com respectivas métricas topológicas e identificação de usuários mais influentes das comunidades;
- Q4) *Aspectos Psicológicos*: Caracterizamos o estilo linguístico de cada grupo e analisamos se existem diferenças em aspectos psicológicos pré definidos;
- Q5) *Fontes de Informação*: Para investigar a fonte de informação que os grupos mais consomem e propagam, classificamos a origem dos links em categorias e menções intra/entre grupos;
- Q6) *Demografia*: Para comparar o gênero e idade de eleitores de partidos de ambos os lados estimamos a demografia de um grupo influenciado politicamente de forma automática a partir de dados do perfil de seus usuários seguidos;

O restante desse capítulo está organizado conforme a seguir. A Seção 4.1 explica

Figura 4.1: Framework de Análise



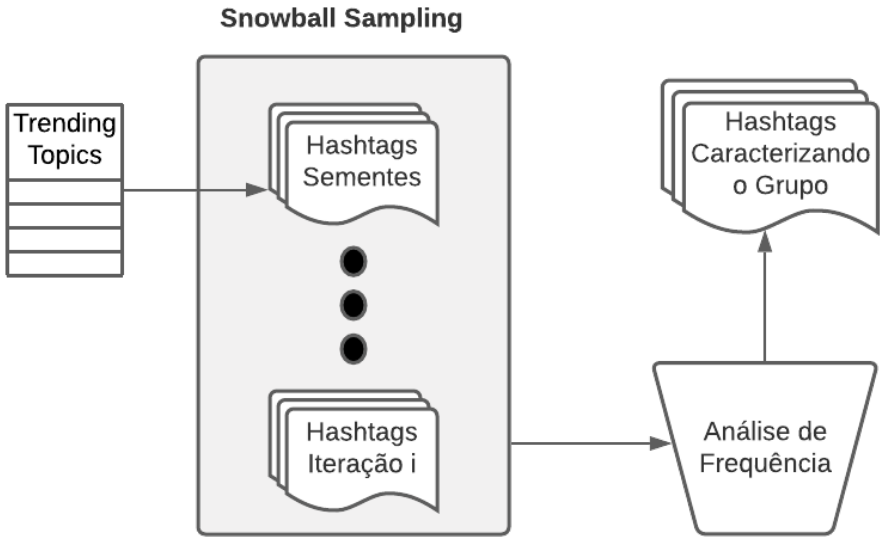
como é estruturada a formação de grupos e coleta de dados. A Seção 4.2 apresenta o cálculo do índice de polarização política (Q1). A Seção 4.3 explica a utilização complementar de duas técnicas de modelagem de tópicos para caracterizar as preocupações dos grupos (Q2). A Seção 4.4 mostra o processo adotado para caracterizar as redes de cada grupo, assim como a seleção das comunidades de interesse (Q3). A Seção 4.5 aponta os aspectos psicológicos e categorias que os compõe (Q4). A Seção 4.6 apresenta o método idealizado para analisar o fluxo de informações (Q5). A Seção 4.7 explica como os dados demográficos são inferidos (Q6).

4.1 Casos de Estudo e Coleta de Dados

Para a aplicação do framework deve-se buscar dados de grupos cuja natureza política se deseje estudar, para então analisar a influência desta polarização no comportamento dos mesmos. Partindo de um cenário específico onde há posicionamentos contrários, busca-se caracterizar dois ou mais grupos com posicionamentos distintos. Além disto, é importante incluir um grupo de controle (neutro) para confirmar um comportamento distinto dos polarizados.

No Twitter, a formação destes grupos pode ser realizada de diversas formas, com ou sem supervisão. Neste trabalho, propomos uma abordagem não supervisionada baseada na coleta de *tweets* com palavras chave que reconhecidamente caracterizem cada

Figura 4.2: Processo de formação de grupos



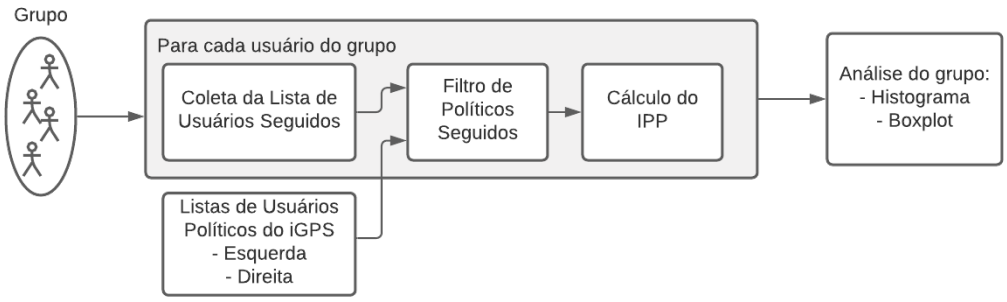
Fonte: O Autor

grupo (*hashtags*), uma técnica amplamente utilizada na literatura para construção de grupos pelas redes sociais (JUNGHERR, 2014; JUNGHERR, 2016). A Figura 4.2 ilustra o processo de construção dos grupos e coleta dos *tweets*. O processo inicia com uma fase de inspeção dos *Trending Topics* nos períodos onde as discussões sobre o assunto supostamente polarizado a se investigar está mais ativo, procurando hashtags relacionadas a favor ou contra. O conjunto de hashtags obtido é o conjunto semente, com o qual se procede a coleta de *tweets* com os termos presentes, e a partir destes é aplicado o processo de *snowball sampling* (amostragem bola de neve). Busca-se manualmente hashtags relacionadas ao assunto pesquisado que formarão mais um conjunto de hashtags para coletar e iterar. Ao final do *snowball sampling*, realiza-se uma análise de frequência para selecionar as hashtags que geraram mais *tweets* a favor e contra um assunto, e estas compõem os conjuntos finais. Para a coleta de um grupo neutro sobre o assunto, são coletados *tweets* com termos que se relacionam ao assunto, sendo excluídos todos aqueles que possuem alguma das hashtags que formam os grupos com posicionamentos a favor ou contra.

A coleta resulta em um conjunto de *tweets* de cada grupo planejado, e então são coletadas a foto e data de criação do perfil de cada autor de *tweet*, e a lista de usuários que segue.

Para melhorar o desempenho da análise textual nos *tweets* são aplicadas ações clássicas de pré-processamento: normalização, remoção de pontuação, caracteres especiais, hashtags, menções e URLs (DENNY; SPIRLING, 2018; AGGARWAL; ZHAI, 2012).

Figura 4.3: Cálculo do IPP



Fonte: O Autor

Também há descarte de *tweets* com menos de três termos, considerando que não há uma construção mínima de sentença a ser analisada.

Para retirar certas formas de promoções artificiais de engajamento em assuntos, removemos perfis de duas formas:

- Uso de software para remoção automática de bots. Neste trabalho utilizamos o API Botometer[1];

- Remoção de contas criadas recentemente. Neste trabalho, definimos usuários criados no máximo há 30 dias antes do início das coletas.

4.2 Índice de Polarização Política

A abordagem utilizada na formação e coleta dos grupos baseia-se em encontrar conjuntos de hashtags com um claro posicionamento de acordo com o grupo planejado. As caracterizações e ações executadas destes grupos podem ser dependentes de sua polarização política, configurando então uma importante informação a ser descoberta. Neste trabalho propomos um Índice de Polarização Política (IPP), uma métrica que combina as abordagens propostas em (BARBERá et al., 2015) e (GARIMELLA; WEBER, 2017), calculando um índice de polarização baseado em políticos de esquerda e direita que são seguidos. A lista de políticos seguidos é obtida através da coleta dos usuários que um usuário segue no Twitter, e então comparada com listas de políticos de direita e esquerda. A Figura 4.3 ilustra este processo de cálculo do IPP dos indivíduos de um determinado grupo.

[1] https://rapidapi.com/OSoMe/api/botometer-pro

A Equação 4.1 expressa o cálculo da métrica do IPP. Para cada usuário da amostra é coletada sua lista de usuários seguidos e calcula-se a proporção entre número de políticos de direita seguidos (PDS) e o total de políticos seguidos, isto é, o somatório de PDS e PES (políticos de esquerda seguidos), sendo estes políticos de direita e esquerda retirados do iGPS (Seção 3.2.2). Para garantir o ajuste do cálculo, soma-se 1 para PDS e 1 para PES, garantindo o correto IPP de neutralidade (50%) caso o usuário não siga políticos, ou caso siga o mesmo número de políticos de cada lado. Quanto mais próximo de 0 for o IPP de um usuário, mais orientado à esquerda é este usuário. Analogamente, quanto mais próximo de 100 for o IPP, maior orientação à direita.

$$IPP = \frac{1 + PDS}{2 + PDS + PES} * 100 \tag{4.1}$$

Para a formação das listas de políticos utilizadas pelo framework utilizamos o iGPS, que situa usuários políticos em um espaço ideológico variando de direita à esquerda conforme o grau de polarização. Selecionamos os 165 políticos do iGPS posicionados mais à direita e os 165 mais à esquerda. Este ponto de corte da quantidade foi definido para que ambos os lados não englobassem políticos situados na zona neutra, onde indivíduos possuem alcance em ambos os lados ideológicos.

Finalmente, considerando o conjunto de todos os usuários de um grupo, utilizamos histogramas e boxplots para avaliar sua distribuição e caracterizar a existência de uma polarização política.

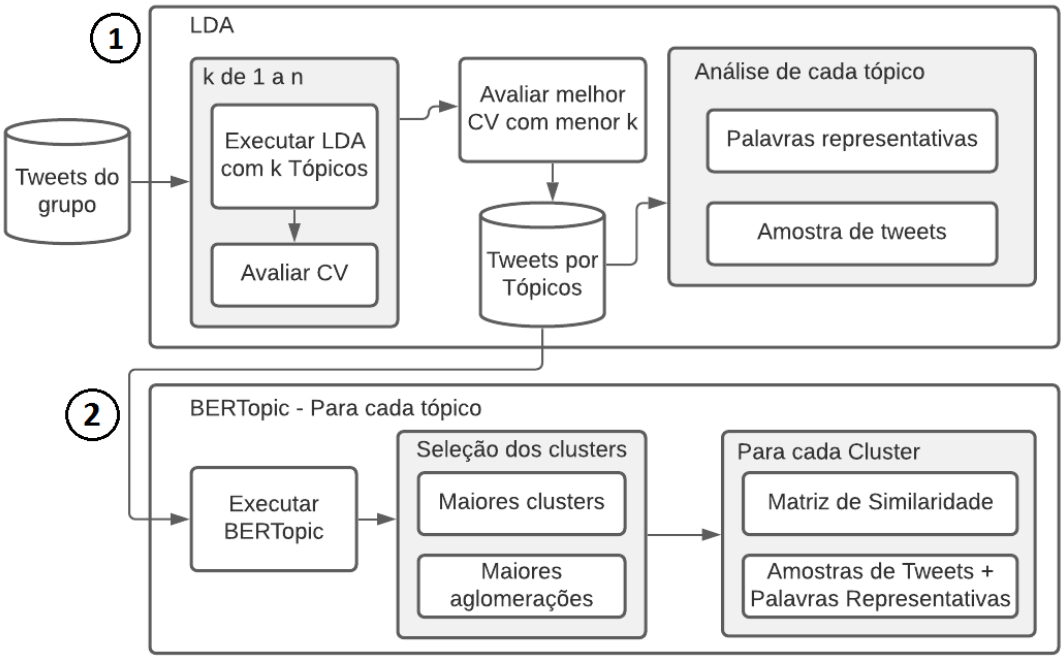
4.3 Modelagem de Tópico

A divisão dos *tweets* de cada grupo por tópicos auxilia na compreensão das preocupações e assuntos em comum, e com isso na caracterização da racionalidade presente nos conjuntos de usuários. Porém, é natural que discussões conduzidas por grupos politicamente opostos tenham um grande volume de *tweets* envolvidos, tornando a mensuração e entendimento dos tópicos presentes uma difícil tarefa. Para gerenciar a complexidade da tarefa de análise, propomos o uso combinado de duas técnicas de modelagem de tópico para formar dois filtros dos tópicos de assuntos presentes nos *tweets*: LDA e BERTopic.

As duas técnicas podem ser vistas como complementares por duas razões:

- **Critérios diferentes para agrupamento:** os diferentes critérios usados para o agrupamento de documentos fornecem subsídios distintos para a interpretação dos

Figura 4.4: Modelagem de Tópicos com LDA e BERTopic



Fonte: O Autor

resultados. Enquanto o LDA é baseado na co-ocorrência de palavras em conjuntos de documentos, estas palavras em isolado necessitam ser interpretadas no contexto do tweet original para a atribuição de seu significado. Como o BERTopic explora *embeddings* contextuais para identificar *tweets* próximos em um espaço vetorial, mesmo com uso de palavras diferentes, é adequado para identificar argumentos semelhantes;

- **Número de agrupamentos encontrados:** enquanto o BERTopic resulta em um número elevado de *clusters* (centenas), o LDA pode ser combinado para um número menor (k), o qual pode ser ajustado por métricas de coerência como o CV. É necessário que a interpretação estabeleça um compromisso entre número de *clusters* e significado.

Propomos então a abordagem combinada de duas técnicas de modelagem de tópicos buscando complementar suas forças e minimizar seus defeitos. A utilização de LDA constrói um filtro inicial de assuntos gerais do conjunto dos *tweets*, enquanto BERTopic constrói os *embeddings* contextuais dos *tweets* de cada tópico e encontra os *clusters* de argumentos similares, apontando as sentenças com significados mais utilizados dentro de cada macro tópico da etapa 1. O método proposto está esboçado na Figura 4.4.

A primeira etapa do processo consiste na utilização do LDA para encontrar um número mínimo de *clusters* coerentes. Para tratar o problema da definição da quantidade k ideal de tópicos propusemos uma forma de estimar baseada na execução do LDA com k variando de 1 a 30. Utilizamos a métrica CV para avaliar as 30 variações de tópicos e conferimos manualmente as configurações com o maiores valores de coerência CV, através dos termos mais representativos de cada tópico. O tamanho do conjunto de tópicos com menor redundância e maior significado após a seleção da etapa anterior é a escolha para o k de cada grupo.

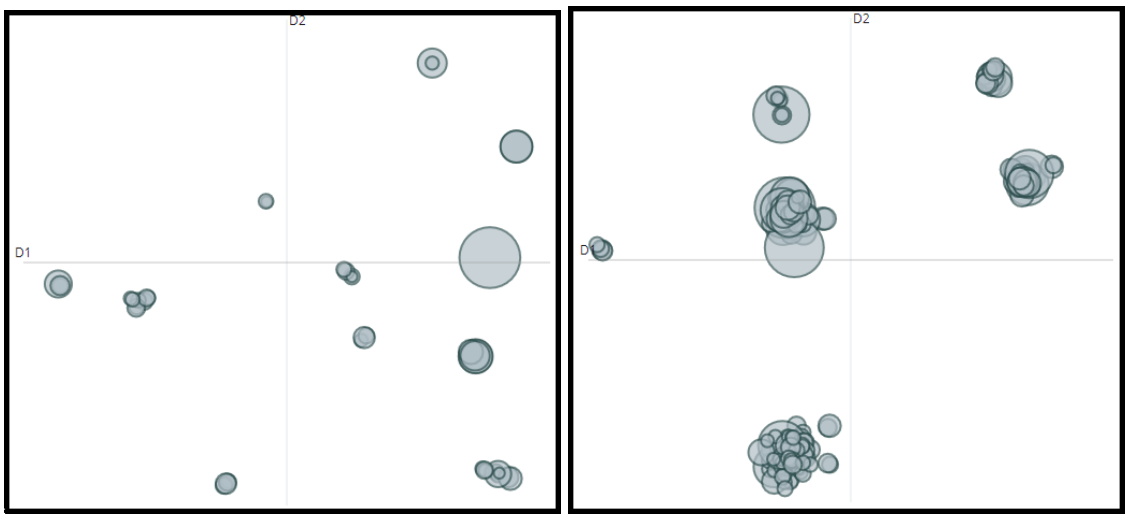
A segunda etapa envolve a aplicação do BERTopic nos macro tópicos encontrados na etapa 1, resultando em um conjunto de *clusters* de *tweets* aproximados por similaridade em um espaço vetorial, ilustrado na Figura 4.5. Para a análise dos principais argumentos utilizados em um tópico há um passo de seleção dos maiores conjuntos semelhantes de *tweets*, podendo ser realizado através de um conjunto de *clusters* menores e próximos no espaço (aglomeração) ou nos *clusters* com mais *tweets*:

- Maiores Aglomerações: utilizando o recurso de visualização da distribuição de *clusters* do BERTopic pode-se localizar manualmente conjuntos de *clusters* que estão próximos a si mas longe em relação aos outros. Pela distância próxima no espaço vetorial, os *clusters* possuem significados também próximos, então se localizam as aglomerações com maior número de *tweets* para serem analisados. A Figura 4.5 (a) ilustra um caso onde é possível aplicar esta abordagem, com o tópico apresentando 42 *clusters* distribuídos em 11 aglomerações, resultando em um conjunto reduzido de *clusters* a analisar por aglomeração;

- Maiores Clusters: a Figura 4.5 (b) ilustra um caso onde há poucas áreas no espaço reunindo argumentos (6), fazendo com que os 165 *clusters* do referido tópico se distribuam nestes poucos centros. A análise destas aglomerações com um número elevado de *clusters* se torna difícil pela grande quantidade de argumentos próximos, mas mesmo assim levemente distantes. Para estes casos, propomos a seleção dos maiores *clusters* para a análise.

Para os casos onde os maiores *clusters* são analisados, plotamos matrizes de similaridade entre estes conjuntos de *tweets* para perceber a diferença ou semelhança entre as principais representações. A matriz de similaridade é um recurso que contribui para a análise neste cenário, visto que há menos *clusters* a analisar, e estes *clusters* apresentam *tweets* de maior similaridade entre si. Para a análise de aglomerações, analisamos

Figura 4.5: Índice de Polarização Política



(a) 42 clusters (b) 165 clusters

analogamente ao LDA, fazendo uso de palavras representativas dos *clusters* e analisando manualmente amostras de *tweets*. Neste trabalho analisamos nos estudos de caso três conjuntos de *tweets* para cada grupo, sejam os três maiores *clusters*, ou três maiores aglomerações.

4.4 Efeitos na Estrutura da Rede Social

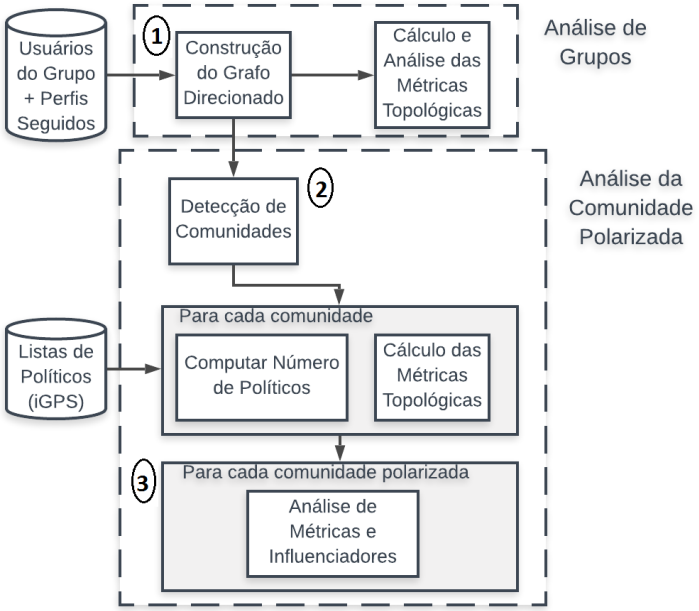
A análise das redes sociais formadas através dos usuários de grupos e suas respectivas listas de usuários seguidos pode trazer descobertas sobre dinâmicas e tipos de usuários influentes para cada público. Propomos como uma das dimensões do framework a análise da estrutura da rede social de cada grupo. O processo proposto para esta análise está representado na Figura 4.6, tendo três passos a executar.

Como passo 1, construímos um grafo direcionado para cada grupo. Os nodos do grafo correspondem aos usuários do grupo e os usuários seguidos por eles, enquanto as arestas direcionadas conectam os nodos de acordo com a lista de usuários seguidos. Calculamos métricas globais para cada rede de grupo para caracterizar sua complexidade:

- Número de nodos, arestas e grau: permite verificar o tamanho e o quão conectada é a estrutura social do grupo;

- Caminho mais curto médio, diâmetro: permite mensurar a extensão da estrutura social do grupo;

Figura 4.6: Análise da Estrutura Social



Fonte: O Autor

- Coeficiente de clusterização: permite observar a tendência do grupo em formar comunidades a partir de conjuntos de usuários com padrões sociais semelhantes.

No passo 2 verificamos a existência de comunidades nas estruturas sociais dos grupos. Para isso, buscamos subgrupos de arestas fortemente conectadas entre si e fracamente com os outros (isto é, comunidades), com o software Gephi. São calculadas para cada comunidade as métricas topológicas: número de nodos e arestas, média do caminho mais curto, diâmetro e o grau médio. Localizamos então as comunidades que contém políticos presentes na lista retirada do GPS Ideológico, descrita na Seção 4.2, em sua estrutura para analisar esta dinâmica naturalmente polarizada.

O passo 3 corresponde à análise das métricas topológicas das comunidades polarizadas. Além das métricas mencionadas, detectamos ainda os usuários centrais das comunidades através das métricas de centralidade:

- *Closeness Centrality*: o quão próximo um nodo está de outros;

- *Betweenness Centrality*: quantifica o número de vezes que um nodo atua como uma ponte para outros pares de nodos;

- Maior *In-degree*: aquele que recebe o maior número de conexões.

4.5 Aspectos Psicológicos Derivados de Características Linguísticas

Pessoas ou grupos apresentam aspectos psicológicos que ajudam a compôr suas identidades, influenciadas por diversos fatores como a classe econômica, grau de escolaridade ou mesmo a polarização política. Os grupos planejados para análise do framework, alinhados com a defesa ou ataque do assunto, formam perfis distintos naturalmente pelo posicionamento, mas também podem ser diferenciados pelos aspectos psicológicos observados dos usuários que os compõe.

A análise de alguns aspectos psicológicos pode trazer caracterizações importantes referentes aos grupos. Propomos para o framework a verificação dos seguintes aspectos:

- **Coesão e união:** este aspecto é utilizado para investigar os grupos com maior senso de união, ou seja, aqueles cujos membros demonstram maior cumplicidade entre seus pares. A união é um aspecto interessante a ser verificado, buscando o grupo que se comporta com maior cooperação e cumplicidade, e fornecendo indícios que um dos lados políticos (ou mesmo diferenciando pela presença de ideologia no grupo) expressa maior apoio entre seus pares. Segundo (TAUSCZIK; PENNEBAKER, 2010), palavras da categoria *we* podem ser utilizadas para promover a interdependência do grupo, e a maior adoção de palavras de concordância (categoria *assent*) pode revelar um maior consenso. Trabalhos relacionados usaram essas categorias de palavras para caracterizar pertencimento e envolvimento (DEMSZKY et al., 2019; CHOUDHURY et al., 2016).

- **Estados Afetivos:** a análise de emoções, tanto positivas quanto negativas, mostra os grupos que expressam mais sentimentos quando inseridos em discussões polarizadas politicamente. A expressão dos estados emocionais constitui um nível semântico relevante para a polarização e pode auxiliar na detecção de níveis ideológicos (DEMSZKY et al., 2019; PREOŢIUC-PIETRO et al., 2017). Um maior percentual de emoções positivas pode evidenciar que um grupo demonstra mais esperança, enquanto que uma maior taxa de emoções negativas pode levar ao entendimento que um grupo mostra-se mais apreensivo ou agressivo na discussão do assunto. Para caracterizar o estado emocional de cada grupo, usamos as categorias de afeto LIWC positivo e negativo (*positive emotions* e *negative emotions*), e as subcategorias negativas *anger*, *anxiety* e *sadness*.

- **Complexidade Cognitiva:** de acordo com (TAUSCZIK; PENNEBAKER, 2010), a

complexidade cognitiva pode ser pensada como uma riqueza de dois componentes de raciocínio: a medida em que alguém diferencia entre múltiplas soluções concorrentes e aquele em que alguém demonstra capacidade de integrar as diferentes soluções. Esses dois processos são capturados pelas categorias de palavras *exclusivity* e *conjunctions*. Também está relacionada a quão sofisticado é o pensamento abstrato ou conceitual de alguém, normalmente associado a uma maior capacidade de discernir entre o conteúdo verdadeiro e o falso. *Prepositions* (e.g., para, com, acima) e *cognitive mechanisms* (e.g., causar, saber, dever) são todos indicativos da capacidade de lidar com uma linguagem mais complexa. Assim como em (SLATCHER et al., 2007; CHOUDHURY et al., 2016), adotamos o uso combinado das classes *exclusivity*, *conjunctions*, *prepositions* e *cognitive mechanisms* para caracterizar os aspectos relacionados à complexidade de cognição revelados por estilo linguístico.

- **Preocupações pessoais:** A mensuração das preocupações pessoais dentro dos grupos pode trazer diferentes óticas de cada um, diferenciando a partir de preocupações alinhadas com a polarização política. Para alinhar os tópicos encontrados automaticamente no corpus com as preocupações dos indivíduos de cada grupo, utilizamos as subcategorias LIWC *Personal concerns*: *work, achievements, money, leisure, home, religiosity, e death*.

Contabilizamos para cada tweet as palavras pertencentes a todas as categorias do LIWC com uma versão para o idioma português do software[2] (inserindo 0 se ausente, 1 se presente). Em seguida, analisamos se havia diferenças significativas no uso proporcional de palavras das categorias do LIWC em geral e aquelas relacionadas aos quatro aspectos psicológicos descritos acima. Usamos o teste Qui-quadrado (alfa = 0,05) para avaliar a significância estatística dessas diferenças.

4.6 Fontes de Informação

Com base nos tipos de fontes utilizadas por seus membros de forma individual, busca-se caracterizar o padrão do grupo. Selecionamos para tal três grandes tipos de fontes de informação:

- Opiniões de outros usuários da rede social, na forma de menções a *tweets* do

[2]http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc

mesmo. Neste tipo de fonte quantificamos menções a usuários de um mesmo grupo e de grupos contrários, visando identificar quando um usuário busca fontes de validação de sua própria câmara de eco e quando busca furar esta bolha;

- Uso de redes sociais, citadas na forma de URL. Através da análise de frequência dos domínios utilizados nas URLs presentes nos *tweets* coletados dos grupos, observamos o uso considerável de três redes sociais: Youtube, Facebook e Instagram. Adicionamos ainda uma categoria "Outros"para considerar diversas redes sociais que não apareciam com grande frequência (e.g. TikTok, Kwai, Telegram, etc);

- Uso de fontes com curadoria jornalística (portais), também na forma de URLs, que referenciam matérias publicadas e validam sua fonte fugindo de informações falsas. Para evitar viés político, utilizamos a lista de sites de notícias indexada em Kadaza[3], um agregador de sites populares em diversas categorias.

Para calcular a proporção de uso do tipo de fonte relativa a *tweets* mencionados, dividimos o dado de interesse pelo número total de *tweets* do grupo que tenham menção a outro *tweet*, assim podemos comparar comportamentos de grupos com número de usuários ou *tweets* diferentes. Por exemplo, se o grupo A possui 40 *tweets* mencionando outros *tweets*, 10 deles mencionam *tweets* do próprio grupo A, então o percentual de uso dos *tweets* do grupo A utilizando uma opinião do próprio grupo é 25%. Para os outros dois tipos de fonte de informação, que se referem a uso de URLs, cada categoria é dividida pelo número total de *tweets* com URL presente.

4.7 Inferência Demográfica

A composição demográfica de uma população ajuda a construir um entendimento de diferenças e semelhanças entre grupos. A realização de censos demográficos nos países buscam insumos para formulação de políticas públicas a partir da imagem construída pelo levantamento, mostrando a importância destes dados nos mais diversos cenários. A coleta de atributos demográficos apresenta duas oportunidades:

- traçar e analisar o perfil demográfico dos grupos politicamente polarizados coletados no Twitter;

[3]https://www.kadaza.com.br/noticias

- comparar os perfis demográficos destes grupos com a composição de eleitores retiradas de pesquisas eleitorais no último pleito.

Os atributos demográficos que propomos extrair automaticamente são gênero e faixa de idade, a partir das imagens de perfil dos usuários coletados. A tarefa de extrair estes atributos das imagens pode ser efetuada de duas formas: via anotação ou automática. O processo de anotação, mais caro e demorado, pode apresentar mais precisão, visto que é realizado por humanos analisando caso a caso. Contudo, a latência e custo deste processo manual são grandes desvantagens. O processo de extração automática, resolve as desvantagens do processo de anotação manual, apresentando melhores vantagens para grandes volumes de dados, sendo mais rápido e convergindo em bons resultados. Trabalhos como (ELSHERIEF; BELDING; NGUYEN, 2017), (WALTER; BECKER, 2018) e (HARB; EBELING; BECKER, 2020) propõe o uso de API Face++[4], uma ferramenta que apresentou acurácia de 85% em experimentos de avaliação (FAN et al., 2014).

4.8 Considerações Finais

Este capítulo apresentou o framework proposto para coleta de dados e análise segundo as dimensões de análise abordadas. Estas foram projetadas para responder as perguntas de pesquisa visando caracterizar a influência da polarização política no comportamento de grupos.

O framework inclui técnicas computacionais para cada dimensão, o que viabiliza a análise em larga escala de grupos a partir de *tweets* a um custo reduzido e com baixa latência. Também não está restrito a questionários específicos, permitindo seu emprego em quaisquer posicionamentos que possam ter viés político. Os capítulos 5 e 6 apresentam a aplicação do framework em 2 estudos de caso no contexto da COVID-19: isolamento social e vacinação, respectivamente, mostrando sua generalidade. Limitações do framework que possam ameaçar a validade do estudo são destacadas no Capítulo 5.

[4]https://www.faceplusplus.com/

5 AMEAÇAS À VALIDADE

Esta seção discute as ameaças à validade do nosso framework. A principal ameaça é a forma como os grupos foram definidos, por esta fase ser essencial e o início de cada análise realizada. O uso de hashtags[1] para coleta automática de grupos nas redes sociais é uma forma amplamente utilizada para esse fim, porém, pode levar a diferentes tipos de bias. Em primeiro lugar, as hashtags podem não representar a população-alvo, pode se mitigar este risco por meio de uma inspeção cuidadosa das hashtags frequentes. Outro risco é que os *tweets* possam estar inseridos no contexto de uma hashtag porque refutam uma ideia representada pela hashtag (falsos positivos). Porém, a quantidade de usuários no caso de falsos positivos é possivelmente pequena dentro do total de usuários que compõem cada grupo e não deve afetar os padrões gerais identificados. Destaca-se também que este processo de coleta de dados não representa a totalidade dos posicionamentos dos grupos, somente uma amostra dos que se dispõe a engajar-se com as hashtags específicas no Twitter. Esta amostra, entretanto, apresenta um subconjunto variado de usuários, considerando a adesão e popularidade do Twitter pelo mundo.

Outra ameaça de validade são os políticos selecionados para representar a polarização política. Mitigamos o risco usando iGPS, que se baseia em um modelo amplamente consolidado (BARBERá et al., 2015). Outra ameaça são os usuários que são politicamente ativos, mas seguem políticos de ambos os lados e, portanto, são considerados neutros. Modelos mais complexos podem ser avaliados no futuro, que consideram, além dos políticos seguidos, o conteúdo textual dos *tweets* como um refinamento (PREOŢIUC-PIETRO et al., 2017).

Sobre o cálculo dos aspectos psicológicos, um usuário pode dar *retweet* manual (isto é, escrever "RT"e copiar o tweet de outra pessoa) de um usuário com o posicionamento oposto ao seu no assunto. Esta prática é utilizada para refutar o *tweet*, porém em nosso estudo faria com que a contabilização dos aspectos psicológicos tenha um ruído do originado do grupo de posicionamento contrário. Ainda sobre os aspectos psicológicos, não há a detecção de ironia ou sarcasmo, ocasionando também um cálculo incorreto para o *tweet* específico.

A estratégia de anotação automática de dados demográficos é outro ponto que ameaça a validade da análise planejada com o framework. Usuários utilizando imagens de perfis de outras pessoas ou mesmo suas antigas inserem dados falso-positivos. Ainda,

[1] https://help.twitter.com/en/using-twitter/how-to-use-hashtags

há um questionamento sobre a precisão da ferramenta para estimativa de determinados biótipos de indivíduos.

A detecção de comunidades apresentada na Seção 4.4 apresenta uma limitação quanto à quantidade de nós e arestas das redes. O software utilizado não realiza a detecção com redes muito grandes, então, realizamos combinações de amostras das redes par a par para ser possível a aplicação da detecção. Este processo resulta em uma representação amostral da rede que se assemelha à imagem geral, mas as comunidades detectadas nesta representações podem não conter figuras de centralidade principais da rede ou mesmo apresentarem métricas topológicas levemente diferentes da rede geral.

Por fim, é de conhecimento geral que a audiência do Twitter pode não representar características da população em geral, principalmente em análises como esta, que representa um enquadramento do público dessa rede social. A população do Twitter é restrita por questões como renda, adesão à internet, adesão ao próprio Twitter, faixa etária, etc.

6 CASO DE ESTUDO: DISTANCIAMENTO SOCIAL

Este capítulo apresenta o estudo de caso sobre o isolamento social, com base no cenário brasileiro observado em março de 2020, período inicial da pandemia no Brasil. Os resultados iniciais foram publicados em (EBELING et al., 2020), os quais motivaram diversas melhorias no framework (EBELING et al., 2020; EBELING et al., 2021).

6.1 Contexto

O surgimento da pandemia deu início a um intenso debate sobre a condução das medidas de combate ao vírus até uma solução permanente. Enquanto o Ministério da Saúde defendeu o isolamento social, uma prática sendo adotada ao redor do mundo, o presidente Jair Bolsonaro e sua base governista apoiaram uma medida de isolamento menos radical, difundindo também o uso de medicamentos sem eficácia comprovada como a cloroquina, alertando que a medida proposta pelo Ministério da Saúde poderia causar consequências na economia. Este dilema entre vidas e economia acabou tendo grande repercussão em um país já dividido politicamente por acontecimentos da última década, e acentuado pela eleição de 2018. Neste contexto, o governo central lançou em 27 de março de 2020 uma grande campanha contra o isolamento social ("O Brasil não pode parar"[1]), apoiada e criticada por muitos brasileiros.

6.2 Coleta de Dados

Para investigar o comportamento polarizado pró e contra isolamento, analisamos características de três grupos ligados a medidas de isolamento social adotadas em razão da pandemia da COVID. Através do método de coleta proposto na Seção 4.1, encontramos no final de março de 2020 representações dos grupos pró e contra isolamento maciçamente em uma hashtag para cada.

- *Cloroquiners*: para representar este grupo, escolhemos a hashtag #OBrasilNãoPodeParar, referente à campanha de mesmo nome lançada pelo governo federal. A campanha argumenta que o isolamento social traz efeitos mais devastadores à po-

[1] https://noticias.uol.com.br/ultimas-noticias/agencia-estado/2020/03/26/planalto-lanca-campanha-o-brasil-nao-pode-parar-contra-medidas-de-isolamento.htm

pulação pelas consequências à economia;

- *Quarenteners*: para capturar uma reação polarizada, identificamos a hashtag #OBrasilTemQuePararBolsonaro[2], que representa oposição à campanha governamental;

- *Neutros*: como os dois grupos acima têm claramente um viés político, procurou-se também um grupo independente destes, representados pelas hashtags #FiqueEmCasa e #FicaEmCasa. Foi observado que o foco maior deste grupo é no isolamento, a *priori* sem polarização política.

A coleta de *tweets* e perfis de usuários envolvidos foi realizada com a API GetOldTweets[3], que possibilita a captação de *tweets* antigos. A coleta ocorreu entre 22 de março de 2020 a 07 de abril de 2020, período em que foi observada uma maior utilização destas hashtags. A Tabela 6.1 mostra, para cada grupo, as respectivas hashtags, o volume de *tweets* coletados e o número de usuários do grupo.

Tabela 6.1: Hashtags e números coletados por grupo

Grupo	**Hashtags**	**Nº Tweets**	**Nº Usuários**
Cloroquiners	#OBrasilNãoPodeParar	74.395	20.572
Quarenteners	#OBrasilTemQuePararBolsonaro	31.060	10.769
Neutros	#FicaEmCasa, #FiqueEmCasa	201.499	102.309

Para investigar a influência dos perfis de bots para impulsionar essas hashtags artificialmente, aplicamos a API Botometer. Dado um perfil do Twitter, esta API analisa as características da conta e retorna uma pontuação de probabilidade relacionada ao comportamento do robô. Devido à grande demanda recebida pela API no momento do desenvolvimento desta pesquisa, restringimos a análise a amostras de usuários selecionados aleatoriamente: 3.750 usuários Cloroquiners (18,2%) e 2.792 usuários Quarenteners (25,9%). Encontramos uma quantidade semelhante de bots em ambas as amostras: 6,46% dos usuários da amostra Cloroquiners e 5,94% dos usuários da amostra Quarenteners.

Considerando as limitações da API, complementarmente analisamos o número de perfis criados em 30 dias ou menos antes das utilizações das hashtags. Essa análise foi motivada pelo fato de que a atualidade de uma conta é um forte indicador de perfis de robôs. Encontramos 5,98% de perfis recentes entre o grupo Cloroquiners e 5,54% entre os Quarenteners. Assim, com base nesses dois critérios, concluímos que não há diferença

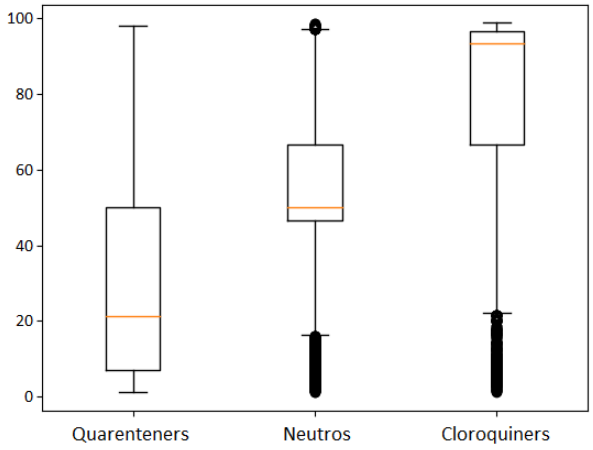
[2]https://getdaytrends.com/trend/%23OBrasilTemQuePararBolsonaro/
[3]https://github.com/Jefferson-Henrique/GetOldTweets-python

significativa no uso (ou uso potencial) de robôs entre os usuários que representam essas posturas. Removemos todos os robôs identificados e contas suspensas de nossa análise.

6.3 Índice de Polarização Política

O boxplot da Figura 6.1 mostra a distribuição da métrica de polarização em cada grupo, calculado como descrito na Seção 4.2.

Figura 6.1: Boxplots de distribuição da polarização de usuários



a) Cloroquiners

Com mediana de 92,3, este grupo concentra metade dos usuários da amostra acima deste percentual e apenas 25% (primeiro quartil - Q1) tem IPP menor que 66,29. Usuários com IPP abaixo de 21,5 são considerados outliers. Claramente este grupo é altamente polarizado à direita. A Figura 6.2 (a) mostra um histograma com a distribuição dos IPPs dos usuários deste grupo.

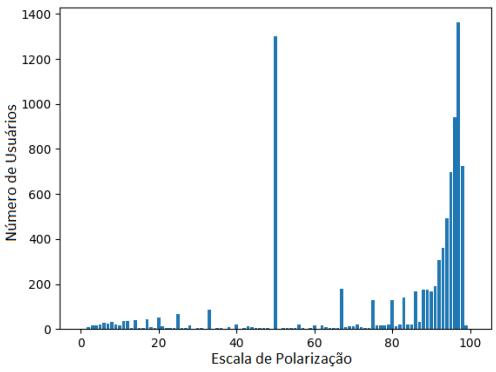
b) Quarenteners

Este grupo tem caráter ideológico mais heterogêneo, com IPP variando entre 0,6 e 97,3 (Min/Max, respectivamente). Dada a amplitude do intervalo, não há *outliers*. Mesmo assim, 75% dos usuários (terceiro quartil - Q3) possui orientação política mais à esquerda (49,9 ou menos), e 50% dos usuários tem índice abaixo de 21 (mediana). A Figura 6.2 (b) mostra o histograma de distribuição dos IPPs dos usuários deste grupo.

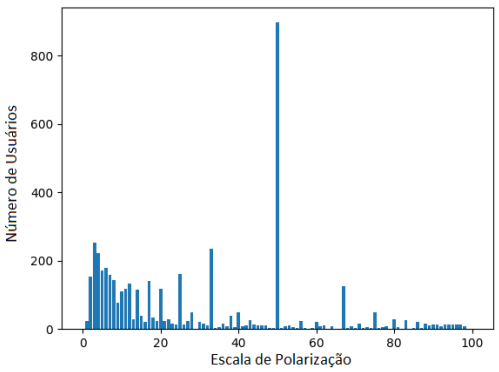
c) Neutros

Este grupo tem uma distribuição similar de usuários em cada um dos lados, com mediana de 50%. Pelo menos 25% (Q1) dos usuários têm IPP igual ou inferior a 46,25, e

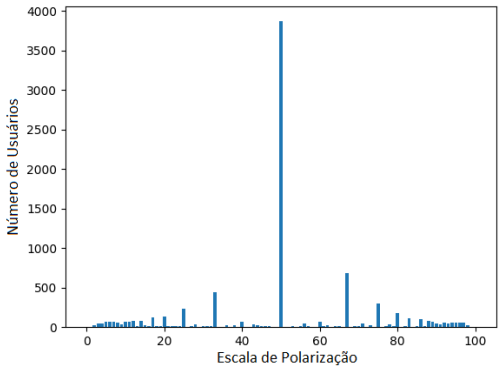
Figura 6.2: Índice de Polarização Política



(a) Cloroquiners



(b) Quarenteners



(c) Neutros

75% dos usuários (Q3) têm valor menor ou igual a 66,3. Os valores min/max são respectivamente 16,4 e 96,5, sendo valores inferiores/superiores a estes considerados *outliers*. A Figura 6.2 (c) mostra o histograma de distribuição dos IPPs deste grupo, observa-se a distribuição de forma proporcional dos Neutros em ambos os lados políticos.

d) Discussão

Conclui-se que o grupo pró-isolamento possui um viés político prevalentemente de esquerda, enquanto o grupo contra isolamento apresenta um viés de direita e mais extremista. Ambos os grupos possuem diversos usuários com IPP neutro (50), enquanto no grupo de Neutros este índice é majoritário para os usuários, com poucos indivíduos se distribuindo de forma espelhada nos lados políticos. Assim, confirma-se que os grupos Cloroquiners e Quarenteners são politicamente polarizados, e o grupo Neutro funciona como controle.

6.4 Assuntos Comentados

Confirmado que os grupos são polarizados, passamos a analisar as principais preocupações de cada grupo de acordo com o método descrito na Seção 4.3.

6.4.1 Análise usando LDA

A Tabela 5.2 mostra os tópicos encontrados, o número de *tweets* e usuários que abordam cada tópico, a densidade (proporção de *tweets* por usuários), e as seis (6) palavras mais representativas de acordo com o peso a ser associado ao tópico. Em seguida, com base nas palavras mais relevantes e na inspeção manual de uma amostra de *tweets* relacionados, conjecturamos sobre a preocupação central de cada tópico.

Tabela 6.2: Tópicos por Grupo

Tópico	**Tweets**	**Usuários**	**Densidade**	**Palavras - Cloroquiners**
0	12855	**5352**	2,40	trabalhar, vamos, povo, quer, dinheiro, trabalho
1	11351	4009	2,83	risco, país, quarentena, mundo, casa, fome
2	**15293**	3984	**3,83**	presidente, bolsonaro, brasil, tudo, pronunciamento, deus
Tópico	**Tweets**	**Usuários**	**Densidade**	**Palavras - Quarenteners**
0	4634	**2159**	2,14	bolsonaro, mortes, governo, saúde, campanha, todos
1	**5105**	1715	**2,97**	genocida, presidente, vírus, urgente, carreata, povo
2	4799	2027	2,36	brasil, parar, agora, vamos, bozo, contra
Tópico	**Tweets**	**Usuários**	**Densidade**	**Palavras - Neutros**
0	8064	2804	2,80	melhor, mãos, sempre, água, cuide, lave
1	10159	2944	3,45	casa, ficar, pode, fica, saúde, ajudar
2	9851	3114	3,16	isolamento, social, ainda, pessoas, vírus, corona
3	9755	**5105**	1,91	brasil, covid, bolsonaro, país, casos, mortes
4	12292	3655	3,36	todos, vamos, vida, deus, passar, amor
5	12229	3148	3,88	quarentena, aqui, fazer, tudo, agora, amigos
6	**13207**	3159	**4,18**	semana, quero, noite, coisas, sair, música
7	7930	4309	1,84	hoje, live, jorge, galera, instagram, parabéns

a) Cloroquiners

Os Cloroquiners estão preocupados com a economia, as consequências econômicas da distância social e a política. O Tópico 0 aborda a necessidade de voltar ao trabalho, usando termos para denotar ações (*vamos, agora*), assuntos (*cidadãos, pessoas*) e motivação (*trabalho, dinheiro, emprego)* . O Tópico 1 compara a distância social no Brasil e no mundo (*país, mundo*) e destaca as consequências econômicas da distância social (*fome, risco*). O Tópico 2 expressa apoio ao presidente, com menções (*presidente*), referências ao slogan de sua campanha[4] (*brasil, deus*), e suas ações (*discurso, verdade*). O Tópico 2 engloba o maior número de *tweets*, com uma média de 3,83 *tweets* por usuário, mos-

[4]https://politica.estadao.com.br/blogs/fausto-macedo/tribunal-extingue-acao-e-mantem-bordao-de-bolsonaro-brasil-acima-de-tudo-deus-acima-de-todos/

trando o envolvimento dos apoiadores do Bolsonaro. O Tópico 0 inclui o maior número de usuários, com uma média de 2,4 *tweets* por usuário.

b) Quarenteners

Os Quarenteners expressam claramente sua oposição à campanha do governo. O Tópico 0 critica a campanha, com menções de alvos (*campanha, governo, bolsonaro*), o motivo da crítica (*morte*) e menções de profissionais de saúde (*saúde*). O Tópico 1 expressa preocupações sobre as ações do presidente e seus apoiadores, com menções diretas (*presidente*), adjetivos pejorativos (*genocídio*), pedidos de mudanças no governo (*urgente*) e críticas para os apoiadores do Bolsonaro (*carreata*). O Tópico 2 enfatiza a distância social como meio de impedir a disseminação do vírus, por meio de ações (*pare, vamos*), apelidos depreciativos (*bozo*) e motivação para distância social (*vida*). O Tópico 1 concentra o maior número de usuários, com média de 2,97 postagens por usuário. Os demais tópicos estão associados a quantidades e densidade de postagens semelhantes (2,14 e 2,36 para os tópicos 0 e 2, respectivamente).

c) Neutros

O grupo Neutros é o mais diverso, discutindo vários aspectos relacionados ao vírus e à distância social. O Tópico 0 trata da lavagem das mãos. Os tópicos 1 e 2 discutem a importância da distância social e os riscos de não adotá-la. O tópico 3 trata de informações sobre pandemias, principalmente no Brasil e ações governamentais. O tópico 4 concentra-se nas mensagens de esperança, positivismo e fé. Os tópicos 5 e 6 tratam das consequências da distância social, como rotinas diárias, monotonia e anseio por atividades ao ar livre. O tópico 7 trata do entretenimento virtual. Os tópicos que abordam explicitamente a distância social (1, 5 e 6) são os que concentram o maior número de postagens, com densidade por usuário de 3,45, 3,88 e 4,18, respectivamente.

d) Discussão

Essa análise inicial evidencia o engajamento político dos Quarenteners e Cloroquiners, visto que os tópicos que apoiam/rejeitam o presidente e suas ações (Tópico 2 para os Cloroquiners e Tópico 1 para os Quarenteners) concentram a maior densidade de postagens por usuário. Os usuários do grupo Neutros não parecem incorporar um viés político ao expressar suas principais preocupações: o impacto prático do isolamento social e do entretenimento virtual.

Tabela 6.3: Clusters de Argumentos dos Tópicos

Grupo	Tópico	#Clusters	#Aglomerações
Cloroquiners	0	127	21
	1	106	18
	2	132	20
Quarenteners	0	55	9
	1	42	11
	2	58	10

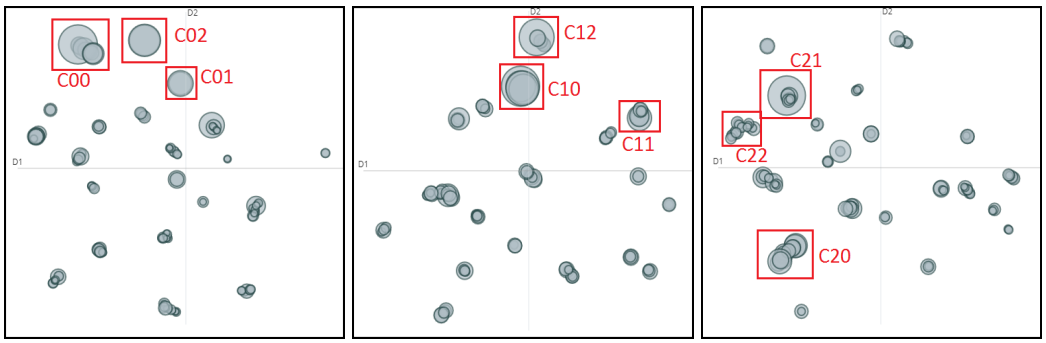
6.4.2 Análise usando BERTopic

A segunda parte de análise de tópicos, utilizando BERTopic, focou apenas nos grupos polarizados. A Tabela 6.3 quantifica para cada tópico LDA dos Cloroquiners/Quarenteners o número de *clusters* encontrados com os parâmetros default do BERTopic. Confirma-se assim que analisar centenas de tópicos para cada grupo é inviável.

Buscamos então dentro de cada tópico LDA os *clusters* com maior número de *tweets*, a fim de buscar os argumentos mais representativos de cada grupo. As Figuras 6.3 e 6.4 apresentam a distribuição dos *clusters* de cada tópico de Cloroquiners e Quarenteners, respectivamente, no espaço vetorial. Com a a análise visual do BERTopic pode-se perceber que os *clusters* encontrados se encontram espalhados pelo espaço, mostrando certa diversidade dos argumentos utilizados. Selecionamos visualmente as áreas onde haviam mais *tweets* próximos, independente do número de *clusters*, para mensurar os principais conjuntos de argumentos mesmo que houvesse alguma pequena dissimilaridade. Na Figura 6.3, por exemplo, a aglomeração C02 é a terceira área com maior número de *tweets* contando um único *cluster*, já a aglomeração C00 possui 11 *clusters* sobrepostos na área, configurando na maior concentração de *tweets*. A Tabela 6.3 apresenta o número de aglomerações por tópico de cada grupo.

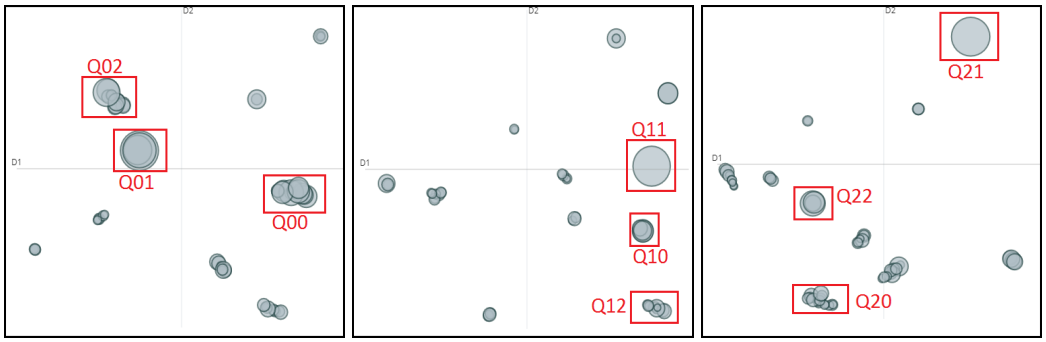
As Tabelas 6.4 e 6.5 sumarizam os resultados para Cloroquiners e Quarenteners, respectivamente, mostrando para cada tópico LDA as aglomerações, o número de *clusters* aglomerados, número de *tweets*, e as palavras mais representativas. Eles também destacam as três maiores aglomerações identificadas para cada tópico. Por convenção, as aglomerações são rotuladas com um acrônimo que identifica o grupo, o tópico e a aglomeração. Por exemplo, a aglomeração C00 denota a aglomeração 0 para o Tópico 0 encontrada para os Cloroquiners. Finalmente, as Tabelas 6.6 e 6.7 apresentam um tweet representativo para cada aglomeração analisada para os Cloroquiners e Quarenteners, respectivamente. No restante desta subseção, detalhamos os argumentos representativos encontrados para cada grupo.

Figura 6.3: BERTopic Clusters - Cloroquiners



(a) Tópico 0 (b) Tópico 1 (c) Tópico 2

Figura 6.4: BERTopic Clusters - Quarenteners



(a) Tópico 0 (b) Tópico 1 (c) Tópico 2

Tabela 6.4: Cloroquiners: Aglomerações mais densas

Tópico	Agl.	#Clusters	#Tweets	Palavras Representativas
0	C00	11	1247	gripezinha, riscos, fome, vagabundos
	C01	2	332	trabalho, contas, dinheiro, famílias
	C02	1	262	Bolsonaro, presidente, parabéns, governo
1	C10	4	702	governadores, imprensa, histeria, higiene
	C11	9	347	colapso, alpinistas, políticos, esquerdistas
	C12	7	326	governadores, prefeitos, ditadores, fascistas
2	C20	17	852	presidente, imprensa, Globo, corruptos
	C21	9	512	Bolsonaro, razão, traidores, dorianos
	C22	19	378	ministros, militares, parabéns, pronunciamento

Tabela 6.5: Quarenteners: Aglomerações mais densas

Topic	Agl.	#Clusters	#Tweets	Palavras Representativas
0	Q00	12	657	economia, vidas, mortes, morrer
	Q01	3	544	bolsonaristas, Trump, hospitais, pandemia
	Q02	12	466	impeachment, semanas, manifestação, quarentena
1	Q10	4	362	presidente, trabalhador, transporte, ricos
	Q11	1	360	bozo, canalha, gado, carreata
	Q12	8	224	casa, vidas, mundo, errado
2	Q20	15	349	monstro, psicopata, presidente, ditador
	Q21	1	305	podemos, juntos, brasileiros, campanha
	Q22	3	275	pare, saúde, assassino, pior

Tabela 6.6: Cloroquiners: Exemplos de argumentos das aglomerações mais densas

Aglomeração	Argumento Exemplo
C00	“Se o Brasil parar, quem não morrer do vírus vai morrer de fome. A vida segue, desligue a televisão e pesquise as estatísticas, comparando a COVID-19 com outras doenças e outras causas de morte no Brasil e no mundo, acordem!
C01	“Mais do que querer, precisamos trabalhar!! As contas já chegaram e o dinheiro acabando. Quarentena vertical já"
C02	“O governo brasileiro, sob o presidente Jair Bolsonaro, está preocupado com milhões de brasileiros que dependem do trabalho para sustentar suas famílias e que agora por causa de uma fanfarra política, causando terror psicológico, decide paralisar"
C10	“O presidente tem razão, apenas em manter os idosos e portadores de doenças crônicas em quarentena. Mantenha boas regras de higiene, evite multidões e proteja seus familiares mais vulneráveis"
C11	“Eles querem nos quebrar, eles querem ver a fome aumentar, mas não vão conseguir"
C12	“E disseram que o governo Bolsonaro seria uma ditadura, só se esqueceram de dizer que os ditadores seriam os governadores!"
C20	“O presidente não passa frio hoje porque tá coberto de razão! E não janta mais por que jantou a imprensa fábrica do fakenews"
C21	“Nosso presidente é maravilhoso! Está realmente com as pessoas! Ele não está escondido no palácio, cercado pela polícia como o covarde de Doriana e outros! Como pode não amar o Bolsonaro?"
C22	“A melhor equipe de ministros que este país já teve! Se não fosse pelo presidente Bolsonaro, nunca teríamos uma equipe assim!"

a) Cloroquiners

A Tabela 6.4 resume as propriedades dos três principais *clusters* de cada tópico, e a Tabela 6.6 apresenta argumentos representativos usados pelos Cloroquiners. O Tópico 0, que apresenta argumentos e motivos para a população retornar ao trabalho, possui a aglomeração com maior número de *tweets* no grupo Cloroquiners, C00, representando 9,7% dos *tweets* neste tópico. A aglomeração C00 engloba 8,6% dos *clusters* encontrados para este tópico, e o argumento central é a preocupação com o impacto do isolamento social na situação econômica (*riscos, fome*). Também minimiza o perigo de COVID para a saúde das pessoas (*resfriado moderado*). A segunda maior aglomeração, C01, é composta por dois *clusters* que argumentam que o retorno ao trabalho é necessário porque a população precisa pagar suas contas (*contas, dinheiro*). Os *tweets* pertencentes à terceira maior aglomeração, C02, elogiam o presidente pelas medidas tomadas para prevenir uma grave crise econômica na pandemia (*parabéns*). Observa-se que a aglomeração C02 representa um único *cluster* e, portanto, uma área muito densa de *tweets*, isto é, muito semelhantes. Essa visão mais detalhada dos argumentos centrais confirma que o Tópico

Tabela 6.7: Quarenteners: Exemplos de argumentos das aglomerações mais densas

Aglomeração	**Argumento Exemplo**
Q00	“Você não vai morrer de fome para parar por 30 ou 60 dias, mas o vírus pode matar milhares nesse tempo. Pare de ser egoísta e escravo do dinheiro"
Q01	“Até Trump, cara. O Brasil gosta muito de esgoto, tanto é que meteu um @#! no Palácio do Planalto"
Q02	“Bolsominions, Você escolheu qual parente entregará ao Coronavirus, simplesmente para agradar ao presidente? Avós? Mãe? Sogra?"
Q10	“Por medo do corona, clientes e comerciantes não arriscarão suas vidas, presidente"
Q11	“Você notou que os imbecis que fizeram a carreata não saíram do carro? Por que não fizeram uma passeata? Vão para a rua, valentões!"
Q12	“Mais uma vez os gado acha que o mundo todo tá errado só o emissário do bem paladino dos bons costumes salvador do Brasil Jair Bolsonaro tá certo"
Q20	“Bolsonaro brinca com a vida dos brasileiros, alguém precisa parar esse homem"
Q21	“Bolsonaro é irresponsável e coloca em risco a vida de milhares de brasileiros para garantir uma narrativa fantasiosa. NÃO dê ouvidos ao presidente, a vida de alguém da sua família pode depender disso!"
Q22	“Ministério da Saúde não foi consultado sobre a campanha criminosa de Bolsonaro contra o isolamento"

0 expressa a necessidade de manter a economia ativa, considerando que as consequências seriam mais prejudiciais à população do que o próprio vírus. O apoio ao presidente está frequentemente embutido nos *tweets* que expressam essas preocupações.

O Tópico 1 destaca os riscos do isolamento social para a economia e compara o Brasil com o resto do mundo. A maior aglomeração (C10) agrupa *tweets* expressando o descontentamento com prefeitos e governadores que adotaram medidas de distância social para combater a pandemia (*governadores*), e com a imprensa que gera pânico excessivo (*imprensa, histeria*), visto que uma boa higiene e isolamento vertical seriam suficientes para controlar o contágio de COVID (*higiene*). A segunda maior aglomeração, C11, critica políticos e esquerdistas (*políticos, esquerdistas*) que querem que o país entre em colapso por causa do desemprego e da miséria (*colapso*), tornando o governo do presidente ainda mais difícil (*escaladores*). Eles também compartilham opiniões de infectologistas que defendem o isolamento vertical, medida defendida pelo presidente. A aglomeração C12 também relata insatisfação com governadores e prefeitos (*governadores, prefeitos*) que estão implementando medidas de isolamento mais rígidas, comparando-os a ditadores que impõem regras que ferem a liberdade da população (*ditadores, fascistas*). Essa análise mais apurada permite entender que a comparação com outros países se refere

ao isolamento vertical, medida de menor impacto econômico adotada por alguns países como a Inglaterra, em contraposição ao isolamento horizontal (social). Os argumentos do Tópico 1 também estão entrelaçados com expressões de apoio ao governo de Bolsonaro e críticas a todos os atores (por exemplo, imprensa, governador, prefeitos) que minam suas tentativas de implementar um modelo mais flexível de combate COVID.

A análise das agregações relacionadas ao Tópico 2 confirma o apoio ao presidente, ao seu governo e às ações que defende. Lembre-se de que este tópico tem o maior número de *tweets* e o maior envolvimento do usuário. O argumento central da maior aglomeração (C20) são as críticas aos opositores do presidente, incluindo a imprensa (*imprensa, Globo*), eleitores, políticos de esquerda ou mesmo opositores de direita. A segunda maior aglomeração, C21, expressa apoio ao presidente e críticas a João Dória nota de rodapé João Dória, governador de São Paulo e possível candidato às eleições presidenciais de 2022. (*traidores, dorianers*), que adotou medidas de isolamento para combater a pandemia no estado de São Paulo. A terceira maior aglomeração, C22, tem como argumento central o apoio ao presidente (*parabéns*), elogiando o próprio presidente, sua escolha de ministros (*ministros*), os pronunciamentos do presidente (*pronunciamento*), e até mesmo taxas de recuperação de casos COVID vinculados a ações governamentais. A análise dos argumentos centrais neste tópico revelou que uma parcela significativa dos *tweets* (8,9%) expressa apoio ao presidente por meio de críticas a uma ampla gama de atores tidos como adversários, desde a imprensa até ex-aliados políticos.

b) Quarenteners

A Tabela 6.5 resume as propriedades dos três principais *clusters* de cada tópico, e a Tabela 6.7 apresenta argumentos de exemplo usados pelos Quarenteners. O Tópico 0 (críticas à campanha governamental) apresenta as maiores aglomerações deste grupo. Q00 é a maior aglomeração neste tópico, com 657 *tweets* distribuídos em 12 *clusters* (21,8% dos *clusters* neste tópico). Esses *tweets* expressam o medo de que o Brasil enfrente uma situação semelhante à da Espanha[5] ou Itália[6], medo de morrer (*morte*, *mortes*) a menos que o isolamento social seja estritamente adotado, e críticas aos empresários que lideram movimentos para manter seus negócios abertos (*economia*). A segunda maior aglomeração (Q01) critica os partidários do presidente (*bolsonaristas*), destaca as crescentes taxas de ocupação nos hospitais (*hospitais*) e destaca que até Donald Trump (*Trump*) reforçou

[5]https://g1.globo.com/bemestar/coronavirus/noticia/2020/04/01/espanha-tem-novo-pico-de-mortes-por-coronavirus-em-um-dia-foram-864-nas-ultimas-24-horas.ghtml

[6]https://www1.folha.uol.com.br/equilibrioesaude/2020/03/por-que-a-italia-tem-mais-mortes-pelo-novo-coronavirus.shtml

a distância social nos Estados Unidos[7]. Essa agregação abrange apenas três (3) *clusters* e, portanto, com argumentos altamente semelhantes. Finalmente, o argumento central da terceira maior aglomeração, Q02, são as consequências de não impor algum nível de distância social para combater COVID (*quarentena, semanas*). Todos esses argumentos fornecem novas percepções sobre as principais razões pelas quais este grupo se opõe à estratégia de Bolsonaro de isolamento vertical como um meio de preservar a economia, expressa com um forte viés político contra Bolsonaro, seus apoiadores e outros atores políticos.

O Tópico 1 tem o maior engajamento no grupo Quarenteners e expressa preocupações relacionadas às ações do presidente. O foco principal da maior aglomeração (Q10) são as críticas aos eleitores do Bolsonaro e as razões pelas quais a população não deve ser prejudicada para manter os negócios abertos (*transporte, trabalhador, rico*). Q11 é a segunda maior aglomeração, representado por um único *cluster* de críticos das passeatas que apoiaram a reabertura de empresas (*carreata, gado*), evento incentivado pelo próprio Bolsonaro. O argumento central da terceira maior aglomeração, Q12, expressa que a aposta de Bolsonaro no isolamento social é inadequada e vai contra o caminho seguido pela maioria dos países do mundo (*mundo, erro*). Essa análise mais detalhada revela que os argumentos neste tópico destacam que vidas não devem ser sacrificadas pela economia, com um julgamento fortemente negativo contra o presidente Bolsonaro, seus apoiadores e seus interesses econômicos.

O Tópico 2 enfatiza a importância do isolamento social. Sua maior aglomeração, o Q20, agrega 15 *clusters* (25,86% deste tópico), o que expressa diversas críticas ao presidente sobre as ações empreendidas em resposta à pandemia. Menções depreciativas ao presidente são as palavras mais representativas neste agregado (*monstro, psicopata, ditador*). Q21, a segunda maior aglomeração, refere-se a um único *cluster* que expressa a necessidade de lutarmos juntos contra esta campanha presidencial (*precisamos, juntos*). A terceira aglomeração estudada (Q22) defende que a saúde deve ser priorizada em relação à economia (*saúde*) e que o isolamento social (*stop*) é a medida adequada a ser tomada (ao contrário da campanha do governo), como endossado pelo próprio ministério da saúde. Como todos os tópicos anteriores, essa análise dos argumentos representativos revela desprezo pelo presidente e por sua campanha.

c) Discussão

Essa análise confirma que as posturas dos Cloroquiners e Quarenteners enfocam o

[7]https://www.nytimes.com/2020/03/29/us/politics/trump-coronavirus-guidelines.html

dilema da economia e saúde, e convive com um forte viés político, com endosso/rejeição ao presidente. A análise dos tópicos de LDA usando o BERTopic revelou que a postura dos Cloroquiners se baseia, por um lado, no argumento de que o vírus não é tão letal e que o isolamento vertical seria uma solução adequada. Eles argumentam que o isolamento vertical minimizaria o impacto na economia e suas consequências na população. Do ponto de vista político, eles apóiam o presidente e suas ações governamentais e criticam uma ampla gama de oponentes, classificando suas preocupações como histeria excessiva. Os Quarenteners, por outro lado, expressam medo quanto ao contágio e suas consequências caso medidas de isolamento social não sejam tomadas. Eles trazem argumentos baseados na experiência mundial no controle da COVID e expressam profundo desprezo pelo presidente.

O resultado da análise, obtendo *insights* forma geral e os principais argumentos que sustentam cada tópico, mostra a capacidade da técnica de modelagem proposta combinando dois modelos.

6.5 Análise da Rede

Avaliamos a estrutura das redes sociais de cada grupo de forma a verificar a influência da polarização utilizando as técnicas propostas na Seção 4.4.

a) Estrutura das Redes

A Tabela 6.8 mostra as métricas topológicas extraídas das redes formadas para cada grupo. É possível observar que a rede dos Quarenteners é a menor pelo número de nodos, porém seu grau médio é o maior, configurando um grupo densamente conectado. O grau médio de Cloroquiners é aproximadamente 1/3 do valor dos Quarenteners, porém a análise do coeficiente de clusterização mostra que Cloroquiners têm maior tendência a formarem comunidades entre seus nodos (28), mostrando ser um grupo com padrões de conectividade similares entre seus nodos, ou seja, distribuem-se de forma mais homogênea em sua rede. O coeficiente de clusterização dos Neutros e o alto número de comunidades mostra que são o grupo social mais disperso e desestruturado.

Tabela 6.8: Propriedades dos Grupos

Grupos	#Nodos	#Arestas	#Com.	Coeficiente de Clusterização	Grau Médio
Cloroquiners	1.316.300	3.913.573	28	0,01859	5,94
Quarenteners	1.145.221	8.541.436	44	0,01362	14,91
Neutros	2.791.367	5.949.930	80	0,00004	4,26

b) Análise das Comunidades

Como pode-se observar pelos coeficientes de clusterização mostrados na Tabela 6.8, a probabilidade dos nodos formarem comunidades nos grafos dos Cloroquiners e os Quarenteners é maior comparada à rede dos Neutros. Comparados a dos Quarenteners, a rede dos Cloroquiners tem menos comunidades.

Examinando as comunidades encontradas para cada grupo, buscou-se aquelas que incluíam políticos, com a premissa que estas seriam mais politicamente influenciadas. Notou-se que nos três grupos existe sempre uma comunidade que concentra estes políticos em igual proporção (direita/esquerda), e que as demais não envolvem políticos, ou em quantidade desprezível.

As propriedades das três comunidades mais polarizadas de cada grupo estão apresentadas na Tabela 6.9. A comunidade polarizada dos Cloroquiners (22,7% destes usuários) é também a maior dentre todas as comunidade polarizadas, com os usuários mais conectados, e maior alcance. No caso dos Quarenteners, encontramos duas comunidades polarizadas. A mais polarizada (19% destes usuários) envolve quantidades significativas de políticos de esquerda/direita, enquanto que uma segunda (12,4%) envolve apenas políticos de esquerda (28), denotando que os usuários politizados deste grupo se engajam de formas diferentes. Comparativamente, a proximidade média dos usuários da comunidade mais polarizada dos Quarenteners e dos Cloroquiners é parecida, mas o alcance da rede é menor se comparados os diâmetros. Para os Neutros, a comunidade polarizada representa 12,2% deste grupo, com usuários mais próximos entre si e menor alcance da informação.

Quando se analisam os nodos de maior influência das redes, nota-se que aqueles com maior número de conexões nos Quarenteners/Cloroquiners correspondem aos adversários políticos nas eleições, enquanto que nos Neutros, a um jornal. Os maiores responsáveis por propagar notícias centralidade de intermediação têm viés político: uma *youtuber* seguidora de Olavo de Carvalho e um político de esquerda. Enquanto os Cloroquiners têm como maior fonte de informação (centralidade de proximidade) um político, nos demais grupos este papel é atribuído a mídias formais.

Tabela 6.9: Propriedades dos Grupos e de sua Comunidade Polarizada

Grupos	#Nodos	#Arestas	Méd. Caminho Mais Curto	Diam.	Grau Médio	Politicos Direita	Politicos Esquerda	Maior In-Degree	Centralidade Intermediação	Centralidade Proximidade
Cloroquiners	299.247 (22,7%)	1443231 (23,3%)	5.35	16	9.64	98	97	jairBolsonaro	ProfPaulaMarisa	BolsonaroSP
Quarenteners	218.066 (19%)	837360 (27,5%)	5.06	14	7.68	94	71	Haddad_ Fernando	50ChicoAlencar	TheInterceptBr
Neutros	339.732 (12,2%)	1233377 (27,5%)	4.68	12	7.26	95	97	g1	nilmoretto	g1

c) Discussão

Pode-se concluir que os grupos dos Cloroquiners e Quarenteners são efetivamente polarizados, sendo a tendência ideológica do primeiro orientada à direita, e do segundo orientada à esquerda, mas de forma menos acentuada e mais diversa. Os padrões de engajamento no Twitter mostram os Cloroquiners como uma comunidade mais fechada e próxima, centrada no repasse de informações via redes sociais, enquanto que os Quarenteners apresentam comportamentos de oposição mais diversos.

6.6 Aspectos Psicológicos

Aplicamos o métodos proposto na Seção 4.5 para verificar as semelhanças e diferenças para o uso das palavras em termos de aspectos psicológicos. Começamos aplicando o teste qui-quadrado nas 64 classes linguísticas do LIWC para comparar as diferenças nos 3 grupos. Todas as classes são significativamente distintas, exceto uma (*exclusividade*, da dimensão de Processos Psicológicos do LIWC).

Em seguida, analisamos cada classe linguística para cada par de grupos. O teste para o par Cloroquiners/Quarenteners não resulta em diferenças para 19 classes; 3 classes no par Quarenteners/Neutros e 2 no par Cloroquiners/Neutros. A semelhança no uso de cerca de 30% das classes do LIWC entre Cloroquiners e Quarenteners mostra uma considerável tendência a construções linguísticas parecidas nos grupos polarizados. Na comparação com os Neutros, esta porcentagem é de 3,12% para os Cloroquiners e 4,68% para os Quarenteners. A Tabela 7.9 apresenta os percentuais de uso das categoria utilizadas para investigar 4 aspectos psicológicos, discutidos a seguir:

- **Coesão**: A classe *nós* é significativamente mais presente nos *tweets* dos Cloroquiners, um indício de que este grupo possui maior coesão. O maior uso percentual da classe *concordância* nos Cloroquiners é significativo, e reitera a ideia de que seus indivíduos possuem maior senso de grupo. Não há diferenças significativas no percentual de uso de *concordância* entre Quarenteners e Neutros. Assim, temos evidências de que os Cloroquiners são um grupo mais marcado pela coesão e união.

- **Emoções**: Enquanto os Quarenteners têm o maior percentual de uso de palavras na categoria *Emoções Negativas* e menor em *Emoções Positivas*, os Neutros têm comportamento oposto (diferenças de 5 e 15,21 pontos percentuais, respectivamente). Cloroquiners seguem a mesma tendência dos Quarenteners, mas em percentual levemente menor. Quando examinadas as subcategorias negativas, Quarenteners ex-

Tabela 6.10: Percentuais de categorias LIWC para cada um dos grupos

Dimensão	Categoria	Neutros	Cloroquiners	Quarenteners
Senso de grupo	we	0.06	**0.08**	0.04
	assent	0.06	**0.09**	0.06
Preocupações Pessoais	work	0.34	**0.41**	0.29
	money	0.25	**0.27**	0.23
	leisure	**0.25**	0.13	0.12
	home	**0.19**	0.06	0.06
	health	**0.18**	0.14	0.13
	death	0.05	0.07	**0.08**
Emoções	anger	0.16	0.20	**0.23**
	sadness	0.21	**0.20**	**0.20**
	anxiety	0.10	**0.12**	**0.12**
	negative emotion	0.39	0.41	**0.44**
	positive emotion	**0.57**	0.51	0.41
Complexidade cognitiva	exclusive	**0.50**	**0.50**	**0.50**
	conjunction	**0.62**	0.54	0.55
	preps	**0.80**	0.65	0.67
	cognitive mechanisms	**0.92**	0.86	0.86

pressam significativamente mais *raiva* que Cloroquiners, e percentuais estatisticamente similares de *ansiedade* e *tristeza*. Estes são indícios de que o grupo dos Neutros é o menos alheio a relatos traumáticos, e que as emoções positivas são confirmadas pelos diferentes tópicos discutidos pelo grupo na Seção 6.4. Nos grupos polarizados, os maiores percentuais de emoções negativas contribuem à ideia da descrição de eventos traumáticos como forma de persuasão à ideia de seu grupo, ou simplesmente estar ligado ao pessimismo das ideias defendidas (condução das ações de controle da pandemia nos Quarenteners[8], e economia do país nos Cloroquiners).

- **Complexidade Cognitiva**: considerando as quatro categorias utilizadas para descrever este aspecto, nota-se não haver diferença significativa no uso da categoria *exclusividade* entre os grupos. O grupo dos Neutros apresenta os maiores percentuais de uso das outras três classes (*mecanismos cognitivos, conjunções e preposições*). Entre Cloroquiners e Quarenteners, apenas a diferença no uso de preposições é estatisticamente significativa. Pode-se concluir que o grupo de Neutros consegue expôr suas ideias em *tweets* com narrativas mais coerentes, complexas e concretas comparadas aos 2 grupos politicamente polarizados, os quais possuem índices semelhantes para Complexidade Cognitiva. Estes achados estão consistentes com

[8]https://daliaresearch.com/blog/democracy-perception-index-2020/

(PENNYCOOK et al., 2020), que relata que a ideologia não está relacionada com as crenças sobre a COVID, e sim à aspectos de cognição.

- **Preocupações Pessoais**: a utilização das classes desta dimensão do LIWC confirmaram os tópicos vistos na Seção 6.4: *trabalho* e *dinheiro* têm maiores percentuais entre Cloroquiners; *lazer*, *saúde* e *casa* em Neutros; e *Morte* mais ligado aos Quarenteners, ainda que em percentuais próximos.

Conclui-se que os grupos polarizados diferem em termos de Preocupações Pessoais e Coesão de grupo, mas mostram-se mais próximos quando comparados com os Neutros nos aspectos envolvendo Emoções e Complexidade Cognitiva. A negatividade fornece evidências que o posicionamento é marcado pelo descontentamento, e que a baixa complexidade cognitiva influencia mais a percepção sobre a pandemia que a orientação política (PENNYCOOK et al., 2020).

6.7 Fontes de Informação

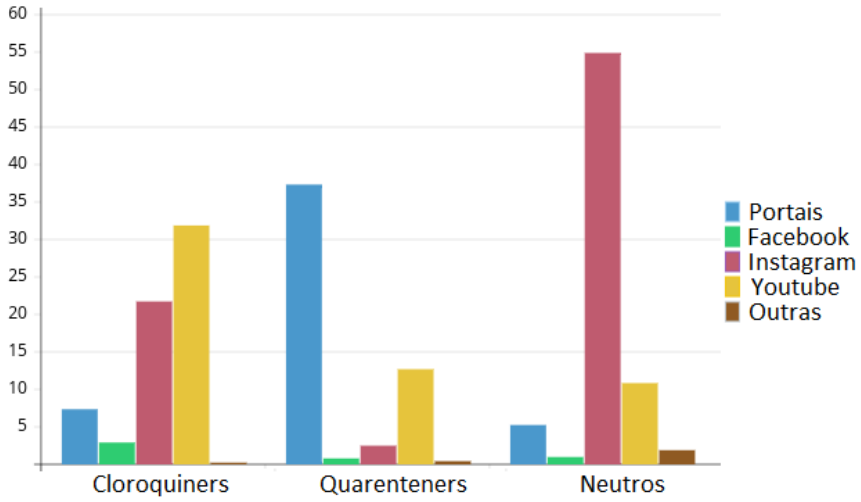
Através do método apresentado na Seção 4.6, calculamos os percentuais de utilização das fontes de informações pelos grupos. As Figuras 6.5 e 6.6 representam respectivamente as proporções de utilização de endereços de sites e menções a *tweets* de outro usuário.

a) Portais de Notícias e Redes Sociais

A Figura 6.5 mostra a distribuição dos endereços web coletados nos *tweets* de cada um dos grupos analisados, divididos em portais de notícias e redes sociais (Facebook, Instagram, Youtube e Outros). Os portais de notícias são a fonte de informação mais referenciada no caso dos Quarenteners (37,24%), denotando que este grupo tem uma preocupação significativa em fundamentar a sua perspectiva em factos reais e nas suas repercussões. Por outro lado, Cloroquiners e Neutros contam com as redes sociais como sua principal fonte de informação. O tipo de link mais frequente de Neutros refere-se ao Instagram (31,78%), seguido do Youtube (10,75%). Uma inspeção de amostra forneceu evidências de que os Neutros usam esses links para conexão com amigos e atividades de lazer/rotina. Para os Cloroquiners, a mídia social mais frequente é o Youtube (31,78%), seguido do Instagram (21,64%). Uma inspeção de amostra revela que esse grupo costuma usar o Youtube para disseminar conteúdo contra o distanciamento social, conteúdo que não é publicado por veículos de notícias regulares e o uso de argumentos de terceiros para

apoiar sua discussão. Surpreendentemente, o uso do Facebook, muito popular no Brasil, não é representativo. A maior taxa de uso (2,8%) está relacionada aos Cloroquiners.

Figura 6.5: Uso de links por grupo



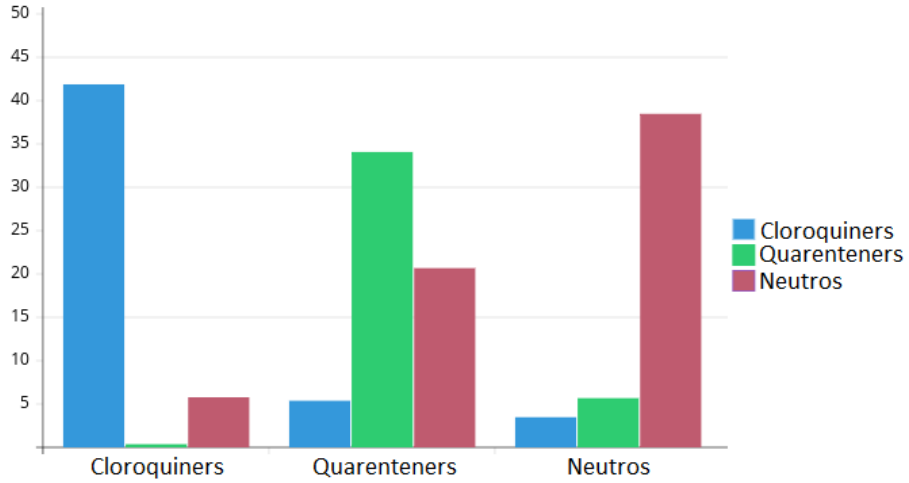
b) Menções

A Figura 6.6 apresenta a proporção de menções a *tweets* escritos por usuários dos grupos, em relação às menções de *tweets* em geral. É nítido que todos os grupos mencionam com mais frequência membros de sua própria comunidade. Os Cloroquiners apresentam o menor percentual de menções aos usuários de outros grupos, com 0,29% de menções aos Quarenteners. O número de menções a Cloroquiners também é pequeno tanto em Quarenteners quanto em Neutros, embora em maior proporção (5,34% e 3,44%, respectivamente). Os Neutros fazem referências comparáveis aos Cloroquiners e Quarenteners. Por fim, os Quarenteners fazem referências significativas aos usuários Neutros. Uma maior comunicação entre os grupos pode denotar uma tendência a uma interação mais aberta ou até mesmo uma maior demanda por debate, mostrando que os Quarenteners procuram discutir mais com outros grupos.

c) Discussão

Por mais que todos os grupos possuam comportamentos de câmara de eco, os Cloroquiners manifestam de forma acentuada, contando com fontes de informação e opiniões que corroboram as crenças do grupo. O uso de mídias sociais e menções a eles próprios revelam que eles estão menos interessados e abertos a outros pontos de vista. De certa forma, os Quarenteners também reproduzem esse comportamento, pois buscam argumentos externos alicerçados em fatos reconhecidos como verdadeiros e os difundem entre si para validar seus argumentos.

Figura 6.6: Menções entre grupos



6.8 Demografia

Com base no método apresentado na Seção 4.7, foi possível realizar a mensuração das informações demográficas a partir de imagens de perfil para 14.480 usuários do movimento de Cloroquiners (70,38%), 7.481 perfis dos Quarenteners (69,46%) e 72.129 usuários dos Neutros (70,5%). As Figuras 6.7 e 6.8 mostram a distribuição dos usuários em termos de idade e gênero para cada grupo, usando as proporções dentro de cada grupo. Em relação à idade, o grupo dos Neutros é o mais jovem, seguido dos Quarenteners. Nestes dois grupos, o percentual de usuários com menos de 30 anos é de 51,13% e 42,25%, respectivamente, enquanto para os Cloroquiners é de 31,66%. A faixa etária de 40-59 anos é muito mais representativa nos Cloroquiners (34,72%), seguida pelos Quarenteners (26,33%). Todos os grupos são semelhantes nas faixas de 30 a 39 anos e acima de 70. Em relação ao gênero, Quarenteners e Cloroquiners têm envolvimento masculino superior (60,6% e 58,1% respectivamente) do que o grupo de Neutros.

Concluímos que o perfil de Neutros pode ser comparado em idade à média dos usuários do Twitter[9], e em gênero, à população brasileira, na qual 51% é feminino[10]. Os Cloroquiners podem ser aproximados em gênero e idade aos dados demográficos dos eleitores de Bolsonaro[11]. No entanto, os Quarenteners não são comparáveis em gênero ou idade aos eleitores de Haddad, o candidato da oposição nas eleições presidenciais de 2018[12].

[9] https://p.widencdn.net/kqy7ii/Digital2019-Report-en

[10] https://brasilemsintese.ibge.gov.br/populacao/distribuicao-da-populacao-por-sexo.html

[11] https://exame.com/brasil/homem-branco-e-conservador-um-perfil-dos-manifestantes-pro-bolsonaro-em-sp/

[12] https://g1.globo.com/politica/eleicoes/2018/eleicao-em-numeros/noticia/2018/10/03/pesquisa-datafolha-veja-perfil-dos-eleitores-de-cada-candidato-a-presidente-por-sexo-idade-escolaridade-renda-

Figura 6.7: Demografia dos Usuários - Gênero

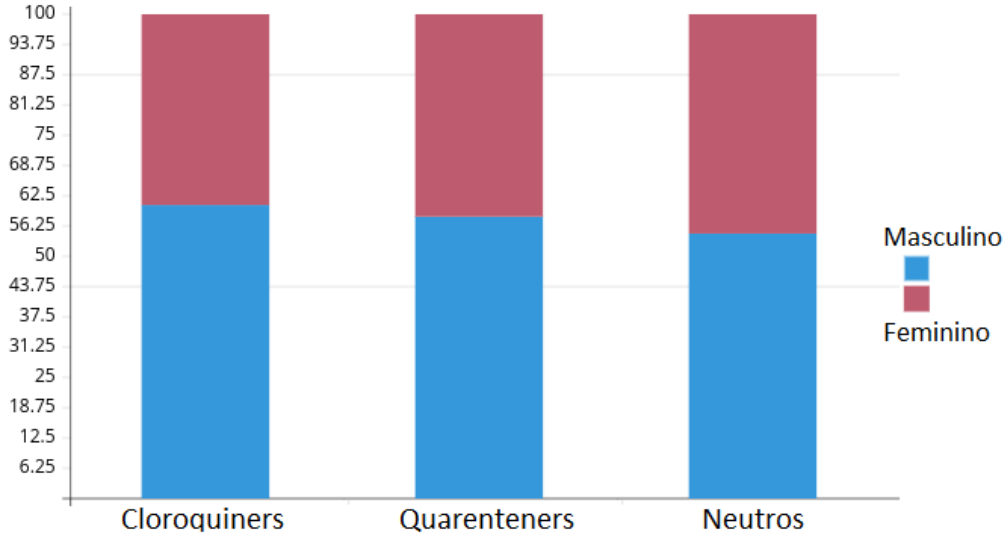
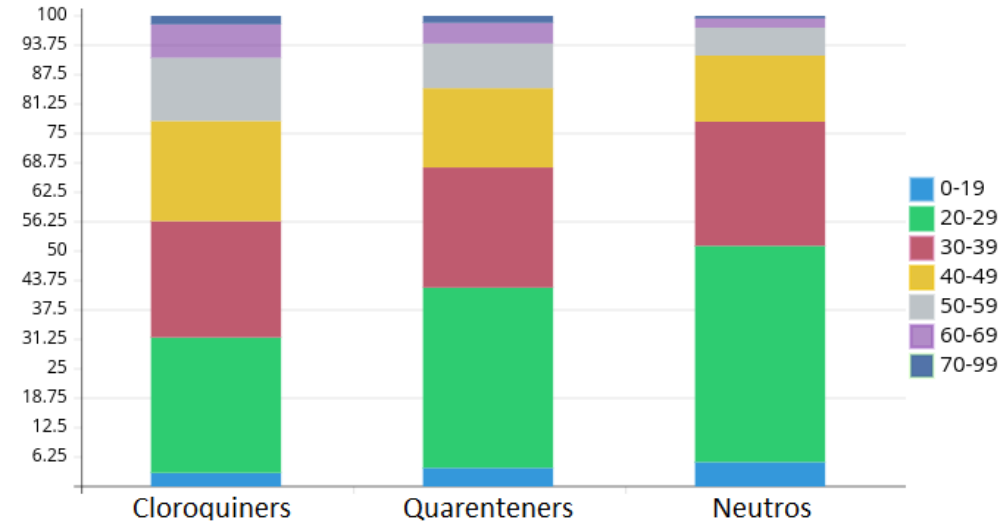


Figura 6.8: Demografia dos Usuários - Idade



6.9 Considerações Finais

A utilização do framework neste estudo de caso trouxe evidências de que as posturas de cada grupo em relação ao distanciamento social são baseadas em orientações políticas, e o comportamento do grupo visa apoiar esse ponto de vista de maneiras distintas. O grupo mais polarizado é representado pelos Cloroquiners, fortemente orientados para a direita e com demografia semelhante aos eleitores do presidente. Quarenteners são polarizados para a esquerda e são ligeiramente mais heterogêneos.

Os temas que os diferenciam refletem a polarização no apoio ou rejeição do presidente no estabelecimento do dilema entre a vida e a economia. Ambos são semelhantes

e-regiao.ghtml

em termos de complexidade cognitiva e emoções negativas, apresentando evidências de que a expressão de seus pontos de vista é devido ao descontentamento. Ambos os grupos polarizados apresentam um comportamento de câmara de eco, porém Cloroquiners de forma mais acentuada, utilizando *tweets* de seu próprio grupo para reforçar e embasar seus argumentos.

7 CASO DE ESTUDO: VACINAS

Este estudo de caso visa demonstrar a generalidade do framework proposto em um segundo contexto polarizado relacionado à COVID-19, a saber, vacinação. Os resultados foram relatados em (EBELING et al., 2022).

7.1 Contexto

No segundo semestre de 2020 começaram a surgir notícias de vacinas em fases de teste após um acelerado processo de confecção, mostrando ao mundo que o término da pandemia poderia estar próximo. No Brasil, presidente, governadores e prefeitos começaram a movimentar-se em negociações com farmacêuticas para aquisição de lotes de vacinas futuras, bem como em direção a medicamentos que supostamente poderiam ser utilizados como tratamento precoce ao vírus. Enquanto o presidente apostou na utilização de um tratamento precoce, governadores tentaram negociações diretas para seus estados com diferentes empresas farmacêuticas para aquisição de vacinas. Dentre o período de negociações ocorreram declarações de ambos os lados sobre a eficácia e risco de cada opção, surgindo uma discussão possivelmente guiada pela polarização política.

João Dória, governador de São Paulo, foi uma das figuras atuantes em relação à negociação das vacinas, idealizando um programa estadual de imunização. Seu governo investiu esforços no desenvolvimento da vacina Coronavac, em parceria com o Instituto Butantan e a farmacêutica chinesa Sinovac. Como possível candidato à eleição presidencial de 2022, Jair Bolsonaro minou repetidamente os esforços de Dória relacionados à imunização, muitas vezes proferindo termos depreciativos e xenófobos (o que resultou no atraso de insumos farmacêuticos chineses para produzir Coronavac[1]). O embate entre João Dória e Jair Bolsonaro tem sido travado principalmente nas redes sociais.

7.2 Coleta de Dados

Para investigar o comportamento polarizado pró e contra vacinação, analisamos características de grupos ligados a estes posicionamentos. Utilizando a abordagem proposta na Seção 4.1, encontramos *hashtags* que descrevem dois grupos com posicionamen-

[1] https://brasil.elpais.com/brasil/2021-05-06/butantan-afirma-que-ataques-de-governo-bolsonaro-a-china-ja-atrapalham-vacinacao.html

tos a favor e contra a vacinação, além de um grupo Neutro. Ainda, durante o processo de análise de frequência das *hashtags* do grupo com posicionamento contra a vacinação notamos a utilização de muitas *hashtags* específicas à vacina Coronavac, então decidimos separá-las como um novo grupo para estudar o comportamento particular deste. Os grupos são os seguinte:

- *Pro-vaxxers*: aqueles que são a favor da vacinação;
- *Anti-vaxxers*: aqueles que são contra a vacinação em geral;
- *Anti-sinovaxxers*: aqueles que são contra a vacina Coronavac;
- *Neutros*: aqueles que comentam sobre vacinas, e que não usam as *hashtags* pró/contra vacinação.

Tabela 7.1: Hashtags e números coletados por grupo

Grupo	**Hashtags**	**Nº Tweets**	**Nº Usuários**
Pro-vaxxers	#EuVouTomarVacina, #VacinaBrasil, #VacinaÉAmorAoPróximo, #VacinaJá, #VacinaNoBrasil, #VacinaParaTodos, #VacinasPelaVida, #VemVacina, #VacinaUrgenteParaTodos	160.867	100.847
Anti-vaxxers	#EuNãoVouTomarVacina, #VacinaNão, #VacinaObrigatóriaNão, #NãoVouTomarVacina	32.876	15.647
Anti-sinovaxxers	#VachinaNão, #VacinaChinesaNão, #VachinaObrigatóriaNão, #VachinaNãoPresidente	17.810	7.067
Neutros	"vacina", "vacinação" (sem *hashtags* dos grupos)	19.558	18.396

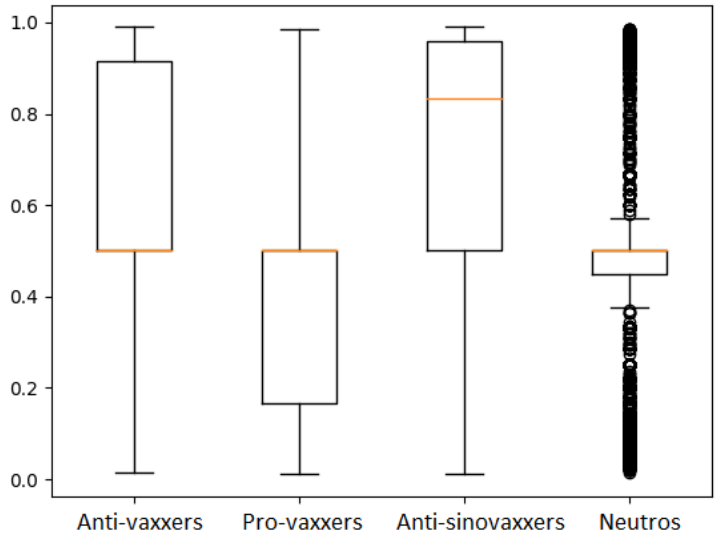
A coleta de *tweets* e perfis de usuários envolvidos foi realizada com a API Snscrape[2], que possibilita a captação de *tweets* antigos. A coleta ocorreu entre 1º de janeiro de 2020 e 1º de abril de 2021, período que cobre a pandemia desde seu início, todas fases de desenvolvimento e aprovação das vacinas, assim como o 1º trimestre de vacinação em 2021. Excluímos bots identificados usando o API Botometer[3], bem como usuários suspensos. Descartamos tweets com menos de três termos. A Tabela 7.1 mostra o volume de *tweets* coletados e o respectivo número de usuários por grupos.

[2]https://github.com/JustAnotherArchivist/snscrape
[3]https://rapidapi.com/OSoMe/api/botometer-pro

7.3 Índice de Polarização Política

A Figura 7.1 apresenta a distribuição dos grupos em um boxplot formado pelos IPPs dos usuários, calculado conforme a Seção 4.2.

Figura 7.1: Boxplots de distribuição da polarização de usuários



a) Pro-vaxxers

O grupo dos Pro-vaxxers possui IPP médio de 37.46, Q3 e mediana em 50, e Q1 em 16.6, mostrando que o IPP médio do grupo está situado na parte esquerda da reta de polarização (alinhado com a esquerda política), e que 75% dos usuários do grupo é visto como neutro ou de esquerda. A Figura 7.2 (a) mostra um histograma com a distribuição dos IPPs dos usuários deste grupo.

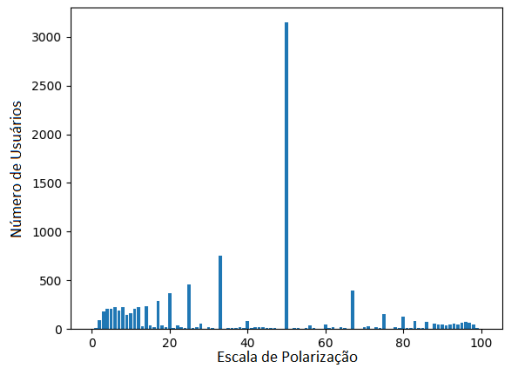
b) Anti-vaxxers

Os grupos de posicionamento contra vacina apresentam certo comportamento espelhado com Pro-vaxxers, com os usuários situados em sua maioria na parte direita da reta de polarização. Anti-vaxxers possui comportamento inversamente proporcional aos Pro-vaxxers: polarização média de 62.41 e 75% dos usuários do grupo neutros (IPP de 50) ou alinhados à direita. A Figura 7.2 (b) mostra o histograma referente à distribuição dos IPPs dos usuários deste grupo.

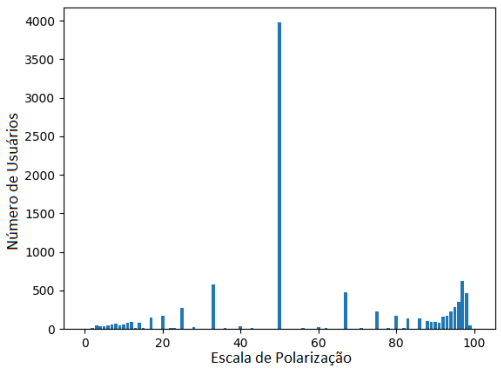
c) Anti-sinovaxxers

O outro grupo de posicionamento contra a vacinação apresenta um nível maior de polarização: seu IPP médio é maior que dos Anti-vaxxers (71.51), e possuem 50% de seus usuários na faixa de polarização entre 82.27 e 98.97. Anti-vaxxers populam a mesma

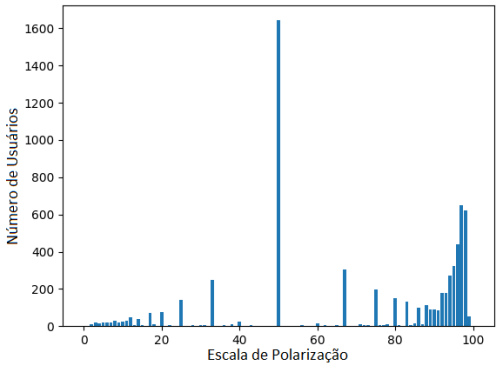
Figura 7.2: Índice de Polarização Política dos grupos



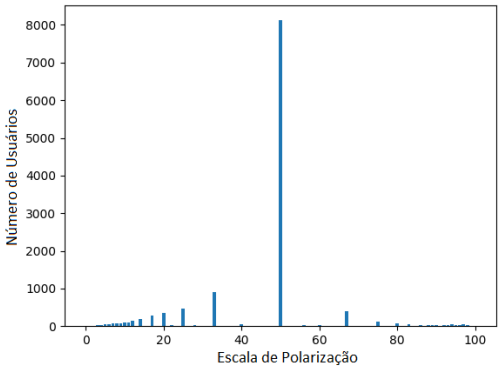
(a) Pro-vaxxers



(b) Anti-vaxxers



(c) Anti-sinnovaxxers



(d) Neutros

quantidade de usuários na faixa de IPP entre 50 e 98.98. A Figura 7.2 (c) apresenta o histograma referente à distribuição dos IPPs dos usuários deste grupo.

d) Neutros

O grupo de Neutros apresenta mediana de 50 no IPP e trata como outliers todos os usuários com IPP abaixo de 45 e acima de 50, atestando a neutralidade política do grupo.

e) Discussão

A partir dos histogramas das Figuras 7.2 pode-se notar que há uma evidente tendência dos grupos anti-vacina concentrarem-se na direita ideológica, e o grupo dos Pro-vaxxers pela esquerda ideológica. Também observa-se a distribuição de forma proporcional dos Neutros em ambos os lados políticos. Comprova-se a partir do cálculo do IPP que as resistências à vacinação não estão ligadas apenas a inseguranças e crenças pessoais sobre sua eficácia, mas têm no Brasil, no caso da COVID-19, um forte componente político. Nos movimentos contrários à vacinação, há uma polarização mais extremista no grupo específico contra a vacina chinesa, podendo apontar um apoio massivo a Bolsonaro e rejeição a João Dória, o qual teria como grande publicidade e impulso para as eleições

de 2022 seu vínculo com a vacina Coronavac. Os Anti-vaxxers possuem polarização à direita também, mas não de forma acentuada como Anti-sinovaxxers, levemente espelhando a distribuição dos Pro-vaxxers.

7.4 Assuntos Comentados

Para compreender os diversos pontos de vista dos grupos analisamos os tópicos de assuntos comentados pelos usuários utilizando LDA, assim como *tweets* que representam a base de argumentos centrais mais utilizados dentro de cada tópico, via BERTopic. A análise foi conduzida de acordo com o método descrito na Seção 4.3.

7.4.1 Análise usando LDA

A Tabela 7.2 mostra a visão geral dos tópicos obtidos com LDA, com percentual de *tweets* e usuários do grupo associados ao mesmo, densidade (*tweets*/usuário), número de tópicos obtidos com BERTopic e palavras mais influentes na associação ao tópico. Cada tópico teve seu sentido investigado através de uma amostra de seus *tweets* e das palavras influentes.

a) Anti-vaxxers

Os usuários do grupo estão preocupados com sua liberdade em não tomar vacina, assim como a definição de uma figura para canalizar seus protestos (João Dória, governador do estado de São Paulo). O Tópico 0 reúne descontentamentos relacionados à obrigatoriedade da vacina, lembrando principalmente que o poder de decisão sobre seu corpo é de cada cidadão. O Tópico 1 levanta questões sobre a segurança de um processo de desenvolvimento da vacina, interpelando ao presidente que não sejam usados como cobaias para algo inseguro. O Tópico 2 expressa o descontentamento dos usuários com medidas vistas como autoritárias. O Tópico 3 expressa a disputa entre apoiadores do presidente e de João Dória, chamando por manifestações da população contra qualquer tipo de obrigatoriedade. Enquanto o Tópico 0 apresenta o maior número de usuários associados (40,6%), o Tópico 3 possui o maior número de *tweets* (32,1%) e densidade (1,84%), mostrando que muitos usuários manifestam seu descontentamento com uma imposição de vacinação, porém os *tweets* criticando a vacina chinesa são realizados de forma mais coordenada.

Tabela 7.2: Tópicos por grupo

Top.	**Tweets**	**Users**	**Dens.**	**Clusters**	**Anti-vaxxers: Palavras**
0	25,9%	**40,6%**	1,34	164	corpo, regras, respeito, queremos
1	21,8%	28,9%	1,58	147	vacina, cobaia, liberdade, Bolsonaro
2	20,2%	25,4%	1,66	133	vamos, ponto, final, ditadores
3	**32,1%**	36,6%	**1,84**	173	brasil, vacina, pessoas, Doria
Top.	**Tweets**	**Users**	**Dens.**	**Clusters**	**Anti-sinovaxxers: Palavras**
0	30,9%	40,1%	1,94	113	doria, ditador, china, queremos
1	29,9%	47,7%	1,58	104	china, vai, isso, brasil
2	**39,1%**	**48,8%**	**2,01**	155	vacina, chinesa, presidente, contra
Top.	**Tweets**	**Users**	**Dens.**	**Clusters**	**Pro-vaxxers: Palavras**
0	18,2%	**18,8%**	1,54	510	brasil, aprovada, pessoas, vacina
1	16,8%	16,9%	1,57	467	pronta, vamos, esperando, jacaré
2	18,8%	15,6%	1,92	485	ciência, bolsonaro, fora, vacinação
3	19,9%	16%	1,98	508	vida, feliz, vacinado, vamos
4	**26,3%**	18,4%	**2,28**	665	coronavac, anvisa, hoje, primeiro
Top.	**Tweets**	**Users**	**Dens.**	**Clusters**	**Neutros: Palavras**
0	**24,2%**	**25%**	**1,03**	534	covid, onde, vamos, queremos
1	22,5%	23,5%	1,01	485	quando, pega, braço, pegar
2	19,8%	20,5%	1,02	454	contra, chinesa, pode, amanhã
3	18,5%	19,4%	1,01	467	queremos, sextou, caminhada, logo
4	14,9%	15,6%	1,01	412	cura, deus, investir, barato

b) Anti-sinovaxxers

O grupo aborda a rivalidade entre Bolsonaro e Dória visando em uma possível disputa política nas eleições presidenciais de 2022. O Tópico 0 ataca João Dória, alegando que suas imposições relacionadas à COVID-19 possuem razões escondidas com a China. O Tópico 1 de forma geral apresenta preocupação com efeitos colaterais de uma vacina originada da China. O Tópico 2 questiona questões de segurança no desenvolvimento da vacina chinesa, se posicionando contra qualquer obrrigatoriedade por este motivo. O Tópico 2 apresenta o maior número de usuários associados (39,1%), maior número de *tweets* (48,8%), e portanto a maior densidade (2,01%), mostrando que seu posicionamento é baseado em preocupações no processo de desenvolvimento da vacina e nas reais intenções de João Dória.

c) Pro-vaxxers

Os assuntos gerais dos usuários do grupo se resumem em ansiedade da imunização, crítica ao governo federal e presidente, e celebração dos resultados de pesquisas de centros científicos brasileiros no desenvolvimento da vacina da COVID. O Tópico 0 apresenta comentários dos marcos iniciais da vacinação no Brasil, parabenizando os institutos Fiocruz e Butantan que iniciaram a produção de vacinas da COVID no Brasil. O Tópico 1 apresenta expectativa pelas fases da vacinação chegarem nas faixas de idade dos usuários,

para então receberem a oportunidade da imunização. O Tópico 2 critica o negacionismo perante vacinação do presidente brasileiro, e atenta a importância da ciência. O Tópico 3 é formado por *tweets* expressando expectativa pela chegada da vacina no Brasil, assim como críticas ao presidente e seus apoiadores. O Tópico 4 critica a falta de ações efetivas do presidente Bolsonaro que contribuiu para o aumento da taxa de mortes da pandemia, mas também há comemoração com o início da vacinação e uma esperança de voltar à rotina pré pandemia. O Tópico 0 possui o maior número de usuários engajados (18,8%), porém o Tópico 4 possui o maior número de *tweets* (26,3%) e densidade (2,28%), apresentando indícios de que os usuários estão em uma maior parte exaltando o início da vacinação, porém o engajamento em si é sobre o fato expectativa da vacinação como uma forma de voltar à rotina pré pandemia.

d) Neutros

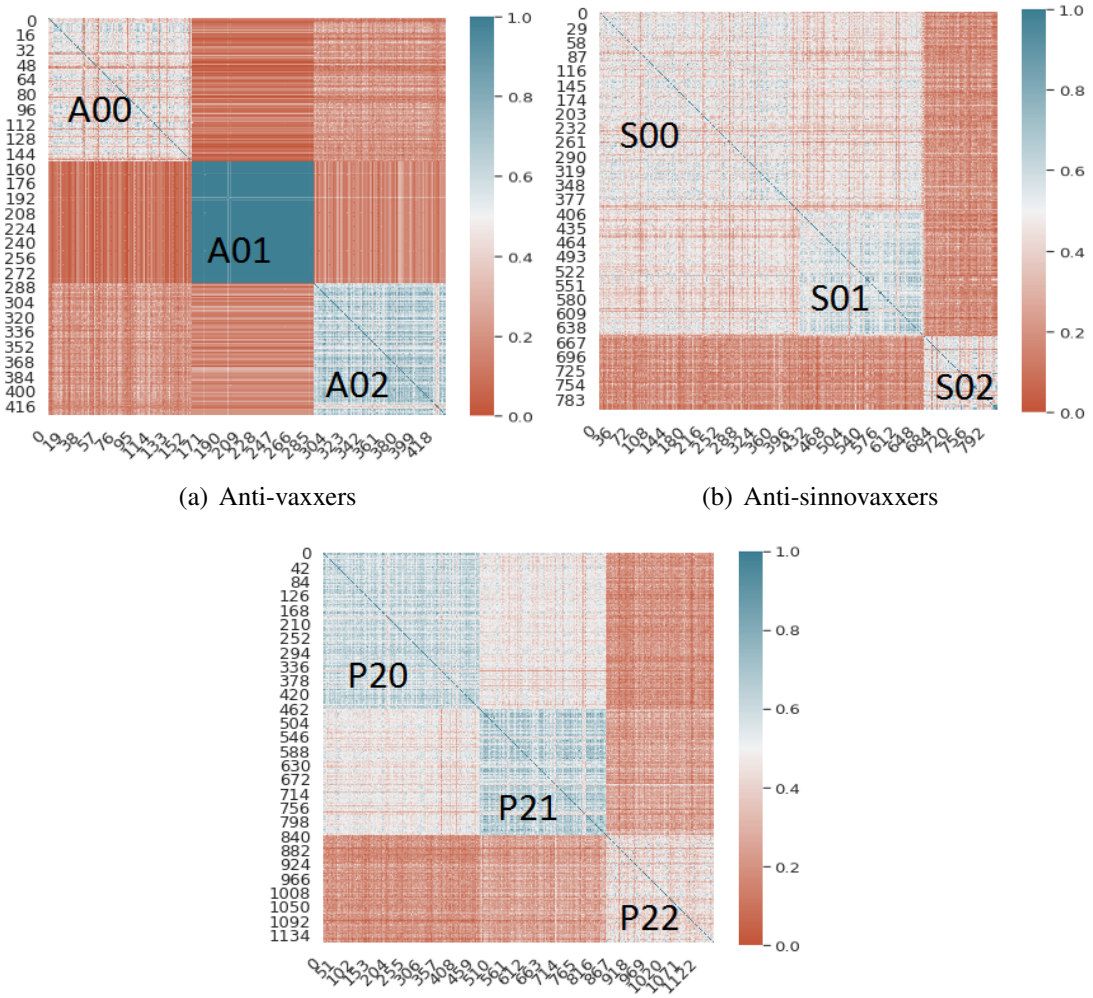
O grupo apresenta assuntos relacionados com a expectativa da vacina e, com ela, a volta da rotina pré pandemia. O Tópico 0 exprime preocupação com a demora da chegada e oferta da vacinação para a população, alertando que esta demora agrava ainda mais a situação no país. O Tópico 1 acompanha o andamento da aplicação de vacinas pelo mundo, comparando com o atraso no Brasil. O Tópico 2 apresenta preocupação com o comportamento descuidado de uma parcela de pessoas em relação à utilização de máscaras e promoção de aglomerações. O Tópico 3 reúne comentários sobre lazer e atividades de celebração a serem realizadas quando a pandemia terminar. O Tópico 4 apresenta preocupações sobre o plano de imunização brasileiro e seus elementos. O Tópico 0 apresenta o maior número de usuários (25%) e *tweets* (24,2%), expressando a expectativa dos usuários do grupo com a chegada da vacinação em massa.

7.4.2 Análise usando BERTopic

Após a obtenção dos tópicos, ocorre a segunda etapa de investigação, agora com BERTopic, buscando os *clusters* que são formados por argumentos similares dentro dos tópicos. As Tabelas 7.3, 7.4, 7.5 e 7.6 apresentam os principais *clusters* de cada tópico dos grupos, com identificação, número de *tweets* associados e um tweet representando o argumento central deste *cluster*, respectivamente para Anti-sinovaxxers, Anti-vaxxers, Pro-vaxxers e Neutros. Para Neutros analisamos somente o maior *cluster* dentro de cada tópico, devido à sua neutralidade no ponto de vista, para os demais grupos são apresentados dados dos três maiores *cluster*. Ainda, para ilustrar a diferença de *clusters* dentro

dos tópicos, a Figura 7.3 representa as matrizes de similaridade de três maiores *clusters* de um tópico de (a) Anti-vaxxers, (b) Anti-sinovaxxers e (c) Pro-vaxxers.

Figura 7.3: Matrizes de Similaridade



(a) Anti-vaxxers

(b) Anti-sinnovaxxers

(c) Pro-vaxxers

a) Anti-vaxxers

A Tabela 7.3 complementa a análise, apresentando os argumentos mais representativos de cada tópico. O Tópico 0, que aborda questões relacionadas à obrigatoriedade da vacina, apresenta como principais argumentos expressões de descontentamento com o STF (A00), o mantra “meu corpo, minhas regras"(A01), e xingamentos para a situação de obrigação da vacina (A02). O Tópico 1 levanta questões sobre a segurança das vacinas produzidas de forma tão rápida, com os principais argumentos sendo as parabenizações ao presidente (A10), organização de protestos contra o STF (A11), e outro *cluster* de argumentos contra a obrigatoriedade da vacina (A12). O Tópico 2 expressa o descontentamento dos usuários sobre as medidas autoritárias que cerceam suas liberdades individuais, porém o maior *cluster* do tópico é um falso positivo, ou seja, um *cluster* de

argumentos a favor da vacinação que se posiciona insinuando que grupos anti-vacina são exemplos de seleção natural. Os outros dois *clusters* investigados no tópico são sobre alertas sobre a ditadura do STF e dos governadores (A21) e pedidos de proteção e bênção ao presidente (A22). O Tópico 3, que apresenta o descontentamento dos apoiadores de Bolsonaro contra João Dória, possui os maiores *clusters* de argumentos sobre a população que não quer ser cobaia para uma vacina de origem chinesa (A30), manifestações contra qualquer obrigatoriedade (A31), e insultos a quem queira obrigar uma vacinação compulsória (A32).

Tabela 7.3: Anti-vaxxers: três maiores *clusters*

Top.	**Clust.**	**#Tw.**	**Argumento Representativo**
0	A00	155	“O Brasil vai mostrar que não somos marionetes do STF"
	A01	131	“Meu corpo minhas regras"
	A02	141	“Vacinação obrigatória vá @!"
1	A10	138	“Sempre com Bolsonaro"
	A11	114	“Brasil inteiro em Brasília"
	A12	91	“Quero ver quem é que vai me obrigar a tomar essa !@"
2	A20	250	“Ótimo, a seleção natural tá aí pra isso"
	A21	215	“Acorda Brasil"
	A22	118	“Elevo a Deus uma prece amém"
3	A30	636	“Não à vacina obrigatória, não seremos cobaias da China"
	A31	281	“Tem de fazer igual o povo de Búzios invadir Brasília"
	A32	253	“Eles que sirvam de cobaias"

b) Anti-sinovaxxers

A Tabela 7.4 apresenta os argumentos mais representativos de cada tópico do grupo. O Tópico 0 reúne suspeições tanto contra a vacina chinesa quanto com João Dória. Os maiores *clusters* de argumentos questionam a atual intenção de João Dória com a negociação da Coronavac (S00), chamadas a protestos contra a vacina chinesa (S01), e elogios a Bolsonaro (S02). O Tópico 1 apresenta preocupação com os possíveis efeitos colaterais da vacina chinesa, e os dois principais *clusters* de argumentos rejeitam a origem da vacina. O terceiro maior *cluster* do tópico (S12), porém, é formado por usuários com posicionamentos contrários do grupo, questionando justamente a rejeição de uma vacina pela sua origem. O Tópico 2 questiona a segurança de uma vacina desenvolvida de forma tão apressada, e uma possível obrigatoriedade de aplicar a mesma. Os maiores *clusters* de argumentos do tópico critica a vacinação obrigatória no estado de São Paulo (S20), que brasileiros não podem ser cobaias (S21), e na extrema celeridade de uma vacina passar por todos os estágios de desenvolvimento (S22).

Tabela 7.4: Anti-sinovaxxers: três maiores *clusters*

Top.	Clust.	#Tw.	Argumento Representativo
0	S00	409	“Ditadória, essa vacina de corrupção não vai colar comigo"
	S01	251	“É só mostrar que não iremos tomar vacina chinesa, temos o poder de pressionar contra"
	S02	147	“Bolsonaro: o melhor presidente da história do Brasil"
1	S10	518	“Jamais tomarei essa vacina do Dória"
	S11	212	“Não dá, China nunquinha"
	S12	88	“Tem que ter muita coragem pra negar uma vacina porque é da China mano, se a vacina vier de Chernobyl eu já estou na fila"
2	S20	228	“Diga não à vacina chinesa que o joão Dória quer obrigar o povo de são paulo tomar vacina obrigatória não"
	S21	154	“O povo brasileiro não é cobaia"
	S22	100	“Vacinas feitas às pressas fica pra turminha de vocês"

c) Pro-vaxxers

A Tabela 7.5 apresenta os argumentos representativos do grupo. Os maiores *clusters* de argumentos do Tópico 0, que é um tópico comemorativo pelas fases da vacina, parabenizam as instituições envolvidas com a vacinação, enquanto o terceiro maior (P02) mostra expectativa com o início da imunização. O Tópico 1 aborda o entusiasmo com os novos estágios da campanha de vacinação contra a COVID, com os dois maiores *clusters* de argumentos celebrando a chegada da vacina no Brasil, e o terceiro maior (P12) apresentando expectativa em receber vacina de qualquer origem. O Tópico 2, que critica o negacionismo perante a ciência, possui o maior *cluster* de argumentos comentando sobre o desejo das pessoas em receber alguma vacina o mais rápido possível (P20), enquanto os próximos dois maiores *clusters* são formados por críticas a Bolsonaro e suas ações sobre a pandemia. O Tópico 3, que apresenta expectativa e felicidade sobre a chegada da vacinação, apresenta o maior *cluster* de argumentos sobre a felicidade da disponibilidade da vacina (P30), enquanto o segundo e terceiro maiores são formados por críticas pesadas ao presidente e seu governo. O Tópico 4, que mistura críticas e confiança, possui os maiores *clusters* de argumentos exprimindo esperança para que o início da vacinação faça acontecer a volta do funcionamento das escolas (P40), expectativa em vacinação para retomar protestos contra o presidente (P41) e chamadas de atenção à uma questionável condução de ações na pandemia (P42).

d) Neutros

A Tabela 7.6 mostra os argumentos representativos dos Neutros. O Tópico 0 questiona a demora para a vacina ser disponibilizada, e o maior *cluster* de argumentos (N00) é alinhado com o assunto geral do tópico. O Tópico 1 acompanha o status da vacinação

Tabela 7.5: Pro-vaxxers: três maiores *clusters*

Top.	Clust.	#Tw.	Argumento Representativo
0	P00	483	“Viva SUS, viva os cientistas, viva Butantan, viva Fiocruz, viva a vacina"
	P01	425	“Butantan a serviço do Brasil, trabalhando para salvar vidas"
	P02	326	“Esperei tanto por esse momento"
1	P10	757	“Brasil, a vacina chegou"
	P11	538	“Feliz demais que alegria vem vacina"
	P12	339	“Já estou preparada para tomar qualquer vacina"
2	P20	467	“Eu vou tomar vacina por mim e por todos os brasileiros"
	P21	374	“Até quando vamos pagar o pato da incompetência bolsonarista? Queremos vacina para todas e todos"
	P22	319	“Revoltante ter um despreparado desse na presidência"
3	P30	534	“Já temos a vacina em solo brasileiro"
	P31	405	“Alguém viu o Bolsonaro por aí? Sumiu ontem depois do anúncio da liberação da vacina"
	P32	321	“Seu lixo, você é uma vergonha para esse país, apoia genocida"
4	P40	969	“Maior desejo do dia: vacina urgente para toda a população para que as escolas voltem a funcionar"
	P41	618	“Quero me vacinar, quero ficar livre da ameaça do coronga pra lutar contra a ameaça Bolsonaro"
	P42	300	“E vai seguindo o plano de impunidade parlamentar e genocídio do Brasil"

pelo mundo comparando com o andamento no Brasil, tendo o maior *cluster* de argumentos comparações com países de programas de vacinação bem definidos (N10). O maior *cluster* de argumentos do Tópico 2, que trata da preocupação dos cuidados das pessoas, é formado por críticas a isso, além de desinformação sobre vacinas e sua eficácia (N20). O Tópico 3, que apresenta comentários sobre lazer, tem seu principal *cluster* de argumentos sobre festas nos finais de semana (N30). O Tópico 4 que reúne comentários sobre o plano de vacinação nacional, tem seu principal *cluster* de argumentos parabenizando o Sistema Único de Saúde (SUS), responsável pela logística, organização e aplicação da vacina.

Tabela 7.6: Neutros: maiores *clusters*

Top.	Clust.	#Tw.	Argumento Representativo
0	N00	356	“Onde está a vacina?"
1	N10	223	“Brasil tem 2.4 milhões de vacinados. EUA vacina 2 milhões POR DIA."
2	N20	339	“A maioria dos grupos de risco, sem máscara e aglomeração, pode sair sem vacina"
3	N30	299	“Estou vivendo ou só esperando pela vacina? SEXTOU"
4	N40	312	“Quando a vacina estiver disponível pra você, SE VACINE!"

e) Discussão

A análise geral e específica dos tópicos e argumentos dos grupos confirma que há uma discussão política em background, onde posicionamentos pró-vacina atacam o presidente e o governo federal, e anti-vacina o elogiam e atacam seus possíveis candidatos rivais na próxima eleição. Neutros apresentam um comportamento pró-vacina mais moderado, onde exaltam o início da vacinação e da ciência, porém não fazem ataques como os Pro-vaxxers. Anti-vaxxers e Anti-sinovaxxers não confiam em vacinas sem um desenvolvimento extenso e seguro, e defendem que a escolha de serem vacinados deva ser uma escolha individual. Anti-vaxxers, entretanto, tentam argumentar seu posicionamento e questionam a obrigatoriedade da vacina. Anti-sinovaxxers alimentam seu posicionamento com questionamentos conspiratórios envolvendo uma futura disputa na eleição presidencial de Bolsonaro com João Dória, explicando o IPP neutro superior ao dos Anti-vaxxers.

7.5 Análise de Rede

Para verificar a influência da polarização nas conexões sociais dos grupos, avaliamos a estrutura de suas redes utilizando as técnicas propostas na Seção 4.4. Para esta análise focamos na comparação entre os grupos surgidos de posicionamentos expressos, retirando o grupo Neutro, para explorar e enfatizar a diferença nos tipos de conexões entre estes grupos, além da análise das métricas destas redes e de suas figuras centrais.

a) Estrutura das Redes

A Tabela 7.7 apresenta as métricas gerais das redes formadas pelos grupos ou amostragens destes, no caso de Anti-vaxxers e Pro-vaxxers.

Apesar dos Pro-vaxxers possuírem o maior número de arestas, possuem muito mais nodos que os demais grupos, resultando no menor grau médio (5.74). Isso explica o maior número de comunidades dentre os três grupos, e o menor coeficiente de clusterização. Já Anti-sinovaxxers apresentam a comunidade mais densa e conectada, com grau médio de 8.31 (contra 6.74 nos Anti-vaxxers) e coeficiente de clusterização significativamente maior. Observa-se que grupos anti-vacina formam redes com comunidades mais coesas e portanto em menor número, comparando com o grupo de Pro-vaxxers.

Uma vez que o software utilizado não suportou o tamanho das redes dos dois maiores grupos, correspondentes aos Anti-vaxxers e Pro-vaxxers, adotamos o seguinte método de amostragem para analisar a estrutura das redes. Para os Anti-vaxxers, dividimos aleatoriamente o grupo em três amostras de usuários e construímos os respectivos grafos três vezes usando pares de amostras. Para cada um dos três grafos, calculamos

todas as métricas e identificamos as comunidades e, em seguida, comparamos os resultados dos três. Os grafos variaram em termos de número de nodos (1,4M-1,55M) e arestas (5M-5,5M), mas eram muito semelhantes em termos de comunidades encontradas (média 44, desvio padrão = 1,73). Em duas amostras, identificamos comunidades polarizadas com propriedades semelhantes, com nodos de centralidade idênticos para as comunidades polarizadas. Para os Pro-vaxxers, aplicamos o mesmo método de amostragem, mas devido à quantidade de dados, dividimos os dados em dez vezes, repetindo o processo cinco vezes. Os grafos variaram em termos de número de nodos (1,8M-2,1M) e arestas (5,4M-5,98M) e eram um tanto semelhantes em termos de comunidades (média 74,6, desvio padrão = 8). No entanto, encontramos comunidades polarizadas semelhantes em 4 grafos, tanto em termos de métricas topológicas quanto de nodos de centralidade. Assim, selecionamos as amostras que produziram os gráficos com as métricas topológicas mais semelhantes, incluindo um número semelhante de políticos de esquerda/direita nas comunidades polarizadas e os mesmos nodos de centralidade.

Tabela 7.7: Propriedades dos Grupos

	#Nodos	**#Arestas**	**Avg. Degree**	**Coef. de Clust.**	**#Comunidades**
Anti-vaxxers	1.558.171	5.252.810	6,74	0,004	42
Anti-sinovaxxers	1.039.047	4.320.372	8,31	0,007	38
Pro-vaxxers	2.027.338	5.803.832	5,74	0,003	63

b) Análise das Comunidades

A busca pelas comunidades polarizadas a partir dos políticos da lista resultou em um padrão diferente do caso de Distanciamento Social. Enquanto no primeiro caso existia uma única comunidade polarizada para cada grupo com políticos de esquerda e direita na mesma quantidade, as comunidades polarizadas deste caso ocorreu em pares para cada grupo: uma contendo majoritariamente políticos de esquerda e outra majoritariamente de direita. A Tabela 7.8 apresenta as métricas das comunidades polarizadas encontradas nos grupos.

Tabela 7.8: Propriedades das Comunidades Polarizadas dos Grupos

	#Nodos	**#Arestas**	**Méd. Caminho Mais Curto**	**Diam.**	**Grau Médio**	**Políticos Direita**	**Políticos Esquerda**	**Maior In-Degree**	**Centralidade Intermediação**	**Centralidade Proximidade**
Anti-vaxxers	229.022 (14,7%)	2.056.847 (39,1%)	3.56	9	17.96	150	1	AbrahamWeint	ananiasfernanda	redpillados
	306.478 (19,7%)	859.406 (16,4%)	4.82	13	5.60	8	142	NetflixBrasil	do_genocida	mfox_us
Anti-sino vaxxers	146.407 (14,1%)	1.004.232 (23,2%)	4.33	10	13.71	154	4	gen_heleno	NiltonGNeto	allanldsantos
	270.009 (26%)	605.235 (14%)	5.31	11	4.48	2	138	NetflixBrasil	thabataganga	_FabioReis
Pro-vaxxers	108.544 (5,3%)	270.546 (4,7%)	4.57	11	4.98	123	1	alexandregarcia	fabiofaria	RafaelFontana
	356.366 (17,6%)	1.510.193 (26%)	4.68	16	8.47	24	120	NetflixBrasil	GuilhermeBoulos	g1

Para os Anti-vaxxers a comunidade polarizada de esquerda é maior em questão

do número de nodos comparado com a orientada à direita (5% a mais de nodos da rede geral), porém possui menor grau médio (comunidade de direita é 3.6 vezes maior) e maior caminho mais curto médio (esquerda é 1.35 vezes maior). Estas métricas denotam a forte conexão dos nodos da comunidade da direita. Para Anti-sinovaxxers ocorre o mesmo comportamento: comunidade da esquerda com maior número de nodos (11.9% de nodos a mais) e maior caminho mais curto médio (1.22 vezes), enquanto comunidade polarizada da direita apresentando maior grau médio (3.06 vezes maior). Já para Pro-vaxxers ocorre um comportamento levemente diferente: enquanto número de nodos e caminho mais curto médio continuam com maiores valores para a comunidade da esquerda, o caminho mais curto médio é maior também nesta comunidade (1.7 vezes maior que a comunidade orientada à direita). Pro-vaxxers apresentam a comunidade de esquerda como a mais conectada. O diâmetros das comunidades orientadas à esquerda é maior que seus pares em cada grupo, denotando que as comunidades de direita possuem seus nodos mais próximos e tornam-se um conjunto mais coeso.

Para cada comunidade polarizada inspecionamos também a força das conexões entre diferentes usuários usando subgrafos que os conectam. Examinamos as conexões dentro de grafos contendo apenas políticos (esquerda e direita) e conexões entre subgrafos contendo usuários regulares conectados a políticos (esquerda e direita). Desta forma, podemos avaliar a influência política na disseminação de informação em cada comunidade.

Para entender a existência das duas comunidades polarizadas dentro de cada grupo, apesar de sua clara orientação direita/esquerda, examinamos sua relação com os políticos seguidos. Nas comunidades de Anti-vaxxers e Anti-sinovaxxers orientados para a direita, uma parte dos políticos de direita estão ligados entre si (44% e 26%, respectivamente). Nas comunidades de esquerda, não há conexão entre os políticos de esquerda. Em outras palavras, esses políticos de esquerda são seguidos por usuários de anti-vacinação, enquanto muitos dos políticos de direita não são apenas seguidos, mas também membros dessas comunidades/grupos. Em ambos os grupos, há cerca de 2% de usuários nas comunidades de direita que estão diretamente ligados a políticos de direita, com um grau médio que é significativamente superior em comparação com o grau médio da respectiva comunidade (34,9 e 53,9 para o Anti-vaxxers/Anti-sinovaxxers, respectivamente). Nessas mesmas comunidades, não há conexão dos usuários com políticos de esquerda. Nas comunidades de Anti-vaxxers e Anti-sinovaxxers voltadas para a esquerda, os políticos não estão conectados de forma alguma (direita ou esquerda). O percentual de usuários conectados a políticos também é muito pequeno, variando de 0 a 0,4%. Assim, concluí-

mos que as comunidades de esquerda nesses grupos anti-vacinação são compostas por usuários motivados a refutar ideias de outros grupos, incluindo políticos de esquerda. As conclusões são compatíveis com os argumentos encontrados em S12 e A20, *clusters* de argumentos que na verdade refutam a postura anti-vacinação.

As comunidades dos Pro-vaxxers apresentam tendências semelhantes, mas com intensidades diferentes, pois nem todos os apoiadores de direita são contra a vacinação. Considerando a comunidade de esquerda, os políticos de ambos os lados estão ligados entre si em proporção significativa (90% e 29% para esquerda/direita, respectivamente). Há 0,5% de usuários conectados a políticos de direita e 2,2% de usuários conectados a políticos de esquerda (graus médios 3,8 e 10,7, respectivamente). Assim, a conexão direta com políticos de esquerda é ligeiramente superior em comparação com a média da comunidade. Na comunidade de direita, parte significativa dos políticos de direita estão ligados entre si (64%), e 2,4% dos usuários estão ligados a eles, com um grau médio de 9,8. As conexões com políticos de esquerda são insignificantes. Há duas explicações possíveis para a presença de usuários de direita e políticos nos debates pró-vacina: confiança na ciência contra ideologias[4], ou uma mudança na agenda do Bolsonaro para ser visto como o responsável pela vacinação apesar de sua postura no passado[5], para fins eleitorais.

Observando as comunidades polarizadas de cada grupo encontram-se padrões alinhados com cada ideologia. Para grupos de posicionamento anti-vacina as comunidades polarizadas de direita apresentam uma porção de políticos de direita conectados entre si e com usuários de direita, enquanto as comunidades de esquerda não mostram conexões dos políticos desta ideologia. Pode-se dizer que nestas comunidades enquanto os usuários de direita se inserem como membros nas comunidades polarizadas de direita, os usuários de esquerda se inserem nas comunidades polarizadas de esquerda com motivação de refutar o que os grupos anti-vacina pregam. Para as comunidades polarizadas de Pro-vaxxers ocorre um padrão similar na comunidade de esquerda, porém com menor intensidade para usuários de esquerda e pouco menor em direita. A presença de usuários de direita conectados com políticos de esquerda na comunidade polarizada de esquerda atenta ao fato de que mesmo usuários de direita estão inseridos nos grupos que defendem a vacina, mesmo com a divergência na ideologia média dos usuários de cada grupo.

Por fim, analisamos os principais influenciadores das comunidades polarizadas, representados pelos nodos com os maiores valores em in-degree e centralidade. Em geral,

[4]https://www.cnnbrasil.com.br/nacional/2021/01/23/ datafolha-79-dos-brasileiros-querem-se-vacinar-contra-o-coronavirus

[5]https://hora.com/5946401/brazil-covid-19-vacines-bolsonaro/

os políticos são muito influentes nas comunidades de direita, enquanto que nas de esquerda encontramos meios de comunicação e ativistas científicos. Os usuários com maior número de seguidores (in-degree) em comunidades de direita são políticos (Abraham Weintraub - ex-Ministro da Educação e General Heleno - Ministro chefe do Gabinete de Segurança Institucional) e um jornalista notoriamente de direita (Alexandre Garcia). Para as comunidades polarizadas de esquerda, em todos os grupos, a plataforma de *streaming* Netflix é a mais conectada (maior in-degree), ou seja, os indivíduos destas comunidades estão nas redes sociais não prioritariamente para politizar, mas estão de uma forma natural.

O nodo com maior centralidade de intermediação, ou seja, aquele responsável por espalhar as informações pela rede, é representado nas comunidades polarizadas da direita por dois apoiadores de Bolsonaro (ananiasfernanda and NiltonGNeto) e pelo Ministro das Comunicações (Fabio Faria). Este papel nas comunidades polarizadas da esquerda é representado por um ativista contra o presidente (do_genocida), uma Youtuber cientista (thabataganga), e um político (GuilhermeBoulos). Já o nodo com maior centralidade de proximidade, responsável por ser o nodo central da rede, é representado nas comunidades da direita por usuários também alinhados à direita: um perfil que divulga apps ditos de direita (redpillados) e dois jornalistas de direita (allanldsantos e RafaelFontana). No momento da escrita deste trabalho, o perfil allanldsantos está suspenso do Twitter por uma demanda legal[6], enquanto um segundo perfil do usuário (allannoexilio) está também suspenso por violação de regras da plataforma[7]. Para as comunidades polarizadas de esquerda este nodo é representado por usuários relacionados com a imprensa: um mídia freelancer (mfox_us), um jornalista especializado na área médica (_FabioReis) e um portal de notícias (g1).

c) Discussão

Encontramos indícios de que a polarização política afeta a estrutura das redes sociais no tocante ao posicionamento sobre vacinação. Os grupos possuem suas comunidades polarizadas alinhadas com a ideologia, onde comunidades de direita de Anti-vaxxers e Anti-sinovaxxers e a comunidade de esquerda dos Pro-vaxxers são maiores e mais densamente conectadas comparadas a seus pares. Nos grupos anti-vacina os usuários de comunidades polarizadas à direita agem com um forte padrão de câmara de eco, enquanto

[6]https://www.terra.com.br/noticias/brasil/politica/twitter-encerra-conta-de-allan-dos-santos-a-mando-da-justica,cb3ad1699e8ebadd1a342a92bc3fc1cex9cp4b5u.html

[7]https://politica.estadao.com.br/noticias/geral,twitter-remove-mais-uma-conta-usada-pelo-bolsonarista-allan-dos-santos,70003890301

a comunidade de esquerda apresenta maior pré-disposição a discutir e rebater ideias fora de sua bolha. Pro-vaxxers possuem um comportamento espelhado, porém com uma estrutura de rede mais heterogênea e aberta. Há uma forte influência de políticos de direita nos grupos anti-vacina, enquanto nos Pro-vaxxers há uma mescla de influenciadores contando com políticos e divulgadores de ciências. A análise das comunidades polarizadas apresenta indícios de que os grupos são polarizados e possuem padrões de construção de conexões das comunidades de forma semelhante e alinhada com a ideologia do grupo.

7.6 Aspectos psicológicos

Investigamos as semelhanças e diferenças no uso de palavras que caracterizam os aspectos psicológicos de cada grupo através dos métodos proposto na Seção 4.5. Primeiro, aplicamos o teste estatístico do qui-quadrado em todas as 64 categorias de palavras do LIWC para comparar seu uso nos quatro grupos. As diferenças são estatisticamente significativas para todas as categorias de LIWC. Em seguida, analisamos essas diferenças em pares. A Tabela 7.9 mostra as porcentagens de uso das categorias LIWC usadas para investigar os quatro aspectos psicológicos.

Tabela 7.9: Percentuais de categorias LIWC para cada grupo

Aspecto	Categoria	Neutros	Antisino.	Antivax.	Provax.
Senso de	we	0.02	**0.05**	0.04	**0.05**
Grupo	assent	**0.07**	0.06	**0.07**	0.06
Preocupações	money	**0.30**	0.24	0.23	0.27
	leisure	**0.40**	0.30	0.32	0.22
	home	**0.06**	0.03	0.04	0.04
	health	0.19	0.12	0.12	**0.23**
	death	0.04	0.05	**0.08**	0.06
Emoções	anger	0.15	**0.19**	**0.19**	**0.19**
	sadness	**0.26**	0.22	0.19	0.19
	anxiety	0.09	0.10	0.10	**0.12**
	negative emotions	**0.43**	0.42	0.39	0.41
	positive emotions	0.51	0.50	0.49	**0.55**
Complexidade	exclusive	**0.61**	**0.61**	0.60	0.59
Cognitiva	conjunctions	**0.71**	0.62	0.63	0.65
	prepositions	**0.88**	0.73	0.70	0.80
	cognitive mechanisms	**0.97**	0.91	0.90	0.92

Observamos que os grupos mais semelhantes são os Anti-vaxxers e os Anti-sinovaxxers, onde não há diferenças estatísticas quanto ao uso de palavras pertencentes a 21 categorias do LIWC. Em comparação com os Neutros, todas as diferenças

são estatisticamente significativas, com seis exceções para o par Anti-vaxxers/Neutros, quatro para o par Pro-vaxxers/Neutros e três para Anti-vaxxers/Neutros. O par Anti-vaxxers/Pro-vaxxers apresenta 9 categorias LIWC com diferenças significativa, e o par Anti-sinovaxxers/Pro-vaxxers, 11 categorias. A semelhança dos Anti-vaxxers e dos Anti-sinovaxxers no uso de mais de 30% das categorias de LIWC é uma evidência de que esses grupos compartilham aspectos psicológicos comuns. Sua forte dissimilaridade reforça essa evidência com os grupos Neutros e Pró-vaxxers. Esses achados são consistentes com o efeito de câmara de eco, uma vez que existe uma semelhança no uso de categorias LIWC e aspectos psicológicos compostos por elas em grupos politicamente polarizados, mostrando que os indivíduos se comportam seletivamente (neste caso, verbalmente) para apoiar sua visão de mundo.

Em relação aos aspectos especificamente estudados não encontramos uma caracterização tão específica como no estudo de caso do Isolamento Social. Especificamente:

- **Coesão:** Não há grupo que se destaque em relação aos demais considerando o aspecto do senso de grupo. Enquanto Anti-sinovaxxers e Pro-vaxxers têm porcentagens mais altas na categoria *we*, Neutros e Anti-vaxxers exibem um maior uso de palavras da categoria *assent*;

- **Estados Afetivos:** Para os grupos polarizados, *anger* não tem diferenças significativas, mostra a maior porcentagem entre os grupos. Isso está alinhado com a necessidade de defender as posturas em suas mensagens, enquanto os Neutros permanecem em um estado mais especulativo. Os Pro-vaxxers têm a maior taxa na categoria de *positive emotions*, denotando seu otimismo por acreditar em algo que contém a pandemia COVID-19. Os Neutros apresentam o maior percentual na categoria de *negative emotions*, principalmente na subcategoria *sadness*, o que pode significar que especulações sobre a vinda da vacina e a comparação do Brasil com a situação ao redor do mundo causem alguns frustração para os usuários do grupo. O posicionamento de cada grupo sobre o tema vacinação norteia as emoções extraídas de seus *tweets*, seja de esperança ou angústia, corroborando com os temas encontrados na Análise do Tópico;

- **Complexidade Cognitiva:** Considerando as quatro categorias de LIWC utilizadas para descrever este aspecto, os Neutros apresentam o maior uso dessas classes (*exclusive, cognitive mechanisms, conjunctions, e prepositions*). Entre os demais grupos, o Pró-vaxxers apresenta percentuais mais elevados em três das quatro cate-

gorias. Essa é uma evidência de que os Neutros podem postar *tweets* com narrativas mais coerentes, complexas e concretas, em comparação com os três grupos politicamente polarizados, que são semelhantes em termos de complexidade cognitiva. Essas descobertas são consistentes com (PENNYCOOK et al., 2020), que relata que a ideologia não está relacionada a crenças sobre COVID, mas à sofisticação cognitiva.

- **Preocupações Pessoais:** O uso de classes nesta dimensão LIWC confirmou alguns tópicos encontrados na Seção 5.2.3: *home*, *leisure* e *money* têm os maiores percentuais em Neutros, envolvendo comentários sobre a importância do isolamento contínuo até a vacina, o preço estimado que a vacina poderia ter nas redes de hospitais privados e aguardando o fim da pandemia para retornar ao lazer e às atividades fora de casa. Os pró-vaxxers têm maior percentual em *health*, buscando conscientizar a população sobre a importância da vacina; e *death* está mais ligado aos Anti-vaxxers, comentando a suspeita que a vacinação sem testes suficientes pode trazer.

A análise das quatro proposições dos aspectos psicológicos pesquisados mostra que existem diferenças entre os grupos, embora mais sutis no tocante à polarização. Os grupos polarizados diferem em termos de preocupações pessoais e coesão do grupo, mas são mais próximos do que os Neutros em aspectos que envolvem emoções e complexidade cognitiva. A negatividade fornece evidências de que a defesa de seus pontos de vista decorre do descontentamento, refletindo o reconhecimento de pensamentos contraditórios às suas identidades como prejudiciais. Também é possível constatar que a baixa sofisticação cognitiva influencia mais a percepção da pandemia do que a orientação política. Isso é consistente com outros estudos (e.g., (PENNYCOOK et al., 2020)).

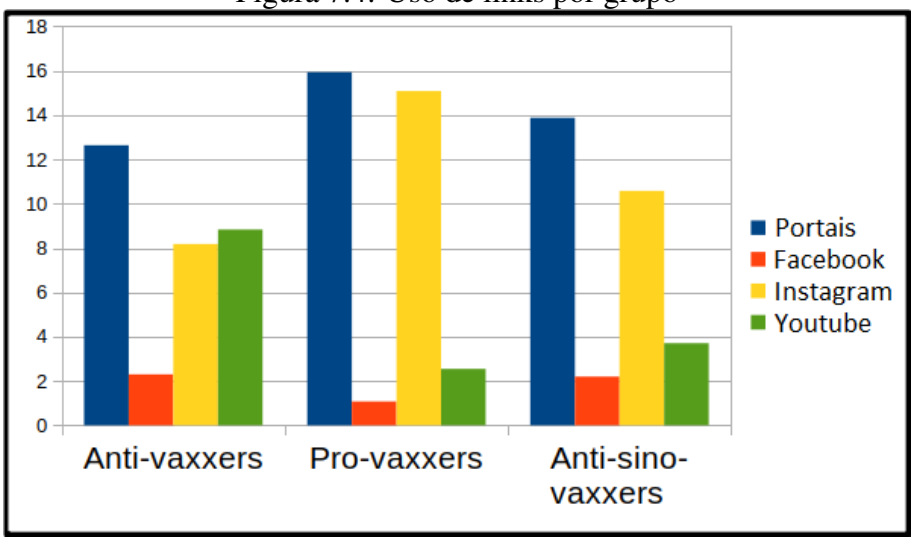
7.7 Fontes de Informação

Finalmente, as fontes de informação dos grupos foram analisadas através da investigação de três formas de disseminação de informação presentes nos *tweets* dos grupos, de acordo com a Seção 4.6. As Figuras 7.4, 7.5 e 7.6 representam respectivamente as proporções de utilização de endereços de sites, menções a *tweets* de outro usuário, e menção a *tweets* de políticos da lista definida pelo GPS ideológico.

a) Portais de Notícias e Redes Sociais

A Figura 7.4 apresenta a utilização de URLs de determinadas categorias (portais de notícias, Facebook, Instagram e Youtube) nos *tweets* dos grupos, dentre todos os *tweets* com algum endenreço virtual. Para ambos grupos observa-se que a categoria mais propagada com links é a de portais de notícias, assim como o Facebook é aquela com menor utilização. Pro-vaxxers e Anti-sinovaxxers possuem o Instagram como a segunda maior categoria utilizada, enquanto que Anti-vaxxers utilizam de forma secundária o Youtube. A soma das categorias de redes sociais nos grupos ultrapassa a utilização de portais de notícias (2.77% superior para Pro-vaxxers, 6.68% para Anti-vaxxers e 2.61% para Anti-sinovaxxers), mostrando que apesar de um esforço em validar um ponto de vista com uma informação acurada, a propagação de informações e opiniões de terceiros sem necessariamente uma conferência continua sendo uma forma comum de comunicação em discussões políticas.

Figura 7.4: Uso de links por grupo

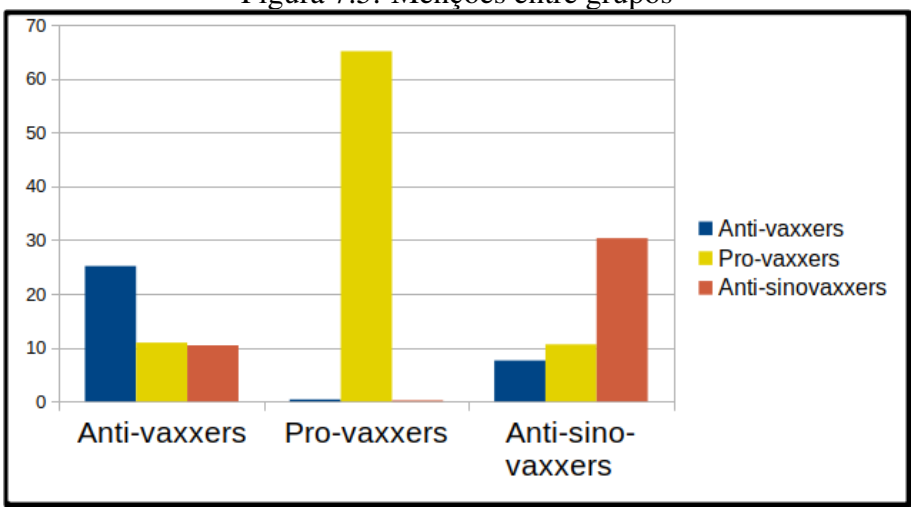


b) Menções de *Tweets*

A Figura 7.5 apresenta a proporção de menções a *tweets* escritos por usuários dos grupos, em relação às menções de *tweets* em geral. Pro-vaxxers é o grupo com maior auto-menção de seus usuários (65.12%), seguido por Anti-sinovaxxers (30.35%) e Anti-vaxxers (25.16%), porém não apresenta uma proporção significante de menções a *tweets* de outros grupos. Os grupos anti-vacina, entretanto, mostram um comportamento espelhado, onde Anti-vaxxers mencionam usuários Anti-sinovaxxers em 10.42% de seus *tweets* e vice-versa (7.59%). Este espelhamento de menções dos grupos anti-vacina podem ser obsevados por outro ângulo: sendo que ambos possuem preocupações e ideologia alinhadas, a soma de suas menções nos dois grupos configura um percentual de menções a grupos com algum posicionamento contra-vacinação. Com esta configu-

ração, Anti-vaxxers apresentam 35.58% de suas menções a *tweets* anti-vacina e Anti-sinovaxxers 37.94%. Deve-se atentar a pequena proporção de menções a outros grupos dos Pro-vaxxers: a quantidade de usuários é 6.4 vezes maior comparado aos usuários dos Anti-vaxxers.

Figura 7.5: Menções entre grupos



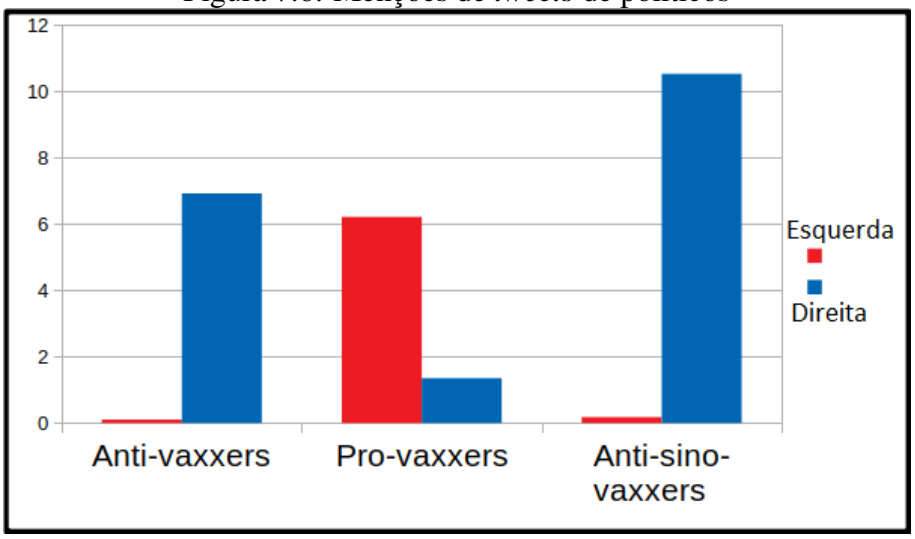
c) Menções de *Tweets* de Políticos

A Figura 7.6 apresenta a distribuição dos *tweets* mencionando políticos de esquerda e direita e aponta um comportamento contrário da análise da Figura 7.5. A análise anterior mostra que Pro-vaxxers pouco mencionam os demais grupos, porém é mencionado pelos mesmos de forma significante, já as menções a políticos mostra que Pro-vaxxers fazem uso em proporção superior a *tweets* de políticos de direita (1.35%), comparando com Anti-vaxxers e Anti-sinovaxxers a políticos de esquerda (0.10% e 0.17% respectivamente). Anti-vaxxers e Anti-sinovaxxers propagam políticos de direita (6.93% e 10.53% respectivamente), e Pro-vaxxers propaga políticos de esquerda (6.21%).

d) Discussão

Observa-se que todos os grupos possuem certa preocupação com a veracidade das informações e propagam grande proporção de links de portais de notícias para defender seus pontos de vista, porém ainda assim há um fluxo de proporção equivalente originado das redes sociais. Todos os grupos possuem um comportamento de câmara de eco, mencionando mais *tweets* de seu próprio grupo e de políticos alinhados com sua ideologia, configurando indícios de que há pouca abertura para discussões de visões diferentes.

Figura 7.6: Menções de *tweets* de políticos



7.8 Demografia

Com base no método apresentado na Seção 4.7, as Figuras 7.7 e 7.8 mostram a distribuição dos usuários em termos de gênero e idade para cada grupo, usando as proporções de cada grupo. Em relação ao sexo, o Anti-vaxxer, o Anti-Sinovaxxer e o Pro-vaxxer são semelhantes, com maior proporção de usuários do sexo masculino (55,21%, 54,64% e 53,72% respectivamente) em relação aos Neutros (40,48%). Em termos de idade, observamos que os Neutros são o grupo mais jovem. Com uma proporção de 72,05% com menos de 30 anos, pode ser comparado em idade ao usuário médio do Twitter[8]. Não há diferença significativa nas distribuições de idade dos Anti-vaxxers e dos pró-vaxxers, em que cerca de 45% dos usuários têm menos de 30 anos. Os anti-sinovaxxers são, em comparação, mais velhos, já que apenas 37,88% dos usuários têm menos de 30 anos.

Concluímos que não há diferença na demografia de Anti-vaxxers e Pro-vaxxers em termos de idade e sexo. Anti-sinovaxxers podem ser aproximados da demografia dos eleitores de Bolsonaro[9]. Os outros três grupos não podem ser aproximados nem da demografia da população brasileira, nem dos eleitores de Fernando Haddad, oposição do Bolsonaro na eleição de 2018.

[8]https://p.widencdn.net/kqy7ii/Digital2019-Report-en

[9]https://exame.com/brasil/homem-branco-e-conservador-um-perfil-dos-manifestantes-pro-bolsonaro-em-sp/

Figura 7.7: Demografia dos Usuários - Gênero

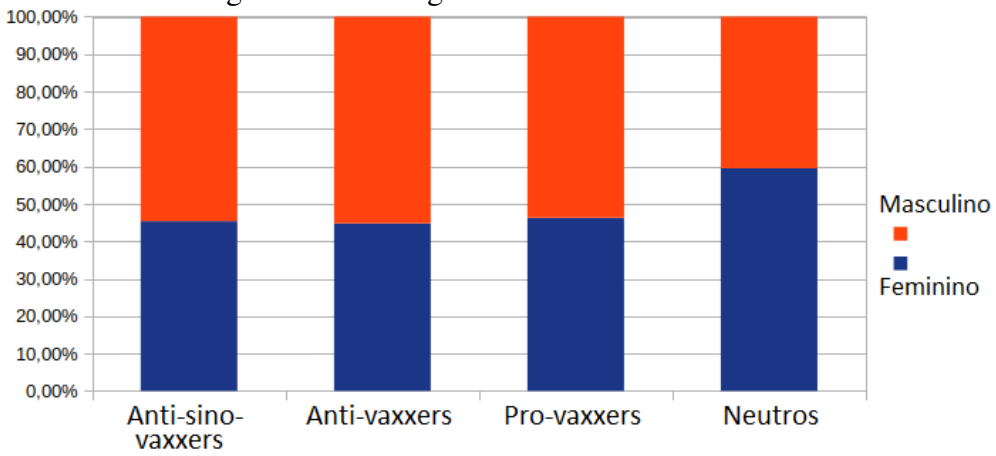
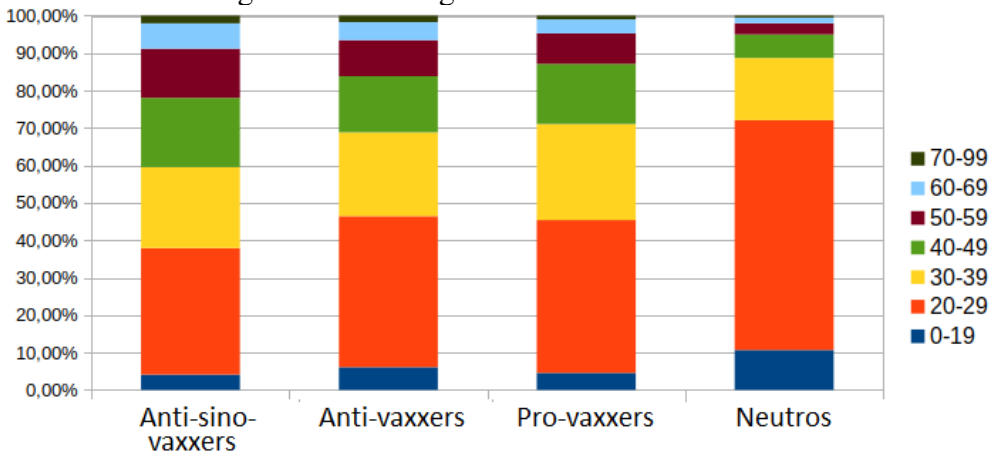


Figura 7.8: Demografia dos Usuários - Idade



7.9 Considerações Finais

A análise deste estudo de caso utilizando o framework proposto aponta que há influência da polarização política nas posturas de vacinação expressas por brasileiros no Twitter. Os grupos posicionados contra a vacinação possuem polarização política à direita, enquanto o grupo pró-vacina é polarizado à esquerda. Os grupos também mostram motivações políticas em seus posicionamentos, uma vez que dois candidatos às eleições presidenciais de 2022 estão usando a imunização COVID como plataforma eleitoral. Nossos resultados contradizem estudos que não observaram um viés político no comportamento anti-vacinação (HORNSEY M. J., 2018; CZARNEK GABRIELA; SZWED, 2020), mas nossa análise se restringe ao cenário específico do COVID brasileiro.

O Anti-sinovaxxers é o grupo mais polarizado, suas preocupações se relacionam com a origem da vacina e questões conspiratórias em relação às intenções políticas de

Doria. Anti-vaxxers e Pro-vaxxers se espelham em relação à polarização política: seus IPPs orbitam em direções opostas à neutralidade com uma distância semelhante, e as preocupações expressas divergem sobre a importância da imunização coletiva. Todos os grupos têm duas comunidades polarizadas, segregadas em termos de políticos de direita / esquerda seguidos. Em geral, enquanto um atua como uma bolha fechada, o outro é composto de usuários pré-dispostos a perfurar a bolha para lançar ideias. As comunidades polarizadas orientadas para a direita dos grupos anti-vacinação são mais densamente conectadas e, em geral, são mais influenciadas por políticos. Os Pro-vaxxers e os neutros compartilham pressa para a imunização, percebida como o único meio de voltar à rotina normal, mas os Pro-vaxxers criticam as ações do governo. Um efeito de câmara de eco foi observado em todos os grupos, principalmente propagando ideias alinhadas com seus próprios pontos de vista.

8 CONCLUSÃO E TRABALHOS FUTUROS

Nesta dissertação apresentamos a estrutura de um framework para análise de comportamento baseado na polarização política. A composição do framework busca providenciar meios para entendimento de seis diferentes dimensões de comportamento que se complementam, pressupondo coleta de dados disponíveis no Twitter. Para ajudar a explicar os padrões de comportamentos polarizados politicamente propomos uma métrica de medida de polarização de usuários, a análise de uma rede social baseado em métricas topológicas de rede, uma análise de dois níveis de granularidade com modelagem de tópico, caracterização de aspectos psicológicos baseados em palavras de estilo em textos de usuários, identificação de fontes de informação disseminadas entre as redes, e inferência automática da demografia de usuários. Para representar o poder de generalidade do framework, também mostramos a análise de dois estudos de caso no contexto da COVID-19: distanciamento social e aceitação de vacinas.

Os estudos de caso analisados mostraram certos padrões entre os grupos polarizado, representando evidências de que o comportamento e crenças de indivíduos no contexto da COVID-19 são afetados pela ideologia. Em ambos estudos o framework apontou que grupos identificados com a direita possuem IPPs mais extremos à direita, demografia próxima aos eleitores de Jair Bolsonaro, e que seus argumentos centrais são formados por elogios ao mesmo, preocupações com imposições de governos municipais e estaduais, e ataques aos opositores do governo central. Já os grupos ligados mais à esquerda disseminam informação pela rede utilizando notícias de grandes portais, são um pouco mais abertos ao debate, e seus argumentos são centrados em críticas ao presidente e preocupação com a saúde da população.

Os padrões observados nos dois estudos de caso configuram indícios de que a aplicação do framework em discussões polarizadas possibilita a caracterização e compreensão de comportamentos de grupos sociais, e com isso suas relações com ideologia política. O framework apresenta uma evolução do trabalho de Stieglitz e Dang-Xuan (2013), agregando outras dimensões de análise e generalizando a utilização para qualquer indivíduo, não focando apenas no contexto de figuras políticas. O trabalho contribui para possibilitar o estudo da influência polarização política em diversas discussões atuais, visto a crescente politização da população ao longo da última década.

A abordagem combinativa entre duas técnicas para quantificar a polarização política conseguiu adaptar-se ao contexto político atual. O uso da técnica proposta em (GARI-

MELLA; WEBER, 2017), combinado com divisão de políticos em cada espectro político com a técnica apresentada em (BARBERá et al., 2015), faz com que uma quantificação seja baseada em políticos que um indivíduo segue, ao mesmo tempo que o espectro destes políticos seguidos seja atualizado de acordo com a dinamicidade das relações deste contexto, principalmente em um país como o Brasil, facilitando a seleção de figuras políticas influentes de referência a esquerda e direita.

Nossa abordagem de modelagem de tópicos, combinando duas técnicas, contribui para as análises de assuntos de forma abrangente, possibilitando um entendimento por etapas. A utilização de LDA, para detectar macro tópicos, conduz a análise a uma primeira granularidade e o entendimento geral dos assuntos sendo abordados de uma forma geral por um grupo. Uma técnica de modelagem que leva em consideração *embeddings* contextuais, BERTopic, é potencializada por um espaço de busca já segmentado fornecido por LDA, possibilitando que argumentos com mesmo significado sejam agrupados. Então, a análise de um assunto após as duas técnicas aplicadas se torna rica por apontar argumentos chave dentro de bolhas de preocupações que compõe um posicionamento de um grupo.

A análise da estrutura social das redes dos grupos, juntamente com a análise das fontes de informação disseminadas neles e por eles, contribui para a análise das influências em grupos polarizados e sua relação com um maior comportamento de câmara de eco. Enquanto a análise da estrutura das redes aborda a influência de indivíduos na questão de conexões sociais, a análise das menções a *tweets* do grupo ou de políticos da mesma orientação política aponta os grupos que se mostram mais fechados à abertura de influência externa e que perpetuam mais a influência interna para validarem suas opiniões.

O desenvolvimento do framework resultou até o presente momento em quatro publicações diretas e um prêmio:

- EBELING, R. et al. Quarenteners vs. Cloroquiners: a framework to analyze the effect of political polarization on social distance stances. Em: **VIII Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)**, 2020 - recebendo o prêmio de melhor artigo do evento;

- EBELING, R. et al. Quarenteners vs. Chloroquiners: A framework to analyze how political polarization affects the behavior of groups. Em: **Proceedings of the 2020 Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)**, 2020;

- EBELING, R. et al. The effect of political polarization on social distance stances in the brazilian covid-19 scenario. Em: **Journal of Information and Data Management (JIDM)**, 2021;

- EBELING, R. et al. Analysis of the influence of political polarization in the vaccination stance: the brazilian covid-19 scenario. Em: **Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)**, 2022 (Em publicação).

Outros dois artigos publicados indiretamente contribuíram para a concepção do framework, instigando a exploração e identificação do uso de técnicas computacionais para analisar comportamentos de grupos:

- HARB, J.; EBELING, R.; BECKER, K. Exploring deep learning for the analysis of emotional reactions to terrorist events on twitter. Em: **Journal of Information and Data Management (JIDM)**, 2019;

- HARB, J. G. D.; EBELING, R.; BECKER, K. A framework to analyze the emotional reactions to mass violent events on twitter and influential factors. Em: **Information Processing & Management (IPM)**, 2020.

Identificamos no desenvolvimento deste trabalho algumas limitações que nos auxiliam a planejar a evolução o framework:

- Cálculo do IPP: a restrição do cálculo com a utilização dos perfis seguidos do usuário pode ser um fator que comprometa a escalabilidade do cálculo, uma vez que elementos nos próprios *tweets* do usuário, com coleta mais facilitada, podem ser explorados para mensuração da polarização política;

- Dados Demográficos: por mais acurada que seja a técnica, há possibilidade da imagem de perfil do usuário não ser atual ou mesmo ser do mesmo, além de não poder considerar a totalidade dos usuários quando muitos, por exemplo, não utilizam imagens;

- Eliminação de Bots: a procura por perfis com comportamento de bot pode ocasionar uma deleção de usuário real que organicamente emula o comportamento de robô;

- Utilização das Fontes de Informação: a análise do uso das fontes de informação é realizada de forma simples, observando a utilização ou não. Há um comportamento

possível que configura no uso da fonte criticando e tratando como uma desinformação, ou mesmo difundindo a fonte com uma interpretação errada.

Apesar dos resultados iniciais, pretendemos continuar o desenvolvimento do framework de forma geral, estudando formas de aprofundar as dimensões já existentes, assim como explorar novos aspectos de análise. Para as dimensões já existentes planejamos a) aumentar os aspectos psicológicos traçados; b) explorar novas métricas de polarização política; c) abordar as diferentes formas de dispersão da informação nas redes. Como novas análises almejamos estudar a mudança de polarização dos usuários ao longo do tempo para identificar características diferentes, assim como fatos que impactam na mudança de crença ideológica.

REFERÊNCIAS

AGGARWAL, C. C.; ZHAI, C. A survey of text clustering algorithms. In: **Mining text data**. [S.l.]: Springer, 2012. p. 77–128.

AJZENMAN, N.; CAVALCANTI, T.; MATA, D. D. **More than Words: Leaders' Speech and Risky Behavior During a Pandemic**. [S.l.], 2020. Available from Internet: <https://ideas.repec.org/p/cam/camdae/2034.html>.

ALSUMAIT, L. et al. Topic significance ranking of lda generative models. In: SPRINGER. **Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases**. [S.l.], 2009. p. 67–82.

ANGELOV, D. Top2vec: Distributed representations of topics. 2020.

BARBERá, P. et al. Tweeting from left to right: Is online political communication more than an echo chamber? **Psychological Science**, v. 26, n. 10, p. 1531–1542, 2015. PMID: 26297377. Available from Internet: <https://doi.org/10.1177/0956797615594620>.

BARRIOS, J. M.; HOCHBERG, Y. **Risk Perception Through the Lens of Politics in the Time of the COVID-19 Pandemic**. [S.l.], 2020. (Working Paper Series, 27008). Available from Internet: <http://www.nber.org/papers/w27008>.

BAZZAN, A. L. C. I will be there for you: clique, character centrality, and community detection in friends. **Computational and Applied Mathematics**, Springer Science and Business Media LLC, v. 39, n. 3, Jun 2020. ISSN 1807-0302. Available from Internet: <http://dx.doi.org/10.1007/s40314-020-01222-7>.

BEDI, P.; SHARMA, C. Community detection in social networks. **WIREs Data Mining and Knowledge Discovery**, v. 6, n. 3, p. 115–135, 2016. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1178>.

BIANCHI, F.; TERRAGNI, S.; HOVY, D. **Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence**. 2021.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. JMLR.org, v. 3, p. 993–1022, mar. 2003. ISSN 1532-4435.

BLONDEL, V. et al. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics Theory and Experiment**, v. 2008, 04 2008.

BOXELL, L.; GENTZKOW, M.; SHAPIRO, J. M. Greater internet use is not associated with faster growth in political polarization among us demographic groups. **Proceedings of the National Academy of Sciences**, v. 114, p. 10612 – 10617, 2017.

BRAMSON, A. et al. Disambiguation of social polarization concepts and measures. **The Journal of Mathematical Sociology**, Taylor & Francis, v. 40, n. 2, p. 80–111, 2016.

BRUIN, W. Bruine de; SAW, H.-W.; GOLDMAN, D. P. Political polarization in us residents' covid-19 risk perceptions, policy preferences, and protective behaviors. **Journal of Risk and Uncertainty**, v. 61, n. 2, p. 177 – 194, 2020.

BURKI, T. The online anti-vaccine movement in the age of covid-19. **The Lancet Digital Health**, v. 2, n. 10, p. e504 – e505, 2020. ISSN 2589-7500.

CATALAN-MATAMOROS, D.; ELíAS, C. Vaccine hesitancy in the age of coronavirus and fake news: Analysis of journalistic sources in the spanish quality press. **Int J Environ Res Public Health**, v. 17, n. 21, 2020.

CHAKRABORTY, A. et al. Who makes trends? understanding demographic biases in crowdsourced recommendations. **Proceedings of the International AAAI Conference on Web and Social Media**, v. 11, n. 1, p. 22–31, May 2017. Available from Internet: <https://ojs.aaai.org/index.php/ICWSM/article/view/14894>.

CHOUDHURY, M. D. et al. Social media participation in an activist movement for racial equality. In: **Proc. of the 10th Intl. Conf. on Web and Social Media (ICWSM)**. [S.l.: s.n.], 2016. p. 92–101.

CINELLI, M. et al. The COVID-19 social media infodemic. **Scientific Reports**, v. 10, n. 1, p. 16598, 2020. ISSN 2045-2322.

CONOVER, M. D. et al. Political polarization on twitter. In: ADAMIC, L. A.; BAEZA-YATES, R.; COUNTS, S. (Ed.). **Proc. of the Fifth International Conference on Weblogs and Social Media**. [S.l.]: The AAAI Press, 2011.

CONWAY, B. A.; KENSKI, K.; WANG, D. The rise of twitter in the political campaign: Searching for intermedia agenda-setting effects in the presidential primary. **Journal of Computer-Mediated Communication**, v. 20, n. 4, p. 363–380, 2015. Available from Internet: <https://onlinelibrary.wiley.com/doi/abs/10.1111/jcc4.12124>.

COSSARD, A. et al. Falling into the echo chamber: The italian vaccination debate on twitter. In: **Proc. of the Int. AAAI Conference on Web and Social Media**. [s.n.], 2020. v. 14, n. 1, p. 130–140. Available from Internet: <https://ojs.aaai.org/index.php/ICWSM/article/view/7285>.

COSTA, L. d. F. et al. Characterization of complex networks: A survey of measurements. **Advances in Physics**, Informa UK Limited, v. 56, n. 1, p. 167–242, Jan 2007. ISSN 1460-6976. Available from Internet: <http://dx.doi.org/10.1080/00018730601170527>.

CURIEL, R. P.; RAMíREZ, H. G. **Vaccination strategies against COVID-19 and the diffusion of anti-vaccination views. arXiv. 2009.13674**. 2020.

CZARNEK GABRIELA, M. K.; SZWED, P. **Political Ideology and Attitudes Toward Vaccination: Study Report. PsyArXiv. June 27. doi:10.31234/osf.io/uwehk**. 2020.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the American society for information science**, Wiley Online Library, v. 41, n. 6, p. 391–407, 1990.

DEMSZKY, D. et al. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In: **Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Ling.: Human Language Technologies**. [S.l.: s.n.], 2019. p. 2970–3005.

DENNY, M. J.; SPIRLING, A. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. **Political Analysis**, Cambridge University Press, v. 26, n. 2, p. 168–189, 2018.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019.

EBELING, R. et al. Quarenteners vs. cloroquiners: a framework to analyze the effect of political polarization on social distance stances. In: **Anais do VIII Symposium on Knowledge Discovery, Mining and Learning**. Porto Alegre, RS, Brasil: SBC, 2020. p. 89–96. ISSN 2763-8944. Available from Internet: <https://sol.sbc.org.br/index.php/kdmile/article/view/11963>.

EBELING, R. et al. Analysis of the influence of political polarization in the vaccination stance: the brazilian covid-19 scenario. **Proceedings of the International AAAI Conference on Web and Social Media**, v. 16, 2022. (in press).

EBELING, R. et al. Quarenteners vs. chloroquiners: A framework to analyze how political polarization affects the behavior of groups. In: **Proc. of the 2020 Conf. on Web Intelligence Conference (WI-IAT)**. [S.l.: s.n.], 2020.

EBELING, R. et al. The effect of political polarization on social distance stances in the brazilian covid-19 scenario. **Journal of Information and Data Management**, v. 12, n. 1, Aug. 2021. Available from Internet: <https://sol.sbc.org.br/journals/index.php/jidm/article/view/1889>.

EFFING, R.; HILLEGERSBERG, J. van; HUIBERS, T. Social media and political participation: Are facebook, twitter and youtube democratizing our political systems? In: TAMBOURIS, E.; MACINTOSH, A.; BRUIJN, H. de (Ed.). **Electronic Participation**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 25–35. ISBN 978-3-642-23333-3.

ELSHERIEF, M.; BELDING, E. M.; NGUYEN, D. # notokay: Understanding gender-based violence in social media. In: **Proc. of the 11th Intl. Conf. on Web and Social Media (ICWSM)**. [S.l.: s.n.], 2017. p. 52–61.

ESTÉVEZ, P. A. et al. Normalized mutual information feature selection. **IEEE Transactions on Neural Networks**, IEEE, v. 20, n. 2, p. 189–201, 2009.

FAN, H. et al. Learning deep face representation. **CoRR**, 2014. Available from Internet: <http://arxiv.org/abs/1403.2802>.

FORTUNATO, S. Community detection in graphs. **Physics Reports**, Elsevier BV, v. 486, n. 3-5, p. 75–174, Feb 2010. ISSN 0370-1573. Available from Internet: <http://dx.doi.org/10.1016/j.physrep.2009.11.002>.

Furini, M. et al. Untangling between fake-news and truth in social media to understand the covid-19 coronavirus. In: **Proc. of the 2020 IEEE Symposium on Computers and Communications (ISCC)**. [S.l.: s.n.], 2020. p. 1–6.

GARCIA, K.; BERTON, L. Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. **Applied Soft Computing**, v. 101, p. 107057, 03 2021.

GARIMELLA, V.; WEBER, I. A long-term analysis of polarization on twitter. In: **Proc. of the 11th Intl. Conf. on Web and Social Media (ICWSM)**. [S.l.: s.n.], 2017. p. 528–531.

GROOTENDORST, M. **BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.** Zenodo, 2020. Available from Internet: <https://doi.org/10.5281/zenodo.4430182>.

GROSSMAN, G. et al. Political partisanship influences behavioral responses to governors' recommendations for covid-19 prevention in the united states. **Proceedings of the National Academy of Sciences**, National Academy of Sciences, v. 117, n. 39, p. 24144–24153, 2020. ISSN 0027-8424. Available from Internet: <https://www.pnas.org/content/117/39/24144>.

HANSEN, D. L. et al. Chapter 3 - social network analysis: Measuring, mapping, and modeling collections of connections. In: HANSEN, D. L. et al. (Ed.). **Analyzing Social Media Networks with NodeXL (Second Edition)**. Second edition. Morgan Kaufmann, 2020. p. 31 – 51. ISBN 978-0-12-817756-3. Available from Internet: <http://www.sciencedirect.com/science/article/pii/B9780128177563000030>.

HARB, J. G. D.; EBELING, R.; BECKER, K. A framework to analyze the emotional reactions to mass violent events on twitter and influential factors. **Information Processing & Management**, v. 57, n. 6, p. 102372, 2020.

HAVEY, N. F. Partisan public health: how does political ideology influence support for COVID-19 related misinformation? **Journal of Comp. Social Science**, v. 3, n. 2, p. 319–342, 2020. ISSN 2432-2725. Available from Internet: <https://doi.org/10.1007/s42001-020-00089-2>.

HAVEY, N. F. Partisan public health: how does political ideology influence support for covid-19 related misinformation? **Journal of Computational Social Science**, p. 1 – 24, 2020.

HOFMANN, T. Probabilistic latent semantic analysis. **Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence**, Morgan Kaufmann Publishers Inc., p. 289–296, 1999.

HORNSEY M. J., H. E. A. . F. K. S. The psychological roots of anti-vaccination attitudes: A 24-nation investigation. **Health Psychology**, v. 37, n. 4, p. 307–315, 2018.

JIANG, J. et al. Political polarization drives online conversations about covid-19 in the united states. **Human Behavior and Emerging Technologies**, v. 2, n. 3, p. 200–211, 2020.

JUNGHERR, A. Twitter in politics: a comprehensive literature review. **Available at SSRN 2865150**, 2014.

JUNGHERR, A. Twitter use in election campaigns: A systematic literature review. **Journal of information technology & politics**, Taylor & Francis, v. 13, n. 1, p. 72–91, 2016.

KALIYAR, R. K.; GOSWAMI, A.; NARANG, P. Mcnnet: Generalizing fake news detection with a multichannel convolutional neural network using a novel covid-19 dataset. In: **Proc. of the 8th ACM IKDD CODS and 26th COMAD**. [S.l.: s.n.], 2021. (CODS COMAD 2021), p. 437.

KHERWA, P.; BANSAL, P. Topic modeling: a comprehensive review. **EAI Endorsed transactions on scalable information systems**, European Alliance for Innovation, v. 7, n. 24, 2020.

KNOWLES, R.; CARROLL, J.; DREDZE, M. Demographer: Extremely simple name demographics. In: **Proceedings of the First Workshop on NLP and Computational Social Science**. Austin, Texas: Association for Computational Linguistics, 2016. p. 108–113. Available from Internet: <https://aclanthology.org/W16-5614>.

LAZARUS, J. V. et al. A global survey of potential acceptance of a COVID-19 vaccine. **Nature Medicine. 10.1038/s41591-020-1124-9**, 2020. ISSN 1546-170X. Available from Internet: <https://doi.org/10.1038/s41591-020-1124-9>.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: **Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32**. [S.l.]: JMLR.org, 2014. (ICML'14), p. II–1188–II–1196.

LE, Q. V.; MIKOLOV, T. **Distributed Representations of Sentences and Documents**. 2014.

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. **Nature**, Nature Publishing Group, v. 401, n. 6755, p. 788–791, 1999.

LELKES, Y. Mass Polarization: Manifestations and Measurements. **Public Opinion Quarterly**, v. 80, n. S1, p. 392–410, 03 2016. ISSN 0033-362X.

LIKHITHA, S.; HARISH, B.; KUMAR, H. K. A detailed survey on topic modeling for document and short text data. **International Journal of Computer Applications**, v. 178, n. 39, p. 1–9, 2019.

LOTAN, G. et al. The arab spring| the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. **International Journal of Communication**, v. 5, p. 31, 2011.

LYU, H. et al. **Social Media Study of Public Opinions on Potential COVID-19 Vaccines: Informing Dissent, Disparities, and Dissemination. arXiv. 2012.02165**. 2020.

MACHADO, C. et al. News and political information consumption in brazil: Mapping the first round of the 2018 brazilian presidential election on twitter. **The computational propaganda project. Algorithms, automation and digital politics. https://comprop. oii. ox. ac. uk/research/brazil2018**, 2018.

MAKRIDIS, C.; ROTHWELL, J. T. The real cost of political polarization: Evidence from the covid-19 pandemic. **Covid Economics**, n. 34, p. 50–87, July 2020.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to information retrieval. **Natural Language Engineering**, Cambridge university press, v. 16, n. 1, p. 100–103, 2010.

MCINNES, L.; HEALY, J. Accelerated hierarchical density based clustering. In: IEEE. **Data Mining Workshops (ICDMW), 2017 IEEE International Conference on**. [S.l.], 2017. p. 33–42.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. In: BENGIO, Y.; LECUN, Y. (Ed.). **1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings**. [s.n.], 2013. Available from Internet: <http://arxiv.org/abs/1301.3781>.

MILOSH, M. et al. Unmasking partisanship: How polarization influences public responses to collective risk. **SSRN Electronic Journal**, 01 2020.

MIMNO, D.; BLEI, D. Bayesian checking for topic models. In: . [S.l.: s.n.], 2011. p. 227–237.

MUELLER, J.; STUMME, G. Gender inference using statistical name characteristics in twitter. In: **Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on Social Informatics 2016, Data Science 2016**. New York, NY, USA: Association for Computing Machinery, 2016. (MISNC, SI, DS 2016). ISBN 9781450341295. Available from Internet: <https://doi.org/10.1145/2955129.2955182>.

NARAYAN, A.; BERGER, B.; CHO, H. Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability. **bioRxiv**, 2020.

ORDUN, C.; PURUSHOTHAM, S.; RAFF, E. **Exploratory Analysis of Covid-19 Tweets using Topic Modeling, UMAP, and DiGraphs. arXiv:2005.03082**. 2020.

PENNEBAKER J. W., F. M. E.; BOOTH, R. J. inguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates, 2001.

PENNYCOOK, G. et al. Predictors of attitudes and misperceptions about covid-19 in canada, the UK, and the USA. doi:10.31234/osf.io/zhjkp. PsyArXiv, 2020.

PREOŢIUC-PIETRO, D. et al. Beyond binary labels: Political ideology prediction of twitter users. In: **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 729–740. Available from Internet: <https://www.aclweb.org/anthology/P17-1068>.

PUERARI, I. et al. Exploratory analysis of electronic health records using topic modeling. **Journal of Information and Data Management**, v. 11, n. 2, p. 131–147, 2020.

QIANG, J. et al. Short text topic modeling techniques, applications, and performance: a survey. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, 2020.

RAO, A. et al. Political partisanship and anti-science attitudes in online discussions about covid-19. **CoRR**, abs/2011.08498, 2020.

RÖDER, M.; BOTH, A.; HINNEBURG, A. Exploring the space of topic coherence measures. In: ACM. **Proceedings of the eighth ACM international conference on Web search and data mining**. [S.l.], 2015. p. 399–408.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In: **Proc. of the 19th International Conference on World Wide Web**. [S.l.: s.n.], 2010. p. 851–860.

SHA, H. et al. Dynamic topic modeling of the COVID-19 twitter narrative among U.S. governors and cabinet executives. **CoRR**, abs/2004.11692, 2020.

SHA, H. et al. Dynamic topic modeling of the covid-19 twitter narrative among u.s. governors and cabinet executives. **CoRR**, abs/2004.11692, 2020.

SLATCHER, R. et al. Winning words: Individual differences in linguistic style among u.s. presidential and vice presidential candidates. **Journal of Research in Personality**, v. 41, n. 1, p. 63 – 75, 2007. ISSN 0092-6566. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0092656606000183>.

SOARES F. B., R. R. V. T. F. G. . S. G. Research note: Bolsonaro's firehose: How covid-19 disinformation on whatsapp was used to fight a government political crisis in brazil. **Harvard Kennedy School (HKS) Misinformation Review**, 2021.

STIEGLITZ, S.; DANG-XUAN, L. Social media and political communication: a social media analytics framework. **Social network analysis and mining**, Springer, v. 3, n. 4, p. 1277–1291, 2013.

TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. **Journal of Language and Social Psychology**, v. 29, n. 1, p. 24–54, 2010.

TOMENY, T. S.; VARGO, C. J.; EL-TOUKHY, S. Geographic and demographic correlates of autism-related anti-vaccine beliefs on twitter, 2009-15. **Social science & medicine**, Elsevier, v. 191, p. 168–175, 2017.

VARGAS-CALDERÓN, V. et al. Using machine learning and information visualisation for discovering latent topics in twitter news. **CoRR**, abs/1910.09114, 2019. Available from Internet: <http://arxiv.org/abs/1910.09114>.

VIKATOS, P. et al. Linguistic diversities of demographic groups in twitter. In: **Proceedings of the 28th ACM Conference on Hypertext and Social Media**. New York, NY, USA: Association for Computing Machinery, 2017. (HT '17), p. 275–284. ISBN 9781450347082. Available from Internet: <https://doi.org/10.1145/3078714.3078742>.

WALTER, R.; BECKER, K. Caracterização e comparação de campanhas promovendo o outubro rosa e o novembro azul no twitter. In: HOLANDA, M.; MONTEIRO, J. M. (Ed.). **XXXIII Simpósio Brasileiro de Banco de Dados: Demos e WTDBD, SBBD 2018 Companion, Rio de Janeiro, RJ, Brazil, August 25-26, 2018**. SBC, 2018. p. 81–87. Available from Internet: <http://sbbd.org.br/2018/wp-content/uploads/sites/5/2018/08/081-sbbd_2018_comp.pdf>.