Capítulo 3

Polarização em Redes Sociais: Conceitos, Aplicações e Desafios

Bruno Hott[1], Bruno P. Santos[1,2], Túlio Corrêa Loures[1], Fabrício Benevenuto[1] e Pedro O.S. Vaz-de-Melo[1]

[1]Departamento de Ciência da Computação, UFMG

[2]Departamento de Ciência da Computação, UFBA

{brunohott,loures.tc,fabricio,olmo}@dcc.ufmg.br, bruno.ps@ufba.br

Abstract

The polarization assessed in social networks has reflected the predisposition of society to the clash of ideas and the recent encouragement of political rivalry in the world. From this context, several questions are raised, such as: Are people becoming more polarized? If so, what are the positive and negative impacts of social media on this process? Is it possible to measure polarization in social networks? The goal of this tutorial is to discuss the current scenario of research in polarization, through a critical overview of the area, its challenges and opportunities. For such, the main concepts and definitions of polarization will be presented. The flow of data collection on polarization, its processing, analysis and knowledge extraction will also be presented. For the latter, a special focus will be given to a taxonomy proposal for polarization metrics in social networks. At end, we will exercise our knowledge in a practical analysis of polarization applied to the Covid-19 topic.

Resumo

A polarização aferida em redes sociais tem refletido a predisposição da sociedade para o embate de ideias e o recente incentivo a rivalidade política no mundo. Deste contexto, diversas questões são levantadas, tais como: As pessoas estão se tornando mais polarizadas? Em caso afirmativo, quais são os impactos positivos e negativos das redes sociais neste processo? É possível medir polarização nas redes sociais? Neste minicurso, o objetivo é discutir o atual cenário de pesquisa em polarização, através de uma visão crítica geral da área, seus desafios e oportunidades. Para tal, os principais conceitos e definições sobre polarização serão apresentados. Assim como o fluxo de coleta dados sobre polarização, seu processamento, análises e extração de conhecimento. Para este último, será dado enfoque especial em uma proposta de taxonomia para métricas de polarização em redes sociais. Ao final, exercitaremos nossos conhecimentos em uma análise prática de polarização aplicada ao tópico Covid-19.

3.1. Introdução

Incipit Vita Nova (uma vida nova começa) para estudos de polarização com o advento da Internet. Essa ferramenta transformou o modo como interagimos e nos comunicamos e, especialmente, potencializou a disseminação de opiniões individuais e coletivas. A polarização tem ganhado enfoque especial da academia ao longo do tempo e, mais recentemente [Moreira et al., 2020; Valensise et al., 2022], também da indústria [Kubin and von Sikorski, 2021], devido ao seu impacto (positivo e negativo) potencial na sociedade (por exemplo, através da economia e política). Este capítulo aborda a Polarização por meio de uma perspectiva teórica e prática. O conteúdo aqui apresentado explora a estrutura, organização, desafios e aplicações dos estudos de polarização com enfoque na Internet, especialmente nas mídias sociais. Nesta seção, iniciamos discussões conceituais sobre polarização, motivamos o leitor com casos recentes e históricos, bem como apresentamos uma proposta de metodologia que guia o capítulo. Para iniciar a discussão, será levantada a seguinte questão: *o que é Polarização?*

A gêneses da polarização moderna aparece ao expressarmos opiniões, desejos ou intenções em veículos propagadores que variam desde a TV até *web blogs* pessoais. No Brasil, por exemplo, os candidatos(as) à presidência do país afirmam e levantam evidências, em debate na TV[1], que as eleições presidenciais de 2022 estão polarizadas. Na Internet, o Facebook tem sido acusado de promover conteúdos que dividem a população, podendo criar ou acelerar polarizações políticas[2]. No caso das mídias sociais, uma motivação da promoção destes conteúdos seria explorar a pré-disposição do cérebro humano em consumir conteúdos que reforçam sua visão de mundo [Klayman, 1995]. Desta forma, a hipótese é que as plataformas estão entregando cada vez mais conteúdos extremos com o objetivo de que os usuários fiquem por mais tempo na plataforma.

Alguns tópicos são naturalmente alvos de debate, como esportes, política, consumo de drogas, entre outros. Empiricamente, essas discussões tendem a fazer com que as pessoas movam suas opiniões para versões mais extremas delas mesmas [Sunstein, 1999]. Por exemplo, aqueles com opiniões contrárias à regulamentação das drogas se tornarão extremamente contrárias depois de interagir com quem compartilha sua visão. As pessoas cada vez mais estão ouvindo suas próprias vozes sendo ecoadas por seus semelhantes. Como consequência, se torna mais difícil para que a população possa resolver os problemas que a sociedade enfrenta conjuntamente, como o aquecimento global, por exemplo [Sunstein, 2018].

Por um lado, a Internet e, em especial, as mídias sociais vem sendo apontadas como estimuladoras da polarização entre os americanos nas últimas décadas ao criarem as chamadas "câmaras de eco" [Iyengar et al., 2012; Lelkes, 2016]. Nessas câmaras, as pessoas são estimuladas por pensamentos análogos, o que as isolam de divergências de opiniões e pensamentos [Bright, 2017; Lima et al., 2018]. Mas é preciso cautela, pois outros autores afirmam que o aumento da polarização pode não estar diretamente ligada ao uso das mídias sociais. Por exemplo, segundo [Boxell et al., 2017], os níveis de

[1]Primeiro debate presidencial 2022 organizado por um pool de emissoras formado entre o Grupo Bandeirantes de Comunicação, a TV Cultura, o jornal Folha de S. Paulo e o portal Uol.

[2]`https://www.washingtonpost.com/opinions/2020/10/26/facebook-algorithm-conservative-liberal-extremes/`

polarização crescem mais em populações que tipicamente não usam medias sociais como, por exemplo, a população idosa.

Segundo o dicionário Priberam[3], polarização é a concentração de ideias em um polo que se opõe a outro. Também existe uma vertente nas ciências sociais que indica que a polarização se dá quando membros de grupos da sociedade se movem na direção dos extremos [Fiorina et al., 2008]. Há quem defina polarização como um processo social, onde grupos se dividem em dois sub-grupos opostos cada qual com posições conflitantes uma as outras, com alguns poucos indivíduos neutros [Sunstein, 1999].

As definições possuem em comum a ideia de que a polarização divide um grupo social por meio da divergência de ideias. E é com esse conceito em mente que trabalharemos ao longo do capítulo. Vale notar que o enfoque deste trabalho se dá no recorte técnico de sua identificação. A análise completa dos impactos sociais e desdobramentos advindos da polarização vão muito além do escopo deste trabalho.

3.1.1. Perspectiva Histórica

Polarização não é um tópico de pesquisa novo, embora tenha recebido grande atenção acadêmica e industrial nos últimos anos. A Figura 3.1 exibe uma linha do tempo de artigos publicados na base de dados *Web of Science* contendo a conjunção das seguintes palavras-chave: *'Political'*, *'Polarization'* e *'Media'*. Estas chaves estão diretamente relacionadas ao conteúdo aqui abordado, apesar de filtrarem somente um recorte dos trabalhos referentes ao tópico polarização.

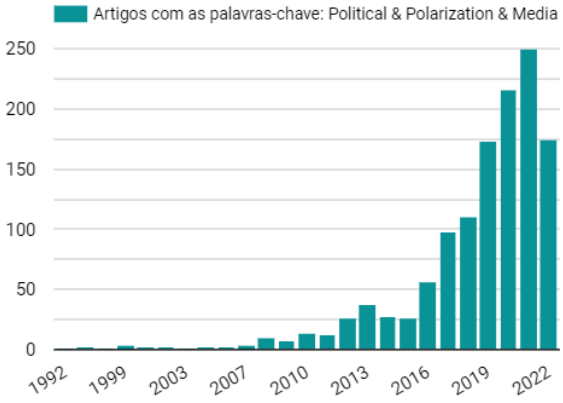

Figura 3.1: Artigos no Web of Science[4]

O artigo mais antigo presente em nossa busca a influência de variáveis religiosas em questões associadas ao aborto aborto [Woodrum and Davison, 1992], um tema ainda controverso em 2022 [Roy and Goldwasser, 2020]. Em contrapartida, o artigo mais recente é de julho de 2022, em que Borah and Singh [2022] realizam investigações sobre como a rede social Twitter tem sido usada para formar enlaces de comunicação inter e extra partidos e seus desdobramentos na divergência de opiniões no discuso político.

O estudo da polarização atrai pesquisadores por diversos motivos. Primeiro pelo valor agregado como, por exemplo, o impacto social na identificação de polarização, no combate a desinformação e informação falsa, na proposta de técnicas para redução da polarização, entre outros. Segundo, por quê a temática é inerentemente multidisciplinar, o que cabe percepções e abordagens diferentes sobre a problemática. Terceiro, por quê há um interseção importante entre a academia e a indústria sobre a temática que busca por soluções cientes da polarização – vide o recente caso da recomendação de conteúdos do

[3] `https://dicionario.priberam.org`
[4] Dados coletados até Agosto de 2022.

Facebook[5].

Do ponto de vista da computação, há diversas frentes de estudo que residem na interseção entre disciplinas distintas. Por exemplo, a aplicação de técnicas de inteligência computacional na identificação, previsão e controle da polarização [Belcastro et al., 2020; Garimella et al., 2021; Ribeiro et al., 2019; Tokita et al., 2021]. Uma lista não exaustiva de áreas de pesquisa que trabalham com a caracterização da polarização dentro do contexto da computação são: i) coleta e processamento de bases de dados [Garimella et al., 2018b]; ii) mineração de opiniões e análise de sentimentos [Rathje et al., 2021], entre outros, com o objetivo de classificar o viés dos usuários ou conteúdo [Bakshy et al., 2015; Garimella et al., 2021; Weld et al., 2021]; iii) desenvolvimento de modelos estatísticos, da teoria dos grafos, da inteligência artificial, entre outros, com o objetivo de classificar e quantificar a polarização [Garimella et al., 2018b; Pergola et al., 2020; Vicario et al., 2019]; iv) visualização e análise de resultados [Jang and Allan, 2018; Roy and Goldwasser, 2020].

3.1.2. Uma taxonomia para o estudo da polarização

Indicamos que a polarização pode ser caracterizada pela divisão das entidades de um grupo em duas partições com posições conflitantes. Logo, antes de analisarmos a polarização de uma população, precisamos definir a posição ou o *viés* de cada uma das entidades que a compõe.

Viés é definido como posicionamento, ou apoio, de um usuário ou declaração com relação a um tópico específico. Por exemplo, ao escolhermos o tópico "legalização das drogas", teremos um conjunto de indivíduos com viés pró-legalização e outro com viés contrário à legalização. Formalizamos o cálculo do viés como mostrado na Equação 1, onde X representa a declaração ou indivíduo a ser analisado com relação a um tópico específico (T). Como saída temos um dos três rótulos: $\{negativo, positivo, neutro\}$.

$$V(X \mid T) = \{\text{Negativo}, \text{Positivo}, \text{Neutro}\} \tag{1}$$

É importante ressaltar que a representação do viés pode ser feita de diferentes formas, como posicionamento binário, discreto de múltiplos níveis (como na Equação 1), ou numérico contínuo, por exemplo, um número real no intervalo entre $[-1;1]$.

O viés pode ser calculado para um indivíduo (ex.: usuário de uma rede social) ou ainda para um conteúdo. Por exemplo, poderemos encontrar um texto em uma página de notícias com viés claramente favorável ao tema de liberação de drogas, em contraponto a uma notícia de outro jornal que aborda o tema com viés negativo. Neste capítulo abordaremos somente viés baseado em conteúdos textuais, porém o mesmo se aplica para outros tipos de mídia, como imagens ou vídeos, por exemplo.

A Figura 3.2 apresenta uma proposta de taxonomia para o estudo sobre polarização e viés. O viés é calculado individualmente para cada usuário ou conteúdo da rede social. Por sua vez a polarização é uma métrica de grupos, que pode ser um grupo de usuários ou um grupo de conteúdos em torno de um tema. Dessa forma fica claro que as métricas ou ferramentas utilizadas para cálculo do viés serão diferentes daquelas para cálculo da polarização. A Figura 3.2 ainda mostra, como exemplo, que o viés pode ser extraído por meio de questionários (*surveys*) passados para cada usuário ou por métodos

[5]Observar nota de rodapé 2

computacionais, como por meio da extração de *hashtags* com vieses conhecidos contidos em*tweets*. As métricas de polarização podem ser calculadas utilizando métricas estatísticas [Akhtar et al., 2019; Morales et al., 2015], de teoria dos grafos [Garimella et al., 2018b] ou ainda por meio de técnicas de inteigência computacional [Al Amin et al., 2017; Roy and Goldwasser, 2020], tais como aprendizado de máquina, ciência dos dados, mineração de informação, entre outras.

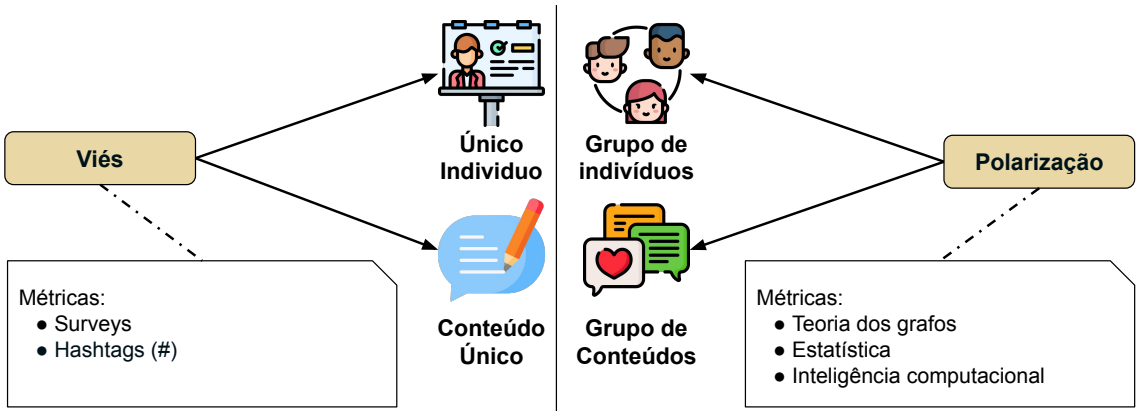

Figura 3.2: Taxonomia proposta de viés e polarização

3.1.3. Metodologia: um mapa do Capítulo

Apresentamos uma metodologia de quatro estágios, como mostrado na Figura 3.3, para realizar análises de polarização em mídias sociais, a saber: i) *Dados sobre Polarização*; ii) *Viés de entidades dos dados*; iii) *Polarização de grupo* e; iv) *Análise de polarização*. A seguir, apresentaremos uma breve descrição de cada estágio e, nas seguintes seções do capítulo aprofundamos discussões sobre cada etapa.

Figura 3.3: Visão geral do Capítulo

Dados sobre Polarização. O propósito deste estágio é capturar e pré-processar os dados referentes a polarização. Dados sobre polarização tipicamente são valores discretos, os quais são derivados de pessoas que expressam seus pensamentos, desejos ou opiniões através de meios de comunicação e interações. No contexto de polarização em mídias sociais, por exemplo, pode-se capturar e modelar dados correspondentes a um tópico de conversação juntamente com um conjunto de entidades relacionadas a ele. Um tópico pode representar um tema ou um assunto de discussão e pode ser operacionalizado (isto é, capturar dados brutos) através de um conjunto de palavras-chave ou *hashtags*, uma comunidade, ou um conteúdo, entre outros. As entidades relacionadas a um tópico consiste nos atores e na interação com aquele tema, discussão ou conteúdo, e pode ser operacionalizado como o conjunto de usuários, postagens, comentários, avaliações ou *likes/dislikes* relacionados ao tópico em questão. Por exemplo, um tópico pode ser representado por uma palavra-chave, como "#ukraine", o que, neste caso, as entidades relacionadas podem

consistir nos *tweets* que contêm aquela *hashtag* ou palavras relacionadas, como "#kyiv" e "#stoprussianaggression". Embora vamos discutir neste trabalho tópicos polarizadores de maneira textual, em princípio, os tópicos podem possuir diversas formas, uma vez que eles representem interações antagônicas entre usuários em torno de um tópico e isso pode surgir através de, por exemplo, mídias em vídeos ou fotos. A Seção 3.2 detalha as diferentes maneiras de realizar a extração e processamento destes dados.

Viés de entidades dos dados. Em uma segunda etapa do processo metodológico, unidades de informação referentes a um tópico de discussão são processadas a fim de se definir um posicionamento (referente aos polos). O objetivo aqui é particionar o conjunto de entidades em três conjuntos disjuntos: i) O conjunto de entidades que suportam o tópico em questão; ii) aqueles que são contrários ao tópico em questão e iii) aqueles que são neutros ao tópico em questão. Lembre que isso pode ser feito tanto de maneira discreta quanto em uma escala contínua. Em termos didáticos, o resultado desta etapa responde a seguinte questão: "assumindo que as entidades se dividem em dois conjuntos opostos de acordo com seu posicionamento com relação ao tópico, quais são esses conjuntos?". A Seção 3.3 detalha as diferentes técnicas para realizar o particionamento das atividades de um tópico.

Polarização de grupo. O terceiro estágio trabalha com os dados pré-processados já contendo a informação do viés de suas unidades. Intuitivamente, a polarização de um tópico que, a depender da técnica empregada, pode ser quantificada, expressa o quão bem separadas as duas partições estão, isto é, o quanto aqueles dois grupos divergem entre si. Apresentaremos diversas métricas para capturar a polarização, incluindo algumas baseadas em representações em baixa dimensão, estatística e teoria dos grafos. A Seção 3.4 apresenta maiores detalhes sobre essas técnicas.

Análise de polarização. A última etapa da metodologia adotada tem por objetivo realizar a análise dos dados e extração de conhecimentos a partir das etapas anteriores. Os tipos de análises e conclusões podem variar bastante a depender do tópico a ser analisado e do objetivo almejado. Alguns temas que podem ser explorados são o impacto das redes sociais na polarização encontrada na sociedade, análises do movimento da polarização ao longo do tempo, dentre outros tantos encontrados na literatura. A Seção 3.5 apresenta uma lista não exaustiva dos tópicos encontrados na literatura para que o leitor utilize como inspiração para seus trabalhos na área.

3.2. Dados no contexto de Polarização

Para derivar informações de polarização a partir de dados dados, é necessário que, primeiro, algumas medidas sejam tomadas. Nesta seção, serão destacadas as principais características e desafios encontrados ao lidar com dados relativos à polarização. Para tanto, serão apresentadas técnicas ou abordagens utilizadas para modelar e pré-processar esses dados. Ademais, será feito uma ligação entre o uso e a extração de informações sobre polarização, as quais são objeto de discussão do restante do capítulo.

3.2.1. A gênesis dos dados sobre polarização

A gênese dos dados, no contexto do estudo de polarização, é de pessoas que expressam seus pensamentos, desejos e/ou opiniões em diversos veículos propagadores de informa-

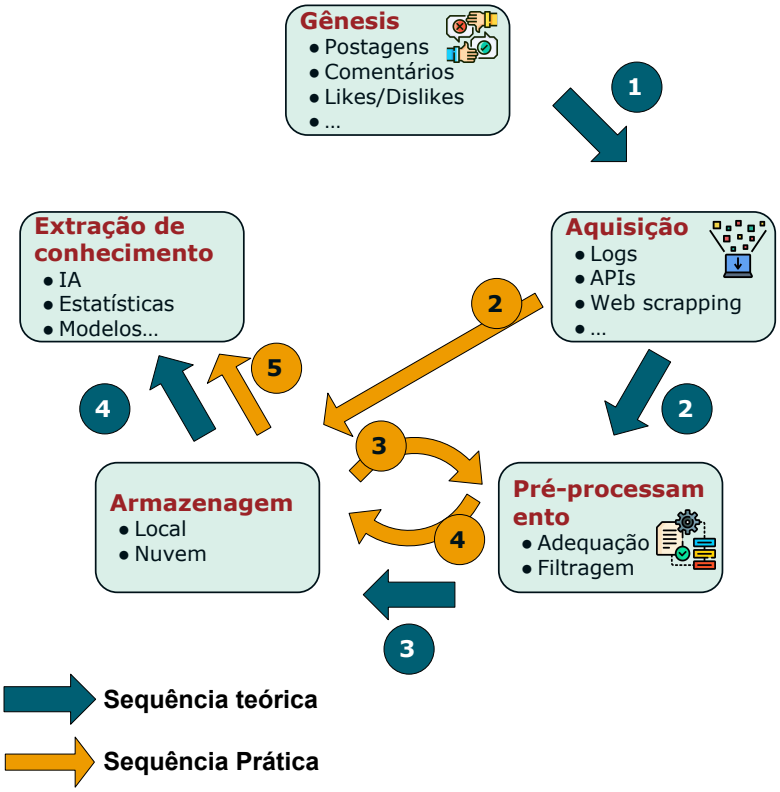

Figura 3.4: Etapas para extração da Polarização a partir de dados brutos

ção, tais como rodas de conversas, veículos de comunicação (TV, Rádio, blogs, jornais) e mídias social. Atualmente, essas últimas estão entre os principais veículos onde pessoas se expressam e propagam informações, normalmente na forma de postagens, comentários, indicação de aproximação através de *likes*, afastamento com *dislikes*, entre outras possibilidades.

Embora as mídias sociais não sejam os únicos veículos onde expressamos nossas opiniões, elas estão sendo massivamente usadas por pessoas como local de expressão [Gokcekus et al., 2021; Milroy and Llamas, 2013; Mitchell, 1974]. Pode-se ainda considerar, sem perda de generalidade, que as mídias sociais têm tido uma capacidade significativa de propiciar conhecimentos sobre posicionamentos de seus inscritos[6], bem como de grupos [Barros et al., 2021; Ferreira et al., 2021; Küçük and Can, 2020].

No contexto de coleta de dados em polarização, será considerado, neste trabalho, mídias sociais como "agregadores de dados de polarização", muito embora essa não seja sua finalidade. Ademais, destacamos que o conteúdo aqui apresentado pode ser extrapolado para outros domínios além das mídias sociais.

[6]Aqui estamos nos referindo a pessoas comuns, mas pode-se extrapolar para entidades genéricas, por exemplo, um conteúdo em vídeo/imagem ou outras quaisquer.

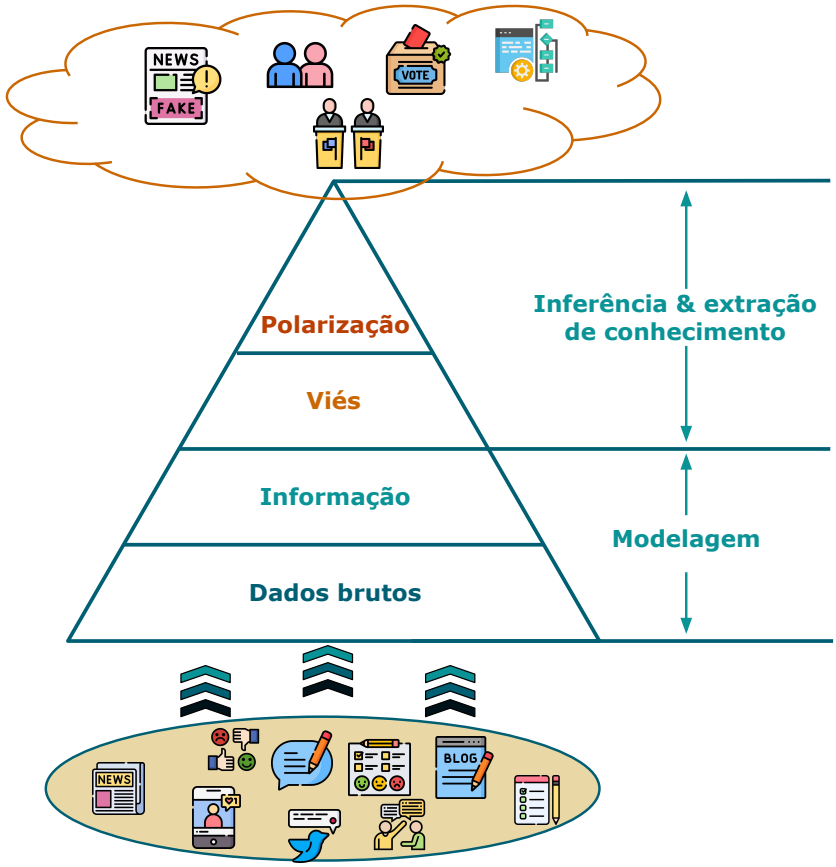

Figura 3.5: Hierarquia dos níveis de conhecimento sobre polarização a partir dos dados brutos

3.2.2. Desafios a partir dos dados

O principal problema aqui abordado é o de extração de conhecimento sobre polarização a partir de dados. Essa extração refere-se ao processo de modelar e analisar dados a fim de operacionalizar a inferência sobre posicionamentos de entidades individualmente ou em grupo. Por exemplo, em Küçük and Can [2020], os autores apresentam diferentes abordagens computacionais para indicar posicionamentos de *postagens individuais* em redes sociais ou textos comuns (ex.: blogs). Já em Ferreira et al. [2021], os autores estão preocupados com o posicionamento de *grupos de indivíduos* que interagem com um tópico de conversação em redes sociais.

Na Figura 3.4 são sumarizadas as etapas ideais e práticas para melhor entender o processo de extração de conhecimento. Iniciemos pelas etapas ideais (ou teóricas), as quais partem da aquisição/coleta de dados brutos sobre polarização (Etapa 1 em verde), passam pelo pré-processamento e armazenagem (Etapas 2 e 3 em verde), até a extração de conhecimentos a partir dos dados brutos (Etapa 4 em verde). As etapas intermediárias de pré-processamento e armazenamento são fundamentais para que os dados estejam aptos a serem utilizados por técnicas de inferência utilizadas na extração de conhecimentos sobre polarização.

Uma outra forma complementar de visualizar o processo é a abordagem hierár-

quica centrada em níveis de conhecimento conforme ilustrada na Figura 3.5, inspirada no trabalho de Santos et al. [2016]. Nessa abordagem, a transformação dos dados brutos é vista como uma hierarquia em que seus níveis são divididos em dois momentos: i) modelagem; e ii) análise e inferência sobre polarização. Na modelagem dos dados, o objetivo é adicionar algum nível de semântica e padronização aos dados brutos, os quais tipicamente são não estruturados, de fontes geradoras diferentes, e ainda podem possuir formatos heterogêneos. Nesta direção, técnicas de pré-processamento como representação de dados, filtragem ou fusão de dados podem ser aplicadas [Ayed et al., 2015; Khaleghi et al., 2013; Santos et al., 2016]. Já a análise e inferência tem por meta aplicar interpretações visando delinear o contexto a partir das informações e bem como extrair conhecimento acerca da situação daquele dado que, neste trabalho, é o *posicionamento* perante os grupos que se deseja estudar.

Embora as etapas intermediárias, entre os dados brutos e a extração de conhecimento sobre polarização, sejam essenciais, elas não são realizadas seguindo a sequência ideal. O que acontece na prática é a sequência destacada em amarelo na Figura 3.4, especialmente do ponto de vista de desenvolvimento. Tipicamente obtém-se os dados brutos e ciclos de pré-processamento e armazenamentos são realizados (em uma espécie de tentativas e erros/acertos) para que então os dados sejam encaminhados à algoritmos e técnicas de extração de conhecimento. O restante da seção tece considerações acerca do fluxo de dados na prática e suas particularidades do ponto de vista introdutório.

3.2.3. Aquisição dos dados

A primeira etapa do processo aqui abordado é o de aquisição de dados. Como os dados sobre polarização surgem de formas distintas, naturalmente as formas de aquisição também serão. Por exemplo, dados como texto opinativos de blogs ou páginas web pessoais podem ser coletados através de *Web Scraping*, enquanto *posts* em redes sociais podem ser adquiridos através de *Application Programming Interfaces (APIs)* públicas disponibilizadas pelos proprietários da rede social. Além disso, há diferentes formas de aquisição de dados sobre polarização, como por exemplo, questionários/formulários, observação, entrevistas, histórias orais, dentre outras. Aqui, serão comentadas as abordagens automáticas *Web Scrapping* e via *APIs*, tipicamente utilizadas para coletar dados da Internet.

O processo de ***Web Scrapping***, em português "raspagem da web", é uma técnica utilizada para extrair conteúdos ou conjuntos de dados diretamente de páginas da web [Khder, 2021]. Esse processo de extração de dados pode ser feito de modo manual ou automático, o qual popularmente se chama de *bot* ou *scraper*. É possível realizar *Web Scrapping* de diversas formas, e dentre elas, três ferramentas gratuitas se destacam: i) O FrameWork Scrapy[7] e ii) Beautiful Soup[8]; iii) Selenium[9]. O primeiro foi originalmente projetado para *Web Scrapping*, embora possa ser utilizado para outras finalidades como, por exemplo, coletar dados de APIs. O segundo, Beautiful Soup, é uma biblioteca para Python que visa fornecer meios para analisar e "caminhar" (*tree travessal*) em páginas web, permitindo recuperar informações específicas contidas na página. Já o terceiro, Selenium, foi primariamente projetado para testar navegadores.

[7] https://scrapy.org/

[8] https://www.crummy.com/software/BeautifulSoup/

[9] https://www.selenium.dev/

No caso de ***APIs***, é comum que grandes portais web e, especialmente, as grandes redes sociais exponham APIs que permitam a integração e expansão dos seus conteúdos e serviços. Essas APIs geralmente seguem a arquitetura e orientações do padrão Web REpresentational State Transfere (REST)[10], que define uma interface comum (tipicamente humano-legível) para integração e acesso a recursos. Esse acesso é realizado através de Uniform Resource Identifiers (URIs)[11] (*endpoints* fornecidos pelo provedor de serviços). REST Web APIs são baseadas no protocolo HTTP e, portanto, se apoia em seus métodos como GET, POST, PUT, DELETE e OPTIONS. Uma questão associada ao uso das APIs REST é a limitação na quantidade de requisições possíveis. Essa limitação varia de provedor para provedor e de recurso para recurso. Ao utilizar essas ferramentas, dados brutos podem ser coletados, tais como reportagens, colunas, manchetes, opiniões, comentários, etc.

Redes sociais como Twitter, Facebook, Instagram, entre outras, expõem APIs para acesso a seus recursos [Batrinca and Treleaven, 2015]. Por isso, dados dessas redes sociais têm sido tão utilizados para o estudo sobre polarização [Arora et al., 2022; Tucker et al., 2018]. Através das APIs dessas redes sociais é possível coletar dados brutos sobre *posts*, *likes/dislikes*, a rede de amizade, tópicos/assuntos em alta, *hashtags*, etc.

3.2.4. Pré-processamento

As tarefas de modelagem e filtragem dos dados são essenciais para transformar dados brutos em informações úteis. Dados sobre polarização podem não possuir uma organização lógica e hierárquica, relacionamentos ou serem completamente desestruturados, o que dificulta sua manipulação. Neste sentido, o desafio da modelagem dos dados brutos é uma representação uniforme para manipular esses dados e garantir que estão seguindo formatos padrões interpretáveis. Já a tarefa de filtragem lida com a eliminação de dados indesejáveis visto que é comum que dados adquiridos possuam imperfeições, erros de inserções, informações irrelevantes, dentre outras, os quais podem eventualmente adicionar vieses prejudiciais às análises ou inferências [Garimella et al., 2018a; Lu et al., 2015; Pannucci and Wilkins, 2010]. A seguir, apresentamos algumas técnicas de representação dos dados e questões de filtragem úteis para a etapa de pré-processamento dos dados.

3.2.4.1. Modelagem dos dados

A seguir listamos, de modo não exaustivo, técnicas para representações conceituais de dados. Salientamos que a aplicabilidade de cada representação pode variar a depender das finalidades da aplicação. Por esse motivo, destacamos os prós e contras de cada técnica. Estudos mais aprofundados podem ser encontrados em Bettini et al. [2010]; Garimella et al. [2018b]; Santos et al. [2016].

Modelagem dos dados: a modelagem dos dados lida com o problema de representar informação (organizada) a partir de dados brutos. Apresentamos, a seguir, uma lista não exaustiva de técnicas frequentemente utilizadas para modelar dados no contexto de polarização.

[10]https://restfulapi.net/

[11]RFC 3986: Uniform Resource Identifier (URI)

- Chave-valor (*key-value*) é uma estratégia para armazenar e recuperar arranjos associativos de informação (ex., dicionários). Nesta abordagem, pares de informação relacionadas (chave e valor) são os elementos básicos. Por exemplo, a chave pode ser o código do usuário em uma rede social e o valor uma lista de mensagens postadas por esse usuário.

- O *Markup Schema*[12] é outra estratégia que visa propor e manter esquemas para dados estruturados na Internet. Esses esquemas são um conjunto de 'tipos' (ex.: *CreativeWork, Book, Movie, Organization, Person*, etc) e cada tipo tem um conjunto de propriedades definidas (ex.: tipo *Book: bookEdition, bookFormat, illustrator*, et). O interessante é que esses esquemas podem ser utilizados em combinação com diferentes formatos de codificação como, por exemplo, RDF, Microdata, JSON-LD.

- A modelagem por *grafos* é bastante comum nos estudos sobre polarização [Belcastro et al., 2020; Coletto et al., 2017; Conover et al., 2011; Garimella et al., 2018a]. Tipicamente redes podem ser construídas a partir de relações entre pessoas (ex.: amizades em redes sociais), interações com postagens, comentários, *likes*, dentre outras. É comum que dados sejam modelados em grafos usando as seguintes abordagens clássicas: lista de arestas, lista de adjacência e matriz de adjacência. No contexto de polarização, cada nó do grafo é normalmente associado com um atributo que denota o posicionamento (ex: a favor) daquele nó em relação ao objeto de estudo (ex: legalização das drogas).

Além dessas três abordagens, outras também são possíveis, como representação usando ontologias, baseadas em objetos, baseadas em lógicas, entre outras [Bettini et al., 2010; Santos et al., 2016].

3.2.4.2. Filtragem dos dados:

O foco principal da filtragem de dados é a eliminação de dados não relevantes visando melhorar a qualidade da informação e, consequentemente, a qualidade das análises e inferências. Muitas questões relacionadas a filtragem dos dados podem acontecer [Rettore et al., 2016]. A seguir apresentamos uma lista não exaustiva de problemas recorrentes em dados sobre polarização que podem requer alguma técnica de filtragem.

- Granularidade dos dados: é a medida no nível de detalhes do dados coletados. Por exemplo, em uma série temporal referente a polarização, a medida de granularidade pode ser baseada na frequência em que os dados são capturados. Se em alta frequência, então tem-se mais detalhes sobre o comportamento das entidades perante aos polos, se menos frequente, tem-se menos riqueza sobre os o comportamento das entidades em torno dos pontos de estudo. Esse aspecto dos dados é importante no contexto de polarização, pois a partir dele, pode-se realizar inferências sobre entidades individuais (ex.: polarização de um único usuário, uma postagem ou comentário, etc.) ou grupos de entidades (ex.: polarização dos usuários de uma rede

[12] https://schema.org/

social). Em [Barros et al., 2021; Ferreira et al., 2021], os autores apresentam series temporais de dados que mostram o movimento dos indivíduos ao longo do tempo em torno dos polos contrastantes.

- Dados vagos: ocorre em conjunto de dados brutos onde seus atributos não estão bem definidos. Atributos com definição aberta/livre, permitem que dados sejam associados de modos subjetivos como, por exemplo, textos de mensagens na rede social Twitter (*tweets*) podem ser usados como dados brutos sobre a opinião de um usuário da plataforma. O conteúdo textual desse *tweet* pode ser vago e não prover informações suficientes para que seja possível detectar o posicionamento polarizador. Por exemplo, no contexto de do tema polarizador legalização de drogas, ao coletarmos *tweets* de um usuário que só faz comentários sobre outros temas (ex.: comida), não será possível identificar o posicionamento deste usuário sobre o tema legalização das drogas, tornando o *tweet* vago, ou em outras palavras, com baixa precisão, para este tema polarizador. Essa característica de baixa precisão também aparece ao utilizarmos expressões vagas, por exemplo, no texto hipotético *"A população teve um aumento* ***"expressivo/considerável/em torno"*** *em no* ***'tema polarizador'***. Embora possamos extrair a informação de aumento, expressões como *expressivo, considerável* não significam nada, o que torna conclusões sobre essa informação vagas, isto é, com baixa precisão.

- Outliers (anomalias): outro aspecto que pode implicar na necessidade de filtragem dos dados é a presença de *outliers*, os quais são pontos de dados que tipicamente diferem significativamente de outras observações [Grubbs, 1969]. No contexto de dados sobre polarização, anomalias podem aparecer de diferentes formas, desde sua inserção gênesis até na sua coleta ou processamento. Um exemplo de anomalias no contexto de polarização são *bots* que produzem spam ao divulgar opiniões extremas sobre um tema polarizador com o intuito de divulgar um posicionamento.

  Esses dados podem, eventualmente, distorcer as análises ao adicionar viés indesejável.

- Dados incompletos: são observações que possuem um ou mais atributos sem valor. Intuitivamente, essas partes ausentes podem gerar inferências e análises incorretas e, portanto, esses dados com partes faltantes podem ser filtrados. Suponha por exemplo que em um estudo sobre polarização, deseja-se estratificar a polarização por faixas etárias, porém, o atributo idade pode não estar presente para todos os indivíduos. Desta forma, teremos uma análise com dados incompletos.

3.2.5. Armazenamento

Para que a grande quantidade de dados gerados sobre polarização possam ser posteriormente analisada e processada, a etapa de armazenamento faz-se essencial. Esse armazenamento aparece na literatura de duas formas principais: armazenamento local ou em plataformas na nuvem voltadas ao armazenamento e, em alguns casos, processamento dessa massa de dados.

Em ambas as abordagens de armazenamento (local ou nuvem) seria desejável que os dados, logo após coletados, fossem adequados a um modelo facilitando a consulta subsequente. Entretanto, o que acontece com frequência, é um salto da aquisição dos dados

brutos para armazenamento utilizando um modelo mais simples e genérico possível e, eventualmente, esses dados passam por ciclos de pré-processamento (adequação a modelo(s) e filtragem) e armazenagem conforme ilustrado na Figura 3.4. É importante notar que a escolha de um Sistema de Gerenciamento de Banco de Dados (SGBD) e esquemas de para armazenagem de dados são etapas importantes no processo, porém estão fora do escopo deste trabalho.

3.2.6. Extração de conhecimento sobre Polarização

Como pode ser observado na Figura 3.5, os pré-processamentos permitem que os dados brutos sejam transformados em informação relevante sobre polarização a partir de uma estruturação em um modelo de representação comum e operações de filtragem. No nível subsequente da hierarquia, o *viés* refere-se a extração de qualquer informação que pode ser utilizada para caracterizar um eventual posicionamento (viés) perante aos possíveis polos/grupos que uma única entidade (postagem, pessoa, comentário, notícia) pode ter. Já o último nível da hierarquia representa a polarização em si, que geralmente é derivada a partir de diversas informações sobre vieses, proporcionando conhecimento global sobre a temática que apresenta contrastes ou agrupamento de opiniões. A etapa de análise e inferência busca definir a polarização e os vieses a partir dos dados coletados.

No restante deste capítulo serão discutidas técnicas para caracterizar viés e polarização. Para tanto, técnicas de inteligência computacional, estatísticas ou modelos matemáticos serão utilizados.

3.3. Viés de entidades dos dados

Quando olhamos para interações em redes sociais, podemos realizar diferentes tipos de análises com relação ao teor do conteúdo submetido e dos agentes participantes. Esses estudos variam tanto com respeito ao tipo de informação que se deseja estudar quanto ao elemento que se pretende observar. Identificar se uma determinada notícia ou postagem contém conteúdo falso, por exemplo, é uma tarefa de extrema relevância nos tempos atuais. Também podemos verificar se um comentário específico é ofensivo, ou identificar se um determinado usuário é um robô (*bot*) ou uma pessoa real. De forma geral, nos referimos a essas características como o viés dos dados.

A princípio, o estudo do viés de dados envolve um vasto conjunto de problemas relacionados. Análise de sentimentos, extração de opiniões, detecção de ironia, classificação de notícias falsas, mineração de argumentos, dentre outros. Cada um desses problemas representa uma área de estudo com amplas possibilidades, e envolve seus próprios métodos, métricas e modelos. Para este mini-curso, nosso maior interesse está no viés representado pelo posicionamento ou estância de um comentário ou usuário em específico (vide Figura 3.2). A partir dessa informação, poderemos realizar análises mais complexas a respeito de como os conteúdos de uma rede social se relacionam e conflitam entre si, nos levando ao conceito de polaridade.

Esta seção irá explorar o conceito de posicionamento, definindo o problema de detecção de posicionamento, comparando com outros problemas relacionados, e citando como ele é abordado em diferentes trabalhos. Desta forma, a seção se encarrega de apresentar soluções e ferramentas que são capazes de classificar um *tweet* ou comentário, ou

um usuário como um todo, com relação ao seu *viés* (muitas vezes, viés político).

3.3.1. O problema de detecção de posicionamento

A detecção de posicionamento [ALDayel and Magdy, 2021; Küçük and Can, 2020] é uma tarefa que trata de identificar o posicionamento (estância, orientação, apoio) que uma entidade presente nos dados representa com relação a um ou mais alvos (proposições, temas, tópicos). De forma geral, um alvo pode ser qualquer tipo de tema ou assunto a respeito do qual um usuário pode se posicionar. Comumente, esses temas podem representar aspectos ideológicos, decisões políticas, organizações, ou até mesmo a indivíduos específicos.

Um posicionamento, para o tipo de problema tratado neste mini-curso, identifica principalmente se o interlocutor está "*A Favor*" ou "*Contra*" àquela determinada proposição. Muitas vezes também são consideradas opções adicionais, como "*Nenhum*" e/ou "*Neutro*" para representar conteúdos que não se posicionam claramente a favor ou contra o alvo em questão. Nesses casos, uma estância "*Neutra*" representa um conteúdo que especificamente não é unicamente a favor nem contra aquele tema, enquanto "*Nenhum*" pode indicar comentários ou usuários que não se posicionam de forma clara naquele assunto ou simplesmente tratam de assuntos não-relacionados. Assim, formalmente, podemos definir o posicionamento de uma determinada entidade com relação a um tópico pela Equação 1, apresentada anteriormente.

3.3.1.1. Tipo de entidade

Podemos tratar o problema de detecção de posicionamento com relação a diferentes tipos de entidades dos dados nos quais avaliamos as estâncias. Há dois principais níveis de entidades que podemos encontrar sendo utilizados na literatura: declaração e usuário.

Quando aplicamos a detecção de posicionamento sobre uma declaração, nosso objetivo é identificar a orientação descrita em um determinado texto individualmente. Essa é uma tarefa de processamento de linguagem natural, e pode envolver textos mais curtos (sentenças), de tamanho médio (parágrafos, comentários, tweets), até textos de tamanhos mais longos (artigos jornalísticos, relatos).

Alternativamente, podemos detectar posicionamento ao nível de usuário. Nesse tipo de tarefa, deseja-se descobrir se cada usuário é, considerando-se todo o seu comportamento na plataforma digital, a favor ou contra o tema alvo. Isso pode incluir não só os textos de comentários e postagens produzidas por aquele usuário, como também outras informações de seu perfil (idade, gênero, etc.), comunidades e conteúdos com o qual interage, e relações com outros usuários em sua rede social.

3.3.1.2. Tipo de alvo

Também podemos diferenciar os trabalhos de detecção de posicionamento de acordo com o tipo de tópico ao qual os dados se referem. Na definição mais básica, uma entidade expressa posicionamento com relação a um tema. Dessa forma, deve-se construir um classificador separado para cada tema que se deseja identificar no conjunto de dados. Mas

também pode ser analisado como uma entidade se posiciona com relação a diversos temas relacionados simultaneamente. Por exemplo, para um estudo sobre eleições presidenciais, podemos analisar tweets e tentar detectar os posicionamentos de cada um com relação a todos os candidatos possíveis simultaneamente. Isso se torna vantajoso em casos onde os conteúdos dos dados costumam colocar os alvos posicionados relativamente um ao outro.

Existe ainda um terceiro tipo de alvo para o problema, no qual, ao invés de um ou mais temas explícitos, busca-se verificar como comentários em notícias se posicionam com relação a alegações feitas nas notícias em si. Normalmente, nesse tipo de problema, o objetivo é detectar se comentários confirmam ou negam determinadas informações dadas nas notícias, e pode ser usado como base para prever sua veracidade e identificar rumores e notícias falsas.

3.3.1.3. Outros problemas relacionados

Em estudos relacionados a conteúdos de mídias sociais, há diversos aspectos a serem abordados. Dentre esses, alguns se destacam por serem relacionados, e muitas vezes confundidos com a detecção de posicionamento. Para clarificar as diferenças entre esses aspectos e análises, e evitar erros conceituais, vamos agora comparar alguns dos principais problemas relacionados

Um dos principais problemas que são comumente utilizados de forma similar à detecção de posicionamento é a análise de sentimento [Liu, 2010, 2012; Ravi and Ravi, 2015]. Há estudos que utilizam o sentimento expresso em conteúdos como forma de determinar o posicionamento [Li and Caragea, 2019], seja diretamente ou como representação secundária. No entanto há diferenças conceituais fundamentais entre posicionamento e sentimento. Análise de sentimento tem como objetivo determinar a polaridade das emoções expressas em um determinado conteúdo, enquanto a detecção de posicionamento pretende identificar se um conteúdo expressa um ponto de vista a favor ou contra determinados tópicos.

De forma geral, podemos dizer que o sentimento é uma informação extraída do conteúdo em sua forma pura, enquanto um posicionamento é colocado como uma relação entre o conteúdo e o tema alvo. Uma simples frase como “estou nervoso!”, por exemplo, indica uma polaridade de sentimento clara, enquanto não necessariamente indica um posicionamento explícito com relação a qualquer tema por si só. Já uma declaração como “estou feliz que esse filme não fez sucesso” indicaria um sentimento positivo ao mesmo tempo que representa um posicionamento negativo com relação ao filme em questão.

Considerando essas diferenças entre as definições dos dois conceitos, é importante destacar que o uso de métricas de sentimento como um único fator determinante para a consideração de posicionamento não é adequado [Aldayel and Magdy, 2019; Mohammad et al., 2017; Sen et al., 2020]. No entanto, isso não quer dizer que não haja relação e interação entre sentimento e posicionamento, ou que não seja possível utilizar os dois conceitos em conjunto para compor uma análise [Tachaiya et al., 2021].

Outros problemas relacionados à detecção de posicionamento, porém distintos, incluem: reconhecimento de emoções [Canales and Martínez-Barco, 2014; Sailunaz et al.,

2018], no qual são identificadas emoções presentes em um conteúdo dentre um conjunto de classes de emoções; detecção de controvérsia [Coletto et al., 2017], que busca medir e identificar tópicos controversos em um conteúdo; previsão de posicionamento [Darwish et al., 2018; Dong et al., 2017], que se preocupa em estimar como usuários (ou grupos de usuários) se posicionariam com relação a temas dos quais esse posicionamento não foi observado, ao invés de detectar um posicionamento já explícito no conteúdo.

3.3.2. Abordagens de detecção de posicionamento

Tendo em vista o contexto mais específico de detecção de posicionamento, conforme definido anteriormente, podemos explorar os diferentes métodos utilizados para essa tarefa. A seguir, apresentamos uma visão geral dos atributos comumente usados nos modelos, assim como os algoritmos aplicados.

3.3.2.1. Atributos

Para abordar a detecção de posicionamento, há uma multitude de sinais que podem ser utilizados, dependendo do tipo de dado estudado. Isso pode incluir desde atributos extraídos diretamente do conteúdo até características da rede social em si, passando por representações geradas por modelos e análises de escolha de vocabulários.

Considerando o conteúdo textual de uma declaração diretamente, há trabalhos que utilizam modelagens dos termos em *bag-of-words* ou n-gramas como um conjunto primário de atributos [Mohammad et al., 2017], assim como outros indicadores como pontuação e tamanho do texto [Kochkina et al., 2017; Lai et al., 2017]. Além de características extraídas diretamente do texto, também é possível considerar métricas como a polaridade de sentimento do conteúdo [Ebrahimi et al., 2016], ou representações do conteúdo em forma de tópicos latentes [Elfardy and Diab, 2016].

Também há trabalhos que se voltam para como o vocabulário difere entre grupos com posicionamentos opostos [Darwish et al., 2020]. Com esse tipo de análise, pode-se por exemplo detectar a perspectiva de certos usuários com relação a determinados temas ao se observar como se comunicam no geral [Beigman Klebanov et al., 2010].

Além disso, o uso de características e atributos retirados a partir da rede também pode ser efetivo. Alguns trabalhos, por exemplo, modelam representações com base no texto do conteúdo em conjunto com características das interações do usuário na rede [Li et al., 2018], assim como atributos indicando as conexões e interações entre usuários na rede [ALDayel and Magdy, 2021], ou outras características retiradas das plataformas como hashtags, re-tweets, URLs, e menções a outros usuários [Darwish et al., 2017; Hamidian and Diab, 2019].

3.3.2.2. Algoritmos

A partir dos atributos extraídos dos conteúdos ou usuários, os estudos de detecção de posicionamento empregam diferentes métodos de aprendizado de máquina para identificar as estâncias daqueles elementos com relação aos temas em questão.

Métodos supervisionados são frequentemente utilizados para esse problema. Por essa abordagem, os dados relativos aos conteúdos ou usuários são anotados de acordo com seus posicionamentos. Normalmente, essas bases de dados são anotadas por especialistas de acordo com rótulos como "A Favor", "Contra"e "Nenhum", como aquela apresentada na tarefa de detecção de posicionamento em *SemEval-2016* [Mohammad et al., 2016]. A partir desses dados rotulados, algoritmos são treinados para aprender os padrões que indicam cada posicionamento. Trabalhos como Elfardy and Diab [2016] e Li and Caragea [2019] exemplificam esse tipo de abordagem, o primeiro com uso de atributos léxicos e semânticos em uma SVM (*Support-Vector Machine*) e o segundo com métodos de aprendizado profundo com uma arquitetura GRU (*Gated Recurrent Unit*).

No entanto, como esses dados rotulados são custosos para se produzir e de difícil obtenção, abordagens semi-supervisionadas e não-supervisionadas também são propostas. O trabalho de Ferreira and Vlachos [2019], por exemplo, usa técnicas de transferência de aprendizado para reutilizar o conhecimento que o algoritmo aprendeu em uma base de dado como ponto de partida para a detecção de posicionamento em dados de outras fontes. Já Zhang et al. [2020] tomaram diversas bases de dado em conjunto para detectar posicionamentos de forma cruzada em tarefas voltadas para temas de sub-grupos diferentes.

Métodos totalmente não-supervisionado também vem sido propostos, primariamente gerando representações dos usuários e conteúdos e aplicando métodos de agrupamento sobre elas. Darwish et al. [2020], por exemplo, aplicaram uma técnica de agrupamento sobre tweets não-rotulados de diferentes tópicos como ponto inicial para anotação dos dados. Outros como Rashed et al. [2020] usaram representações distribuídas de tweets com agrupamento hierárquico para análise de polarização política.

De forma geral, as possibilidades de algoritmos e técnicas para tarefa de detecção de posicionamentos é bastante ampla. Trabalhos como o de Swami et al. [2018] e Tsakalidis et al. [2018] utilizam SVMs. Outros como Kucher et al. [2018] e Ferreira and Vlachos [2016] aplicam regressão logística. Abordando técnicas de aprendizado profundo, são encontradas também diversas possibilidades, incluindo redes LSTMs (*Long Short-Term Memory*) [Rajendran et al., 2018], CNNs (*Convolutional Neural Network*) [Hercig et al., 2017], além de trabalhos utilizando mecanismos de atenção, como aqueles aplicando modelos baseados BERT [Ghosh et al., 2019; Kawintiranon and Singh, 2021]. E de forma geral, o uso de técnicas de agrupamento de modelos também são efetivas [Liu et al., 2016; Siddiqua et al., 2019].

3.4. Métricas de polarização

Como vimos na seção anterior, calcular o viés de um usuário ou conteúdo está ligado a verificar o posicionamento daquela entidade. Por exemplo, um comentário a favor de um tópico ou um usuário que declarou sua posição política. Já a polarização é uma medida de um grupo de usuários (ou um conjunto de conteúdos). Aqui estamos interessados em verificar como uma população está distribuída em torno de um tópico específico. Por exemplo: Em uma universidade a reitora A está passando por uma crise de popularidade. Foi então realizada uma pesquisa que revelou o viés de cada aluno. Destes, 40% dos alunos acreditam que a reitora tem feito um bom mandato, 45% que ela faz um mandato

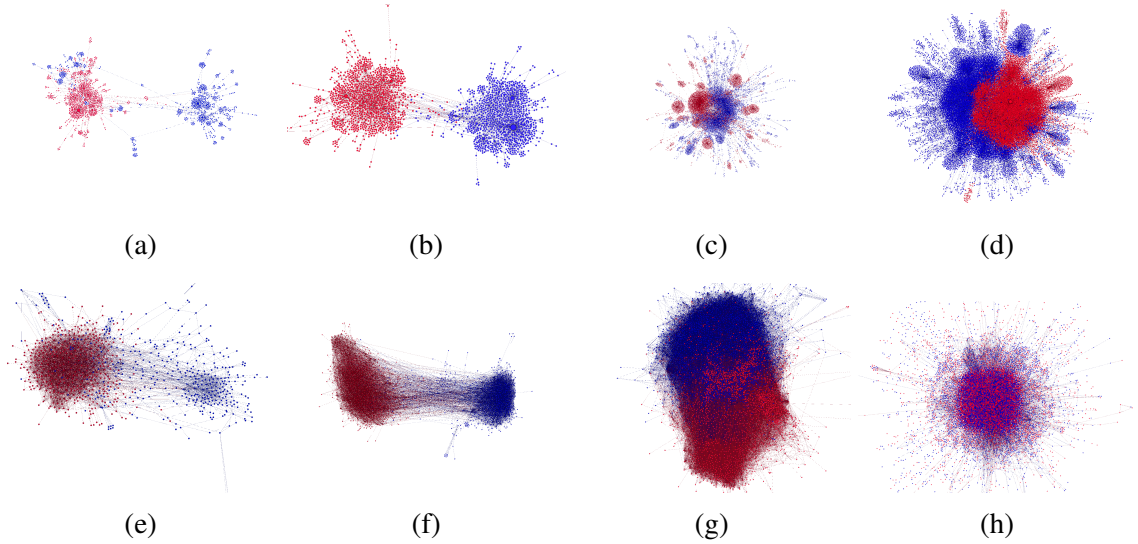

Figura 3.6: Exemplos de redes polarizadas (esquerda) e não polarizadas (direita). Os grafos superiores representam *retweets*, já os grafos inferiores representam *follow*. Mais informações sobre os dados coletados podem ser encontradas em Garimella et al. [2018b].

ruim e 15% ficaram neutros. Neste exemplo observamos que os alunos estão divididos em dois grandes grupos que discordam entre si (e um grupo menor que não tem uma opinião bem formada). Podemos então dizer que a universidade em questão está polarizada com relação ao mandato da reitora?

Outra forma de visualizar (e medir) a polarização é por meio do grafo da rede formado em torno de um tópico. Essa técnica é bastante utilizada no contexto das redes sociais, onde os usuários mantém ligações de amizade entre si, compartilhando e reagindo a postagens de outros usuários. Neste caso, cada usuário é um vértice do grafo e as arestas são formadas por meio das conexões da rede, como amizade e compartilhamento.

Na Figura 3.6 visualizamos alguns exemplos de redes (grafos) polarizadas e não polarizadas. As redes representadas em (a), (b), (e) e (f) são exemplos de redes polarizadas. Note que, em todas elas, temos dois grupos muito bem definidos e com poucas ligações (arestas) ligando seus vértices. Ainda na Figura 3.6, podemos observar exemplos de redes não polarizadas em (c), (d), (g) e (h). Observe que nos exemplos de redes não polarizadas temos os membros dos diferentes grupos com um número maior de ligações entre si, logo, menos isolados e com maior probabilidade de ter acesso a informações "do outro lado da rede".

Na literatura, não existe um consenso de como operacionalizar a quantificação da polarização de uma população. A maioria dos trabalhos podem ser caracterizados como estudos de caso, onde a polarização é identificada em bases de dados específicas e analisadas utilizando conhecimento do domínio (ex: lista de *hashtags* relacionadas a um evento político específico) [Garimella et al., 2018b]. Boa parte destas métricas foram desenvolvidas no contexto de redes sociais e são calculadas por meio de sua rede de comunicação [Conover et al., 2011; Garimella et al., 2018b; Guerra et al., 2013; Coletto et al., 2017; Matakos et al., 2017], outras são independentes da rede [Bakshy et al., 2015;

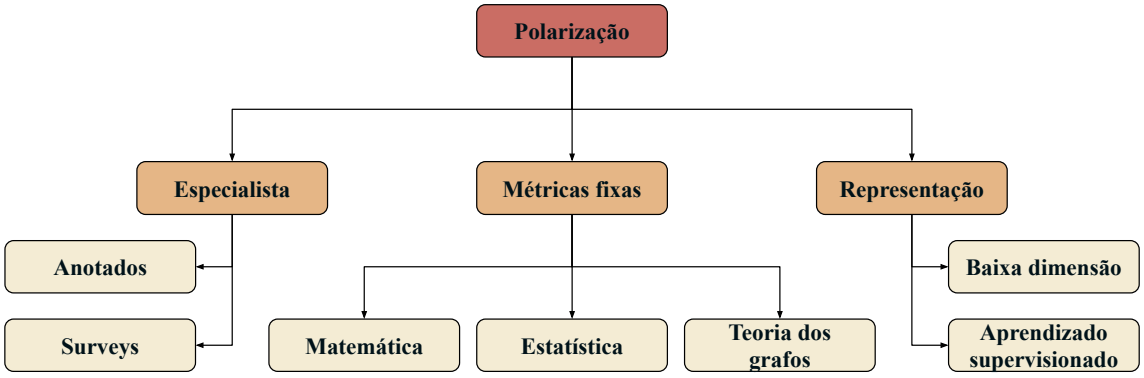

Figura 3.7: Taxonomia das métricas de polarização.

Belcastro et al., 2020; Roy and Goldwasser, 2020; Babaei et al., 2018; Morales et al., 2015; Vicario et al., 2019].

Uma métrica de polarização que não depende da estrutura da rede possui a vantagem de poder ser aplicada em cenários onde a estrutura da rede é desconhecida [Morales et al., 2015; Al-Ayyoub et al., 2018]. Tomemos como exemplo a caixa de comentários de uma famosa página de notícias. Neste caso, as opiniões ali deixadas tem como único ponto em comum a notícia em questão. Não temos informações sobre quais usuários possuem relações de amizade entre si ou ainda a qual viés político ele se alinha. No caso deste exemplo, precisamos de uma métrica de polarização que seja independente da estrutura da rede.

Em outros casos podemos usufruir de informações presentes na rede, como conexões de amizade ou de endosso (*retweet*, por exemplo). Os primeiros trabalhos analisaram a polarização em redes sociais utilizando a modularidade do grafo [Conover et al., 2011]. A modularidade de uma rede quantifica o quanto vértices conectam entre si formando comunidades densas, quando comparado a uma rede aleatória [Newman, 2006]. Outras análises procuram por padrões na rede, como os clusters de usuários [Garimella et al., 2018b], a fronteira entre os grupos [Guerra et al., 2013], entre outras estruturas [Coletto et al., 2017; Garimella et al., 2021]. Nestes casos estamos preocupados com as estruturas geradas na comunidade e como a comunidade está configurada.

Visando organizar as diferentes ferramentas utilizadas na literatura para identificar/quantificar tópicos polarizados, a Figura 3.7 apresenta uma proposta de taxonomia para as métricas de polarização. Dividimos as soluções em três tipos: (1) trabalhos que analisaram a polarização pela visão de um **especialista**, ao anotar os tópicos manualmente [Dori-Hacohen and Allan, 2015; Jang et al., 2016] ou por meio de questionários (*surveys*) [Ribeiro et al., 2019]. (2) Trabalhos que utilizaram **métricas fixas**, como fórmulas matemáticas [Al-Ayyoub et al., 2018; Belcastro et al., 2020], estatísticas [Jang and Allan, 2018; Morales et al., 2015] ou de teoria dos grafos [Garimella et al., 2018b; Guerra et al., 2013], para cálculo da polarização. E (3) trabalhos que utilizaram **representações** em baixa dimensionalidade [Waller and Anderson, 2021] e/ou algoritmos de aprendizado supervisionado [Roy and Goldwasser, 2020; Vicario et al., 2019] para identificação dos tópicos polarizados.

Tabela 3.1: Compilado de trabalhos relacionados em identificação de tópicos polarizados.

Trabalho	Conteúdo	Rede	Ident.	Quant.	Ferramenta
Al-Ayyoub et al. [2018]	✓		✓	✓	Matemática
Akhtar et al. [2019]	✓		✓	✓	Estatística
Babaei et al. [2018]	✓		✓	✓	Matemática
Belcastro et al. [2020]	✓		✓		Matemática
Choi et al. [2010]	✓		✓		Estatística
Dori-Hacohen and Allan [2015]	✓		✓		Anotado
Jang et al. [2016]	✓		✓		Anotado
Jang and Allan [2018]	✓		✓		Estatística
Klenner et al. [2014]	✓		✓		Aprend. superv.
Mejova et al. [2014]	✓		✓		Aprend. superv.
Morales et al. [2015]	✓		✓	✓	Estatística
Popescu and Pennacchiotti [2010]	✓		✓		Aprend. superv.
Ribeiro et al. [2019]	✓		✓		Survey
Roy and Goldwasser [2020]	✓		✓		Aprend. superv.
Tsytsarau et al. [2011]	✓		✓		Estatística
Vicario et al. [2019]	✓		✓	✓	Aprend. superv.
Waller and Anderson [2021]	✓		✓	✓	Baixa dimensão
Yang et al. [2017]	✓		✓	✓	Baixa dimensão
Akoglu [2014]		✓	✓		Estatística
Al Amin et al. [2017]		✓	✓		Aprend. superv.
Coletto et al. [2017]		✓	✓		Teoria dos grafos
Garimella et al. [2018b] (RWC)		✓	✓	✓	Teoria dos grafos
Garimella et al. [2018b] (BCC)		✓	✓	✓	Teoria dos grafos
Garimella et al. [2018b] (EC)		✓	✓	✓	Baixa dimensão
Garimella et al. [2021]		✓	✓		Teoria dos grafo
Gillani et al. [2018]		✓	✓		Estatística
Guerra et al. [2013]		✓	✓	✓	Teoria dos grafos
Shahrezaye et al. [2019]		✓	✓	✓	Teoria dos grafos
Tokita et al. [2021]		✓	✓	✓	Teoria dos grafos

Um compilado dos trabalhos relacionados pode ser visto na Tabela 3.1. Como vimos, os trabalhos são classificados entre aqueles que utilizam informação de conteúdo ou rede. Ainda, somente uma parcela dos trabalhos é capaz de quantificar o nível de polarização presente em um tópico.

Nas próximas seções iremos apresentar em detalhes algumas das principais métricas de polarização utilizadas na literatura. As seções 3.4.1 e 3.4.2 apresentam métricas de quantificação de polarização baseadas em conteúdo, isto é, não levam o grafo da rede em consideração. Já as métricas de polarização apresentadas nas seções 3.4.3, 3.4.4, 3.4.5 utilizam informações da rede para seu cálculo.

3.4.1. Utilizando análise de sentimentos para medir polarização

A primeira métrica de polarização que vamos abordar foi originalmente concebida para trabalhar com análise de sentimentos de tweets. Abordaremos aqui o trabalho de Al-

Ayyoub et al. [2018] que apresenta um conjunto de métricas matemáticas simples. Estas métricas podem ser utilizadas individualmente ou coletivamente na tarefa de análise de polarização, como mostraremos no decorrer da seção.

Razão da quantidade de tweets positivos e negativos (PN). Esta métrica se baseia na premissa de que, em um tópico polarizado, um dos grupos provavelmente utilizaria mensagens que demonstram sentimentos de aprovação ao tópico, e, o outro, mensagens de reprovação. Por consequência, esperamos observar um número de mensagens com sentimentos positivos no mesmo nível de mensagens negativas. Um maior valor para a razão PN representa uma maior grau de polarização. Também podemos utilizar a razão de tweets negativos e tweets positivos como uma métrica oposta a esta.

$$\text{PN} = \frac{|\text{tweets positivos}|}{|\text{tweets negativos}|} \tag{2}$$

Razão entre a quantidade de tweets positivos e negativos (RPN). A RPN é uma melhoria da PN vista anteriormente. Realizando sempre a razão entre o menor valor (tweets positivos ou negativos) e o maior valor, teremos como resultado um valor entre 0 e 1. Sendo que valores maiores representam uma maior polarização.

$$\text{RPN} = \frac{\min\{|\text{tweets positivos}|, |\text{tweets negativos}|\}}{\max\{|\text{tweets positivos}|, |\text{tweets negativos}|\}} \tag{3}$$

Razão entre a quantidade de tweets neutros e a quantidade de tweets positivos e negativos (NPN). Esta métrica se baseia na premissa de que, em um tópico polarizado, temos um pequeno número de mensagens neutras em comparação ao número de mensagens com viés claro. Isso aconteceria pois um número maior de usuários tenderia a demonstrar claramente sua aprovação ou desaprovação do tópico em questão. Nesta métrica, valores menores correspondem a uma maior polarização.

$$\text{NPN} = \frac{|\text{tweets neutros}|}{|\text{tweets positivos}| + |\text{tweets negativos}|} \tag{4}$$

Razão entre a soma dos tweets positivos e negativos com a quantidade total de tweets (PNT). Ainda com a premissa de que tópicos polarizantes obtêm um maior número de comentários com viés claro (positivo ou negativo), a métrica PNT calcula a razão entre a soma dos tweets positivos e negativos sobre o número total de tweets naquele tópico. Seu resultado varia entre 0 e 1, sendo que valores maiores representam uma maior polarização.

$$\text{PNT} = \frac{|\text{tweets positivos}| + |\text{tweets negativos}|}{|\text{tweets totais}|} \tag{5}$$

Métrica PN ponderada pela métrica PNT (PNPNT). Esta métrica é uma combinação entre a métrica PN e a métrica PNT. Ela calcula a razão entre os tweets positivos e negativos levando em consideração a razão entre os tweets com viés claro e o número total de tweets naquele tópico. Como a métrica PNT varia entre 0 e 1, o resultado da métrica PNPNT também varia entre 0 e 1, com valores maiores representando uma maior polarização.

$$\text{PNPNT} = \text{PN} \times \text{PNT} \tag{6}$$

Razão entre os valores de sentimentos positivos e negativos (RPNV). A RPNV é uma modificação da RPN vista anteriormente e leva em consideração os valores de viés calculados para cada tweet. Na RPNV calculamos a razão entre os valores de viés positivos e negativos e não o número de tweets. Valores próximos a 1 implicam que as opiniões contrárias possuem valores próximos, ou seja, que o valor total absoluto dos comentários positivos e negativos estão próximos. Os resultados da métrica RPNV variam entre 0 e 1, sendo que valores maiores representam uma maior polarização.

$$\text{RPNV} = \frac{\min\{\sum \text{valores positivos}, \sum \text{valores negativos}\}}{\max\{\sum \text{valores positivos}, \sum \text{valores negativos}\}} \tag{7}$$

Como podemos observar, as diversas métricas mostradas nesta seção capturam informações diferentes entre si. Cabendo ao leitor escolher a melhor para o seu trabalho ou ainda utilizar várias delas em conjunto. Como as métricas são fáceis de serem calculadas, elas são uma ótima opção para serem utilizadas como *features* para um algoritmo de aprendizado supervisionado. Podemos, por exemplo, elaborar um classificador automático de polarização em grupos de *WhatsApp* utilizando os sentimentos das mensagens que trafegam por ali. O trabalho de Al-Ayyoub et al. [2018] utiliza esse conjunto de métricas como entrada para uma *Support Vector Machine (SVM)* e, assim, classificar tópicos de discussão no Twitter. Além das métricas descritas nesta seção, o trabalho citado também utiliza a métrica *Dipole Moment (DM)* que será discutida na próxima seção.

Outro ponto importante é que as métricas podem ser facilmente adaptadas para outros cenários, como postagens em uma página de jornal ou votação de políticos na câmara de deputados. Basta a adaptação da métrica que calcula o viés de cada entidade. Por exemplo, podemos utilizar palavras-chaves ou ainda a opinião de um especialista.

3.4.2. Momento do dipolo elétrico

Nesta seção abordaremos a métrica introduzida pelo trabalho de Morales et al. [2015]. A métrica inspirada no momento do dipolo elétrico (em inglês, *Dipole Moment (DM)*) tem como objetivo capturar o quão dividido encontram-se os membros de uma população. Para isso parte-se da premissa de que uma população é perfeitamente polarizada se ela pode ser dividida em dois grupos de mesmo tamanho e com as opiniões de seus indivíduos concentradas nos extremos.

Sua inspiração vem da física com uma métrica que calcula a polarização das cargas de um sistema eletromagnético. Para isso ela calcula o grau de separação de cargas positivas e negativas que fazem parte do sistema. Um caso simples é onde temos somente duas cargas, uma negativa e uma positiva ($-q$ e $+q$), o momento do dipolo elétrico é proporcional à distância entre essas duas cargas. Esse caso é análogo a um cenário simples onde temos duas pessoas de opiniões contrárias acerca de um tópico. Logo, a polarização deste pequeno grupo pode ser calculada como o quão distante as opiniões destas duas pessoas se encontram.

Seja X uma variável aleatória que modela a distribuição do viés de uma população acerca de um tópico e X_i o viés de um usuário i de modo que $-1 \leq X_i \leq +1$. Então temos $p(X)$ como uma função de densidade da opinião dos usuários. Primeiramente, iremos calcular o tamanho das populações associadas a cada opinião (negativa e positiva). Seja

A^- a população com viés negativo ($X < 0$), calculamos seu tamanho como a área sob a curva da função de densidade $p(X)$ no intervalo $[-1,0)$ (equação 8). De maneira análoga, calculamos A^+ como a área sob a curva de $p(X)$ no intervalo $(0,+1]$ (equação 9).

$$A^- = \int_{-1}^{0} p(X)dX = P(X < 0), \tag{8}$$

$$A^+ = \int_{0}^{1} p(X)dX = P(X > 0) \tag{9}$$

De posse do tamanho dos grupos, estamos interessados em calcular a diferença absoluta entre eles. Que é facilmente calculada como podemos observar na equação 10. Esta fórmula dá como resultado $\Delta A = 0$ quando a população está perfeitamente dividida em dois grupos de tamanhos iguais ($A^- = A^+$). Por outro lado, $\Delta A = 1$ quando todos os elementos da população concordam entre si ($A^- = 1$ ou $A^+ = 1$).

$$\Delta A = |A^+ - A^-| = |P(X > 0) - P(X < 0)| \tag{10}$$

Em seguida, calculamos o quão distante estão as opiniões de ambos os grupos. Para isso, é calculado o centro de gravidade dos viéses negativos (gc^-) e positivos (gc^+), como podemos observar nas equações 11 e 12.

$$gc^- = \frac{\int_{-1}^{0} p(X)XdX}{\int_{-1}^{0} p(X)dX}, \tag{11}$$

$$gc^+ = \frac{\int_{0}^{1} p(X)XdX}{\int_{0}^{1} p(X)dX} \tag{12}$$

E então calculamos a distância entre as opiniões centrais de cada grupo como a diferença absoluta dos centros gravitacionais gc^- e gc^+, como podemos observar na equação 13. Esta fórmula dá como resultado $d = 0$ quando ambos os grupos concordam integralmente entre si. Por outro lado, teremos $d = 1$ quando ambos os grupos discordam entre si e suas opiniões se concentram nos extremos.

$$d = \frac{|gc^+ - gc^-|}{|X_{\max} - X_{\min}|} = \frac{|gc^+ - gc^-|}{2} \tag{13}$$

Por fim, calculamos o índice de polarização DM à partir dos valores calculados de ΔA e d(equações 10 e 13), como podemos observar na equação 14. Pela equação observamos que a métrica DM é inversamente proporcional à diferença absoluta entre as duas populações ΔA e diretamente proporcional à distância absoluta dos centros de gravidade d. Como o resultado de ΔA e d se encontram no intervalo $[0,1]$, o resultado de DM também se encontra no intervalo $[0,1]$.

$$DM = (1 - \Delta A)d \tag{14}$$

Teremos polarização máxima ($DM = 1$) quando a população estiver perfeitamente dividida ($\Delta A = 0$) e as opiniões discordantes estiverem nos extremos ($d = 1$). Por outro

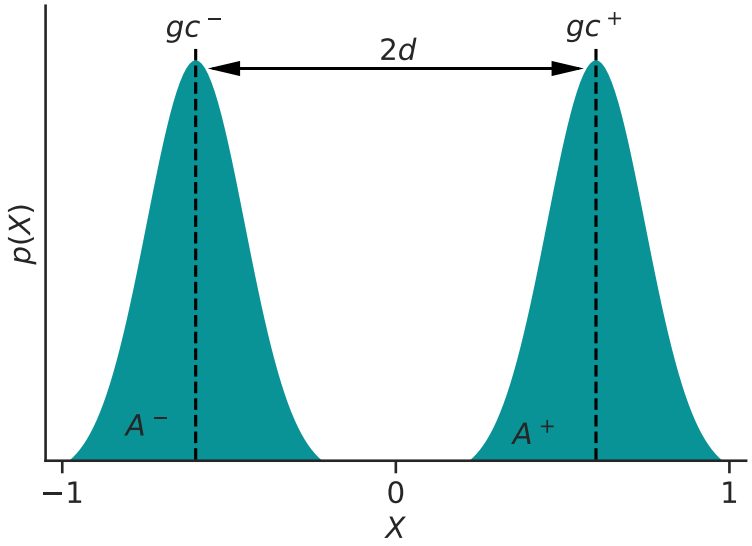

Figura 3.8: Representação da polarização e da métrica DM. Função de densidade de opinião. A representa a área associada a cada ideologia, gc representa o centro gravitacional de cada uma das opiniões e d representa a distância das opiniões.

lado, teremos polarização mínima ($DM = 0$) quando a opinião da população como um todo estiver concentrada em um único ponto, ou seja, quando não há discordância. A métrica poderá apresentar valores entre 0 e 1 quando as populações tiverem tamanhos desiguais ($0 < \Delta A < 1$) e/ou quando a distância entre os viéses das populações for menor que 1 ($0 < d < 1$). A Figura 3.8 ilustra os principais conceitos da métrica DM. Observamos a área que representa o tamanho de cada grupo com seu respectivo viés, assim como a distância entre cada centro de gravidade.

A métrica do momento do dipolo elétrico é muito útil quando o viés dos usuários (ou conteúdos) puderem ser quantificados (valores entre -1 e +1)e o grafo da rede de comunicação não for conhecido. A métrica é de fácil implementação e baixo custo computacional. Além disso, a função de densidade de opinião dos usuários pode ser uma ótima visualização da polarização da população. Uma análise desta função nos dá uma visão de como o tópico em questão foi recebido pela população e qual é a opinião predominante.

3.4.3. Conectividade na fronteira entre grupos antagônicos

A estrutura de uma rede social é afetada pelo contexto e comportamento dos usuários [Easley and Kleinberg, 2010]. Padrões de comportamento, como homofilia [McPherson et al., 2001], alteram a probabilidade que dois usuários se conectem. Em uma rede polarizada, é esperado encontrar padrões que representem a divisão de uma população. Um desses padrões é o antagonismo, isto é, conjuntos de usuários que não apresentarão laços (amizade, compartilhamento, etc) entre si.

Partindo dessa premissa, foram elaboradas métricas de polarização que utilizam informações da rede de modo a extrair informações topológicas do grafo da rede. O primeiro exemplo desse tipo de métrica de polarização é a métrica de *Conectividade de*

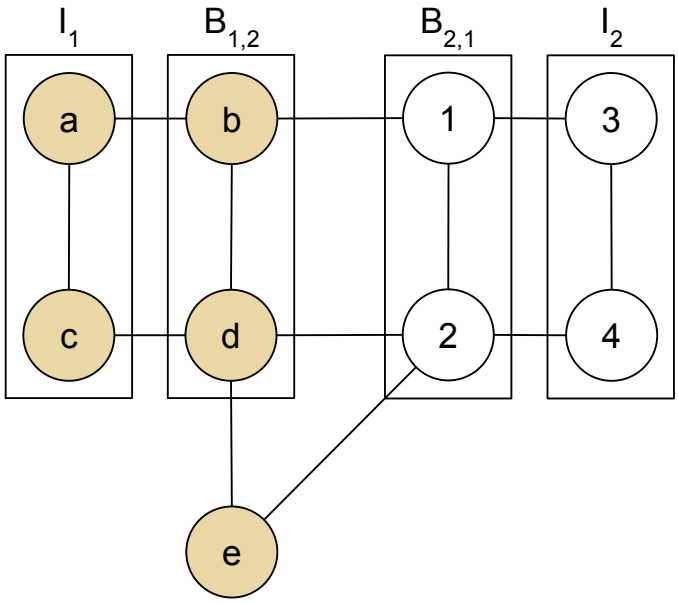

Figura 3.9: Exemplo de uma rede dividida em dois conjuntos G_1 (nós coloridos) e G_2 (nós brancos). Fonte: Guerra et al. [2013].

Fronteira (em inglês, *Boundary Connectivity* – BC) [Guerra et al., 2013]. Essa métrica foca sua análise nos nós que possuem alguma interação com nós do suposto grupo oposto, que aqui denominamos de nós de fronteira.

Seja $G = G_i \cup G_j$ o grafo que representa uma rede dividida em dois conjuntos G_i e G_j ($G_i \cap G_j = \emptyset$). Definimos fronteira de comunidade do grupo G_i, como o conjunto de nós $B_{i,j}$ que satisfaz duas condições:

1. $v \in G_i$ possui ao menos uma conexão com um nó do grupo G_j;
2. $v \in G_i$ possui ao menos uma conexão com um nó do grupo G_i que não possua conexão com os nós de G_j.

Na figura 3.9 temos um exemplo de uma rede dividida em dois grupos G_1 e G_2. Neste exemplo, temos as fronteiras $B_{1,2} = \{b,d\}$ e $B_{2,1} = \{1,2\}$. É importante ressaltar que o nó $c \notin B_{1,2}$ pois, apesar do nó c ter aresta com um nó do outro grupo (nó 2), ele não tem aresta com nenhum nó de G_1 que possua conexão com um nó em G_2. Ainda, nós em G_i que não pertencem à $B_{i,j}$ formam o conjunto de nós internos $I_i = G_i - B_{i,j}$. No exemplo, os nós internos são $I_1 = \{a,c,e\}$ e $I_2 = \{3,4\}$.

Vamos então nos ater aos nós que compõem a fronteira de modo a comparar o grau de preferência destes nós de se conectarem com nós internos ou com nós do outro conjunto. Voltando à Figura 3.9, vamos analisar as conexões do exemplo partindo do nó b. O nó b possui grau três e suas arestas são:

1. $(b,1)$ é uma aresta externa (liga nós de fronteiras opostas).
2. (b,a) é uma aresta interna (liga nós da fronteira a nós internos).

3. (b,d) não é nem uma aresta interna nem externa.

Olhando para as conexões do nó b, ele não nos parece apresentar nenhum tipo de antagonismo, uma vez que se conecta a uma aresta interna mas também a uma aresta externa. Essa mesma análise pode ser extrapolada para os demais nós da fronteira (nós d, 1 e 2). Logo, baseado nos nós da fronteira da rede, o exemplo da Figura 3.9 não possui polarização.

A equação 15 define a métrica *BC* que leva em consideração as escolhas que nós de $B_{i,j}$ fazem ao se conectar com nós de I_i ou $B_{j,i}$. Para cada nó v pertencente à fronteira B ele calcula a razão entre o número de arestas internas que ele possui – $d_i(v)$ – com o número de suas arestas externas – $d_b(v)$ – somadas ao número de suas arestas internas. Essa razão é comparada com a hipótese nula de que cada nó da fronteira tem a mesma probabilidade de possuir arestas com nós internos e nós externos.

$$BC = \frac{1}{|B|} \sum_{v \in B} \left[\frac{d_i(v)}{d_b(v) + d_i(v)} - 0.5 \right] \tag{15}$$

A métrica *BC* poderá apresentar valores entre $-1/2$ e $+1/2$. Um valor *BC* menor do que 0 indica não somente a falta de polarização mas também que um nó na fronteira é mais provável de se conectar com nós do grupo oposto. No exemplo da figura 3.9 temos $BC = 0$ uma vez que os nós da fronteira possuem o mesmo número de arestas com nós internos e com nós do grupo oposto.

A métrica BC apresenta a vantagem de utilizar informações de interação e comunicação entre os usuários, diferentemente das métricas vistas até agora. Além disso, seu foco na fronteira entre os grupos, mensura a interação entre os antagônicos, ou seja, a métrica nos traz uma visão sobre o nível de troca de informações entre os diferentes grupos.

3.4.4. Centralidade do corte da rede

Nesta seção iremos abordar a métrica de centralidade do corte da rede (em inglês, *Betweeness Centrality Controversy score* – BCC) [Garimella et al., 2018b]. Essa métrica vai analisar o conjunto de arestas presentes no corte que se forma ao particionar um grafo em dois grupos opostos X e Y. Para isso, utilizamos a noção de centralidade de arestas (em inglês, *edge betweeness*) e como a centralidade do corte difere das demais arestas.

A centralidade de uma aresta e é calculada pelo que chamamos de "intermediação" (em inglês, *shortest-path betweeness centrality* – $bc(e)$) [Brandes, 2008]. A equação 16 define a intermediação de uma aresta e, onde $\sigma_{s,t}$ é o menor caminho entre s e t, e $\sigma_{s,t}(e)$ é o menor caminho que necessariamente passa pelo vértice e.

$$bc(e) = \sum_{s \neq t \in V} \frac{\sigma_{s,t}(e)}{\sigma_{s,t}}, \tag{16}$$

Consideremos um grafo $G = (V,E)$ polarizado com grupos X e Y opostos e bem definidos ($X \cup Y = G$ e $X \cap Y = \emptyset$). Neste caso, o conjunto de arestas $C \subseteq E$ do corte $C = (X,Y)$ agiria como uma "ponte" levando informações entre os grupos. Fica então

fácil imaginar que o caminho mínimo entre pares de vértices de grupos opostos devem conter alguma aresta do conte C. O que leva a valores altos de intermediação (eq. 16) das arestas em C. Por outro lado, quando pensamos em um grafo onde os grupos não estejam fortemente separados, entendemos que existirão outras arestas pelas quais a informação pode passar. O que leva a valores de intermediação das arestas em C que sejam similares ao restante das arestas do grafo.

Vamos então transformar essa ideia em uma métrica que compara a distribuição do cálculo de centralidade das arestas do corte com a centralidade das demais arestas do grafo. Para isso é computada a divergência de Kullback-Leibler (d_{KL}) [Thomas and Joy, 2006] entre a distribuição de centralidade dos dois conjuntos de arestas. A divergência de KL é uma medida de distância entre duas distribuições de probabilidade (mais detalhes estão fora do escopo deste texto). A métrica *BCC* pode ser vista na equação 17 e é calculada como a inversa da divergência KL.

$$BCC = 1 - e^{-d_{KL}}, \tag{17}$$

O valor de *BCC* estará mais perto de 1 para tópicos polarizados e perto de 0 para tópicos não polarizados.

3.4.5. Random Walk Controversy (RWC)

A última métrica de polarização que iremos abordar almeja mensurar a facilidade do acesso da informação por usuários de grupos opostos. Se supormos uma rede polarizada com poucas ligações entre os grupos, o acesso ao grupo oposto será dificultado. Ao contrário, em uma rede não polarizada temos um número maior de arestas ligando membros de grupos distintos, e um maior tráfego de informação entre os grupos.

Utilizando caminhamento aleatório, a métrica *Random Walk Controversy* (RWC) [Garimella et al., 2018b] pretende calcular a probabilidade de um usuário acessar informações do grupo oposto. Considerando dois usuários que irão caminhar aleatoriamente sobre a rede, a métrica RWC (eq. 18) é definida como a diferença de probabilidade de que ambos terminem no grupo em que começaram e a probabilidade de que ambos terminem em grupos opostos aos quais começaram.

$$RWC = P_{XX}P_{YY} - P_{XY}P_{YX}, \tag{18}$$

onde $P_{AB}, A, B \in \{X, Y\}$ é a seguinte probabilidade condicional:

$$P_{AB} = P[\text{começou na partição A}|\text{terminou na partição B}]. \tag{19}$$

O resultado da métrica RWC ficará próximo a 1 quando a probabilidade de cruzar os grupos é baixa, o que implica uma alta polarização. Por outro lado, o resultado da métrica ficará próximo a 0 quando a probabilidade de cruzar os grupos é comparável a de terminar do mesmo lado, o que implica uma baixa polarização.

3.5. Análises sobre polarização

Chegamos no último estágio da metodologia que é a análise da polarização à partir dos dados coletados e processados. Como já discutido, a polarização é uma métrica amplamente utilizada em ambientes políticos, mas outros cenários também são possíveis, como

esportes [Guerra et al., 2013], discurso de ódio [Almerekhi et al., 2020], desinformação [Vicario et al., 2019; Watts et al., 2021], entre outros.

Logo, a análise da polarização deve levar em consideração as particularidades das áreas e os objetivos a serem alcançados. Nesse sentido, esta seção discute alguns dos tópicos de pesquisa encontrados na literatura. A seguir veremos uma lista dos tópicos abordados nesta seção.

1. *Baselines* para validação e comparação de métricas de polarização.
2. Utilização de polarização como *features* em áreas relacionadas, como hate-speech, toxicidade, fake-news.
3. Análises do impacto nas redes sociais de fatos que acontecem na vida real, por exemplo a morte de uma celebridade ou a corrida eleitoral em um país.
4. Previsão de resultados de eleições.
5. Análises de comportamento dos usuários em torno da polarização.
6. Impacto (ou influência) dos algoritmos das mídias sociais na sua polarização.
7. Abusos do uso da polarização de forma deliberada para ganho próprio.
8. Ferramentas e soluções para conter a polarização.

3.5.1. *Baselines* das métricas de polarização

Dada a diversidade de técnicas e métricas propostas na literatura para identificar polarização, alguns trabalhos direcionaram seus esforços em validá-las e compará-las. Como vimos neste curso (seção 3.4), alguns trabalhos focaram em métricas de polarização que utilizam o grafo da rede [Guerra et al., 2013; Garimella et al., 2018b]. Outros vão comparar as métricas baseadas em conteúdo [Al-Ayyoub et al., 2018; Morales et al., 2015].

Alguns trabalhos exploraram métricas que são independentementes do conteúdo [Garimella et al., 2018b; Guerra et al., 2013]. Outras métricas ainda demandam de um especialista [Jang et al., 2016; Ribeiro et al., 2019]. De toda forma, estudos nesta área se mostraram importantes pois ainda não existe um consenso sobre como medir polarização.

3.5.2. Polarização como *features*

As métricas de polarização não necessariamente devem ser vistam como um fim em si mesmas. Podemos explorar seu uso como uma ferramenta de apoio em áreas correlatas, como classificação discurso de ódio [Akhtar et al., 2019], toxicidade [Guimaraes and Weikum, 2021] e fake-news [Vicario et al., 2019]. É fácil pensarmos que em tópicos com presença de discurso de ódio, por exemplo, provavelmente teremos uma população dividida (polarizada). Tomemos uma discussão com presença de discurso de ódio ocorrendo nos comentários de uma notícia envolvendo racismo. É esperado que a população em torno dessa discussão se divida entre aqueles que atacam com ódio e aqueles que rechaçam o racismo e a intolerância.

De modo geral, os trabalhos vão percorrer suas bases de dados em busca de tópicos identificados como polarizados. E então se debruçar sobre estes tópicos de várias formas. Por exemplo, o nível de discordância (polarização) entre os especialistas que classificam uma base de dados pode ser utilizada juntamente com o conjunto de treino e gerar melhores classificadores [Akhtar et al., 2019].

Ainda, métricas de polarização podem ser utilizadas para melhorar previsões do aparecimento de mensagens tóxicas em uma conversa [Guimaraes and Weikum, 2021] ou ainda para prever prováveis tópicos alvos de notícias falsas [Vicario et al., 2019]. Todos estes são exemplos de classificadores onde a polarização não é o único aspecto a ser considerado, porém utilizar essas informação trouxe ganho aos classificadores.

3.5.3. Impactos de eventos nas mídias sociais

Outro ponto de interesse é a analise da comunicação nas mídias sociais atreladas a eventos do mundo real. Na literatura, encontramos diversos exemplos com eventos importantes, como a análise da polarização quando da morte do Hugo Cháves [Morales et al., 2015] e, em consequência, a crise política instaurada na Venezuela em 2019 [Horawalavithana et al., 2021]; o processo de Impeachment da presidenta Dilma Roussef no Brasil [Moreira et al., 2020], a corrida eleitoral de Donald Trump em 2016 [Yang et al., 2017], dentre outros.

Cada trabalho abordando como as mídias sociais reagiram aos eventos da vida real levando suas particularidades em consideração. Por exemplo, o trabalho de Moreira et al. [2020] realizou uma análise comparativa entre a polarização de dois segmentos da população: a "elite" (classe política) e a "massa" (população comum). Com essa particularidade, o cálculo do viés da "elite", foi realizado à partir das votações ocorridas na câmara dos deputados. Já o trabalho de Morales et al. [2015] se interessou em analisar a polarização dividindo a população em diferentes regiões geográficas.

3.5.4. Previsão de resultados de eleições

Como vimos, muitas métricas de polarização partem da classificação e quantificação de grupos antagônicos da população. Quando o tópico escolhido é aborto, por exemplo, temos uma ideia do tamanho da população pró-escolha e o tamanho da população pró-vida [Lu et al., 2015]. Quando o tópico é um jogo de futebol, quantificamos o tamanho da população que torce para cada um dos times [Guerra et al., 2013]. E, quando o tópico é uma corrida eleitoral, a análise da polarização em torno de um candidato nos mostrará o tamanho da população que o apoia.

No trabalho de Belcastro et al. [2020], por exemplo, os autores fizeram a classificação dos usuários do Twitter durante as eleições de 2016. Para isso, o viés dos usuários foi calculado como a razão entre o número de tweets compartilhados em apoio a cada candidato. Como resultado, eles obtiveram uma previsão da votação mais acurada do que as pesquisas políticas tradicionais. O mesmo resultado positivo foi observado com os dados da eleição de Donald Trump em 2016.

Poder analisar, e até prever, a preferencia dos usuários tem várias aplicações a depender do tópico escolhido. Uma empresa pode estar interessada em prever a recepção de um novo produto, ou um político, em saber o nível de aprovação da população a um

novo projeto, dentre outras aplicações.

3.5.5. Análise do comportamento dos usuários

Outras análises focam em um melhor entendimento do comportamento dos usuários em uma rede polarizada e quais são suas consequências. Os pontos a serem abordados podem variar como, por exemplo, uma caracterização da rede e dos seus usuários [Bakshy et al., 2015], um melhor entendimento do processo de homofilia e das câmaras de eco [Bright, 2017; Tokita et al., 2021], dos conteúdos que tem maior alcance em compartilhamento [Bakshy et al., 2015; Bright, 2017; Weld et al., 2021] ou tempo de permanência dos leitores em páginas com diferentes vieses [Garimella et al., 2021],

Ainda, alguns autores escolhem trabalhar com simulações das redes e do comportamento dos usuários. Essas simulações possibilitaram avaliar o impacto da polarização na estrutura da rede [Tokita et al., 2021]. E entender o papel das decisões dos usuários ou do algoritmo da rede social [Garimella et al., 2021; Valensise et al., 2022]. As diferentes particularidade entre as redes sociais, como o Twitter ou o Reddit, e seus impactos na polarização, também é um tópico a ser explorado [Weld et al., 2021].

O trabalho de Waller and Anderson [2021], por exemplo, fez uma análise de todas as comunidades do Reddit de forma a analisar a polarização política total da plataforma desde a sua criação. Dentre outras coisas, eles observaram que a polarização começou a aumentar durante o ano de 2016 com as eleições do Trump e não voltaram a diminuir. Ainda, eles observaram que a polarização se dá pelos novos usuários que entram na plataforma primariamente.

3.5.6. Influência dos algoritmos na polarização

Vimos na última seção que indícios apontam um aumento da polarização ao longo dos anos. Uma pergunta importante a ser respondida é se o advento da Internet juntamente com as redes sociais possuem sua parcela de responsabilidade.

Alguns autores exploraram essa questão, por exemplo, o trabalho de Kulshrestha et al. [2017] focou em calcular o viés do algoritmo de busca do Twitter. Para isso, ele calcula os vieses dos tweets que retornam com respostas a determinadas consultas. Ainda, o trabalho de Valensise et al. [2022], criou um modelo de simulação para entender o impacto da escolha dos usuários e do algoritmo das mídias sociais. Como resultado o trabalho concluiu que o algoritmo tem um grande papel na polarização dos usuários.

3.5.7. Abusos do uso da polarização

Outra investigação a ser realizada é a existência de elementos que utilizam a polarização de forma deliberada. O trabalho de Ribeiro et al. [2019] investigou uma série de anúncios políticos disparados por uma agência russa ao povo americano durante as eleições de 2016. O trabalho apontou que esses anúncios foram disparados no Facebook com o intuito de explorar a polarização da rede. Tais anúncios eram direcionados para perfis específicos de usuários, possuinudo um alcance 10 vezes maior do que a média de anúncios na plataforma. Neste caso em particular, os atacantes procuraram perfis de usuários pertencentes a populações específicas (principalmente liberais e negros) com o objetivo de criar discórdia.

3.5.8. Contenção da polarização

Como vimos, existem alguns indícios de que a polarização vem aumentando. O leitor pode então estar se perguntando quais seriam as soluções existentes para mudar este cenário. Na literatura, alguns trabalhos apresentaram ferramentas e soluções para diminuir a polarização presente nas redes sociais. Como exemplos desses trabalhos, temos a recomendação de usuários e conteúdos com viés diferente [Gillani et al., 2018], a elaboração de *feeds* de notícias que não sejam enviesados [Babaei et al., 2018; Jang and Allan, 2018]. Ou ainda, alguns trabalhos propõe mudanças na rede de comunicação ao adicionar conexões entre antagônicos [Garimella et al., 2017a] ou adicionar novos nós na rede [Garimella et al., 2017b]. De modo geral, o objetivo dessas soluções é diminuir as câmaras de eco, promovendo o contato com visões opostas.

3.6. Prática: Covid-19 e Hidroxicloroquina (HCQ)

Entendemos que uma abordagem prática é de suma importância para a fixação dos conteúdos aprendidos. Nesta seção, apresentamos um exemplo prático completo desde a coleta até a medição e análise de polarização em rede social. O código fonte utilizado nesta seção podem ser acessados em: `https://github.com/brhott/webmedia2022-polarization`. Para este estudo prático utilizaremos ferramentas e APIs desenvolvidas na linguagem Python[13]. As ferramentas serão apresentadas no desenvolver da seção.

3.6.1. Base de dados

A primeira fase é a escolha e obtenção da base de dados a ser estudada. Neste exemplo, escolhemos analisar o tópico do uso do medicamento Hidroxicloroquina (HCQ) para o tratamento da Covid-19. A base de dados será composta por um conjunto de tweets acerca do tema. Neste caso não nos preocuparemos com os usuários, realizaremos uma análise de polarização baseada no conteúdo postado. Para coleta, utilizamos a ferramenta *Tweepy* [14] para acesso à API do Twitter [15].

O código de coleta de dados pode ser visualizado no Programa 3.1. Nas duas primeiras linhas importamos as bibliotecas Tweepy e *Pandas* [16] – Pandas é uma biblioteca para manipulação e análise de dados. Na linha 4, inicializamos a biblioteca *Tweepy* com uma chave de acesso denominada `Bearer_Token` – mais informações sobre instalação e configuração da biblioteca *Tweepy* podem ser encontradas em `tweepy.org`. Vamos buscar por 100 *tweets* que contenham as palavras `hydroxychloroquine`, `chloroquine` e `HCQ`; eliminando os *retweets* (linhas 5 e 6). E, com a linha 7, armazenamos o resultado da nossa consulta em um *Dataframe* Pandas.

Aprendemos como realizar uma consulta utilizando a biblioteca *Tweepy*. Porém, para o restante desta seção, vamos utilizar a base de dados disponibilizada no trabalho de Mutlu et al. [2020]. Esta base contém um total de 14.374 *tweets* sobre o uso da Hidroxicloroquina como medicamento para a covid-19. Os *tweets* foram coletados durante todo

[13] `https://python.org`
[14] `https://tweepy.org`
[15] `https://developer.twitter.com`
[16] `https://pandas.pydata.org`

```
import tweepy
import pandas as pd

client = tweepy.Client(Bearer_Token)
query = "hydroxychloroquine chloroquine HCQ -is:retweet"
tweets = client.search_recent_tweets(query=query, max_results=100)
df = pd.DataFrame(tweets.data).set_index('id')
```

Programa 3.1: Download de uma base de dados utilizando a API do Twitter

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
sid = SentimentIntensityAnalyzer() #inicializacao

scores = df['text'].apply(lambda text : sid.polarity_scores(text))
df['compound'] = scores.apply(lambda score : score['compound'])
```

Programa 3.2: Análise de sentimentos da base de dados utilizando a ferramenta *Vader*

o mês de Abril de 2020 e seus vieses foram classificados manualmente. Essa classificação foi feita com relação ao seguinte questionamento: "coroquina/hidroxicoroquina é a cura para o novo coronavírus?". Desses, utilizamos apenas 9.117 tweets que continuavam online, sendo 3.732 tweets classificados como negativos, 3.385, positivos, e, 2.000, neutros. Mais detalhes sobre a base de dados e sua coleta podem ser encontrados no trabalho original.

3.6.2. Viés dos tweets

A segunda fase da metodologia é a de cálculo dos vieses dos dados. Como foi dito, optamos por utilizar uma base de dados onde o viés de cada tweet fora anotado manualmente. Porém, vamos abordar também a análise de sentimentos destes tweets, da mesma forma que alguns trabalhos da literatura [Al-Ayyoub et al., 2018; Vicario et al., 2019]. Nesse sentido, vamos abordar a utilização da ferramenta *Vader* [17] para extrair o sentimento de cada tweet da base de dados.

O código no Programa 3.2 realiza a inicialização e aplicação do *Vader* no texto dos tweets coletados (`df['text']`). Os resultados calculados são armazenados em `df['compound']`. Segundo a documentação da ferramenta, os valores são interpretados da seguinte maneira: valores no intervalo $[-1.0, -0.05)$ denominam um tweet com sentimento negativo; valores no intervalo $[-0.05, 0.05]$, sentimento neutro; e valores no intervalo $(0.05, 1.0]$, sentimento positivo.

3.6.3. Polarização do grupo

A terceira faze compreende na escolha e aplicação das métricas de polarização. Como estamos trabalhando com o conteúdo dos tweets e não com a estrutura da rede, vamos implementar as métricas das seções 3.4.1 e 3.4.2 (Programa 3.3). As linhas 2-4 separam os tweets negativos dos positivos. As linhas 6-13 realizam alguns cálculos intermediários para as métricas, como, por exemplo, a contagem de tweets negativos ou o cálculo do

[17] `https://nltk.org`

```
1  # Separacao entre tweets positivos e negativos.
2  g  = df['compound']
3  gn = df[df['compound'] <= -0.05]['compound']
4  gp = df[df['compound'] >=  0.05]['compound']
5
6  A   = g.count()      # num de tweets totais
7  An  = gn.count()     # populacao de tweets negativos
8  Ap  = gp.count()     # populacao de tweets positivos
9  A0  = A - An - Ap    # populacao de tweets neutros
10 Sn  = abs(gn.sum()) # soma dos valores de sentimento positivos
11 Sp  = gp.sum()       # soma dos valores de sentimento negativos
12 gcp = gp.mean()      # centroide dos tweets positivos
13 gcn = gn.mean()      # centroide dos tweets negativos
14
15 PN    = Ap / An                    # metrica PN
16 RPN   = min(An, Ap) / max(An, Ap) # metrica RPN
17 NPN   = A0 / (An + Ap)             # metrica NPN
18 PNT   = (Ap + An) / A              # metrica PNT
19 PNPNT = PN * PNT                   # metrica PNPNT
20 RPNV  = min(Sn, Sp) / max(Sn, Sp) # metrica RPNV
21
22 dA  = abs(Ap - An) / A # diferenca do tamanho das populacoes
23 d   = (gcp - gcn) / 2  # distancia entre centroides
24 m   = (1 - dA) * d     # metrica do dipolo eletrico
```

Programa 3.3: Cálculo da polarização do grupo com implementações dos trabalhos de Al-Ayyoub et al. [2018] e Morales et al. [2015].

centroide dos tweets positivos. Por sua vez, as linhas 15-20 trazem as diversas métricas que abordamos na seção 3.4.1. Por fim, as linhas 22-24 são responsáveis pela métrica de polarização do dipolo elétrico abordada na seção 3.4.2.

O código apresentado nesta seção utilizou os dados de análise de sentimento calculados pelo *Vader*. Porém, é fácil observar que as linhas 2-4 podem ser facilmente adaptadas para carregar os valores de viés obtidos por outros meios. Na próxima seção também iremos apresentar os resultados de cálculo das métricas utilizando os vieses classificados manualmente (como visto na seção 3.6.1).

3.6.4. Análise da polarização

A última etapa consiste na análise e discussão dos resultados. A Figura 3.10 apresenta a distribuição tweets em torno de seus vieses e nos dá uma ideia do nível de polarização em torno do tópico em questão. A figura da esquerda nos mostra a distribuição dos sentimentos dos tweets, onde podemos observar que os conjuntos de tweets negativos e positivos possuem distribuições parecidas. Quando analisamos os vieses anotados manualmente (figura à direita), observamos que os conjuntos de tweets negativos e positivos possuem tamanhos similares. Pela visualização de ambos os histogramas, podemos esperar uma alta polarização dessa população (população dividida entre dois grupos de posições opostas).

Os resultados das métricas de polarização calculadas podem ser visualizadas na

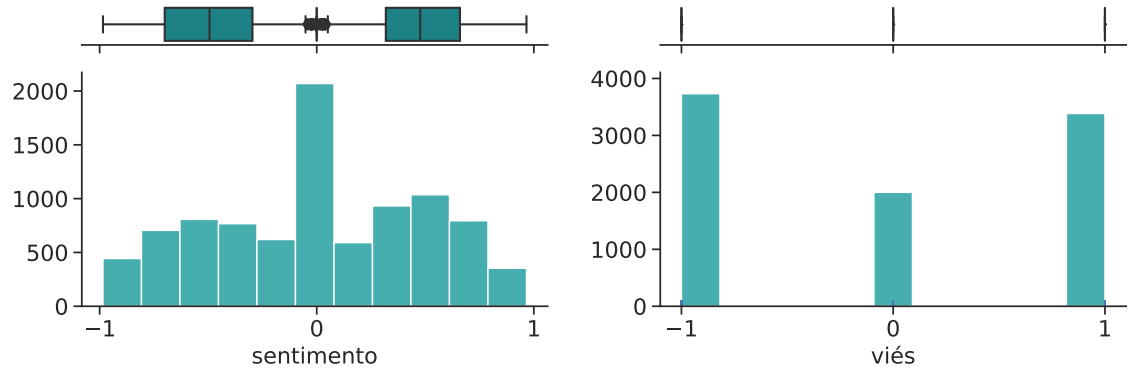

Figura 3.10: Histograma dos vieses dos tweets calculados com análise de sentimentos (esquerda) e manualmente (direita). Acima temos os boxplots de cada um dos grupos: negativo, neutro e positivo.

Tabela 3.2: Resultado das métricas de polarização utilizando vieses calculados por análise de sentimentos (S) e de maneira manual (M)

	PN	RPN	NPN	PNT	PNPNT	RPNV	DM
S	1.11	0.90	0.25	0.80	0.89	0.94	0.47
M	0.91	0.91	0.28	0.78	0.70	0.91	0.96

Tabela 3.2. Como previámos, a polarização encontrada foi alta na grande maioria das métricas. Uma adendo quanto a métrica NPN que é a única das métricas que inversamente proporcional à polarização. Uma segunda observação agora com relação à métrica do dipolo elétrico (DM) onde temos resultados diferentes entre as bases de dados (0.47 e 0.96). Essa métrica leva em consideração o grau de distanciamento das opiniões de cada um dos grupos. Nesse caso, o viés anotado manualmente com somente 3 valores – negativo (-1), neutro (0) e positivo (+1) – colocará a distância entre os grupos negativo e positivo como a maior possível. É importante ressaltar que não é prudente comparar os resultados com bases de dados processadas de maneira diferentes como a análise de sentimentos e a anotação manual.

De toda forma, o conjunto de métricas apresentadas na Tabela 3.2, juntamente com a visualização dos histogramas dos *tweets* apresentados na Figura 3.10 nos trazem valiosas informações acerca da polarização do tópico em questão. Vimos que os conjuntos antagônicos tem tamanhos próximos (PN, RPN), que o número de tweets neutros é baixo com relação aos tweets com vieses claros (NPN, PNT) e que a razão entre o somatório dos vieses de cada grupo fica próxima de 1 (RPNV), ou seja, que os grupos estão enviesados (distância até o centro) de maneira similar entre eles. Todas estas são características de uma população polarizada.

3.7. Considerações Finais

Identificar polarização nas redes sociais ainda é uma tarefa dependente do contexto. Neste capítulo, apresentamos uma revisão bibliográfica com o objetivo de contextualizar e apre-

sentar ao leitor o atual cenário da pesquisa em polarização. Para tanto, apresentamos os principais conceitos e definições da área além de prover ao leitor o ferramental necessário para elaborar suas próprias análises na área de polarização. Para isso, propusemos e apresentamos uma metodologia que passa pela coleta e processamento de dados no contexto da polarização; a classificação do viés das entidades presentes nos dados, com enfoque em texto; a escolha e aplicação de métricas de identificação e, a depender da técnica utilizada quantificação da polarização do tópico em questão; por fim, a análise e interpretação dos resultados da polarização.

Também foi proposta uma taxonomia das métricas e técnicas de polarização em redes sociais. Abordamos métricas elaboradas de diferentes meios, como métricas estatísticas e que utilizam técnicas de teoria dos grafos, o que evidencia a falta um consenso na literatura sobre como operacionalizar a identificação e a polarização de uma população. Nessa mesma linha, este capítulo apresentou diversos exemplos de trabalhos correlatos, seus resultados e suas implicações. Por fim, foi apresentado um exemplo prático de análise de polarização no contexto da pandemia Covid-19 e da discussão em torno do medicamento Hidroxicloroquina.

Agradecimentos. Este trabalho foi parcialmente financiado pelo CNPQ, FAPEMIG e FAPESP.

Referências

Akhtar, S., Basile, V., and Patti, V. (2019). A new measure of polarization in the annotation of hate speech. In *International Conference of the Italian Association for Artificial Intelligence*, pages 588–603. Springer.

Akoglu, L. (2014). Quantifying political polarity based on bipartite opinion networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 2–11.

Al Amin, M. T., Aggarwal, C., Yao, S., Abdelzaher, T., and Kaplan, L. (2017). Unveiling polarization in social networks: A matrix factorization approach. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pages 1–9. IEEE.

Al-Ayyoub, M., Rabab'ah, A., Jararweh, Y., Al-Kabi, M. N., and Gupta, B. B. (2018). Studying the controversy in online crowds' interactions. *Applied Soft Computing*, 66:557–563.

Aldayel, A. and Magdy, W. (2019). Assessing sentiment of the expressed stance on social media. In Weber, I., Darwish, K. M., Wagner, C., Zagheni, E., Nelson, L., Aref, S., and Flöck, F., editors, *Social Informatics*, pages 277–286, Cham. Springer International Publishing.

ALDayel, A. and Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing and Management*, 58(4).

Almerekhi, H., Kwak, H., Salminen, J., and Jansen, B. J. (2020). Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of The Web Conference 2020*, pages 3033–3040.

Arora, S. D., Singh, G. P., Chakraborty, A., and Maity, M. (2022). Polarization and social media: A systematic review and research agenda. *Technological Forecasting and Social Change*, 183:121942.

Ayed, S. B., Trichili, H., and Alimi, A. M. (2015). Data fusion architectures: A survey and comparison. In *2015 15th International Conference on Intelligent Systems Design and Applications (ISDA)*, pages 277–282. IEEE.

Babaei, M., Kulshrestha, J., Chakraborty, A., Benevenuto, F., Gummadi, K. P., and Weller, A. (2018). Purple feed: Identifying high consensus news posts on social media. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 10–16.

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Barros, M. F., Ferreira, C. H., Santos, B. P. d., Júnior, L. A., Mellia, M., and Almeida, J. M. (2021). Understanding mobility in networks: A node embedding approach. *arXiv preprint arXiv:2111.06161*.

Batrinca, B. and Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30(1):89–116.

Beigman Klebanov, B., Beigman, E., and Diermeier, D. (2010). Vocabulary choice as an indicator of perspective. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 253–257, Uppsala, Sweden. Association for Computational Linguistics.

Belcastro, L., Cantini, R., Marozzo, F., Talia, D., and Trunfio, P. (2020). Learning political polarization on social media using neural networks. *IEEE Access*, 8:47177–47187.

Bettini, C., Brdiczka, O., Henricksen, K., Indulska, J., Nicklas, D., Ranganathan, A., and Riboni, D. (2010). A survey of context modelling and reasoning techniques. *Pervasive and mobile computing*, 6(2):161–180.

Borah, A. and Singh, S. R. (2022). Investigating political polarization in India through the lens of Twitter. *Social Network Analysis and Mining*, 12(1):1–26.

Boxell, L., Gentzkow, M., and Shapiro, J. M. (2017). Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences*, 114(40):10612–10617.

Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social networks*, 30(2):136–145.

Bright, J. (2017). Explaining the emergence of echo chambers on social media: the role of ideology and extremism. *Available at SSRN 2839728*.

Canales, L. and Martínez-Barco, P. (2014). Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, pages 37–43, Quito, Ecuador. Association for Computational Linguistics.

Choi, Y., Jung, Y., and Myaeng, S.-H. (2010). Identifying controversial issues and their sub-topics in news articles. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pages 140–153. Springer.

Coletto, M., Garimella, K., Gionis, A., and Lucchese, C. (2017). Automatic controversy detection in social media: a content-independent motif-based approach. *Online Social Networks and Media*, 3:22–31.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Darwish, K., Magdy, W., Rahimi, A., Baldwin, T., and Abokhodair, N. (2018). Predicting online islamophobic behavior after #parisattacks. *The Journal of Web Science*, 4(3):34–52.

Darwish, K., Magdy, W., and Zanouda, T. (2017). Improved stance prediction in a user similarity feature space. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 145–148, New York, NY, USA. Association for Computing Machinery.

Darwish, K., Stefanov, P., Aupetit, M., and Nakov, P. (2020). Unsupervised user stance detection on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):141–152.

Dong, R., Sun, Y., Wang, L., Gu, Y., and Zhong, Y. (2017). Weakly-guided user stance prediction via joint modeling of content and social interaction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1249–1258, New York, NY, USA. Association for Computing Machinery.

Dori-Hacohen, S. and Allan, J. (2015). Automated controversy detection on the web. In *European Conference on Information Retrieval*, pages 423–434. Springer.

Easley, D. and Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press.

Ebrahimi, J., Dou, D., and Lowd, D. (2016). A joint sentiment-target-stance model for stance classification in tweets. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, pages 2656–2665.

Elfardy, H. and Diab, M. (2016). CU-GWU perspective at SemEval-2016 task 6: Ideological stance detection in informal text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 434–439, San Diego, California. Association for Computational Linguistics.

Ferreira, C. H., Murai, F., Silva, A. P., Almeida, J. M., Trevisan, M., Vassio, L., Mellia, M., and Drago, I. (2021). On the dynamics of political discussions on instagram: A network perspective. *Online Social Networks and Media*, 25:100155.

Ferreira, W. and Vlachos, A. (2016). Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Ferreira, W. and Vlachos, A. (2019). Incorporating label dependencies in multilabel stance detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6350–6354, Hong Kong, China. Association for Computational Linguistics.

Fiorina, M. P., Abrams, S. A., and Pope, J. C. (2008). Polarization in the american public: Misconceptions and misreadings. *The Journal of Politics*.

Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2017a). Reducing controversy by connecting opposing views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 81–90.

Garimella, K. et al. (2018a). Polarization on social media.

Garimella, K., Gionis, A., Parotsidis, N., and Tatti, N. (2017b). Balancing information exposure in social networks. *Advances in neural information processing systems*, 30.

Garimella, K., Morales, G. D. F., Gionis, A., and Mathioudakis, M. (2018b). Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27.

Garimella, K., Smith, T., Weiss, R., and West, R. (2021). Political polarization in online news consumption. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 152–162.

Ghosh, S., Singhania, P., Singh, S., Rudra, K., and Ghosh, S. (2019). Stance detection in web and social media: A comparative study. In Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D. E., Heinatz Bürki, G., Cappellato, L., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 75–87, Cham. Springer International Publishing.

Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., and Roy, D. (2018). Me, my echo chamber, and i: Introspection on social media polarization. *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, pages 823–831.

Gokcekus, S., Firth, J. A., Regan, C., and Sheldon, B. C. (2021). Recognising the key role of individual recognition in social networks. *Trends in Ecology & Evolution*, 36(11):1024–1035.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21.

Guerra, P., Meira Jr, W., Cardie, C., and Kleinberg, R. (2013). A measure of polarization on social media networks based on community boundaries. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Guimaraes, A. and Weikum, G. (2021). X-posts explained: Analyzing and predicting controversial contributions in thematically diverse reddit forums. In *ICWSM*, pages 163–172.

Hamidian, S. and Diab, M. T. (2019). Rumor detection and classification for twitter data. *CoRR*, abs/1912.08926.

Hercig, T., Krejzl, P., Hourová, B., Steinberger, J., and Lenc, L. (2017). Detecting stance in czech news commentaries. *ITAT*, 176:180.

Horawalavithana, S., Ng, K. W., and Iamnitchi, A. (2021). Drivers of polarized discussions on twitter during venezuela political crisis. In *13th ACM Web Science Conference 2021*, pages 205–214.

Iyengar, S., Sood, G., and Lelkes, Y. (2012). Affect, not ideologya social identity perspective on polarization. *Public opinion quarterly*, 76(3):405–431.

Jang, M. and Allan, J. (2018). Explaining controversy on social media via stance summarization. *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, pages 1221–1224.

Jang, M., Foley, J., Dori-Hacohen, S., and Allan, J. (2016). Probabilistic approaches to controversy detection. In *Proceedings of the 25th ACM international on conference on information and knowledge management*, pages 2069–2072.

Kawintiranon, K. and Singh, L. (2021). Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online. Association for Computational Linguistics.

Khaleghi, B., Khamis, A., Karray, F. O., and Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44.

Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, 13(3).

Klayman, J. (1995). Varieties of confirmation bias. *Psychology of learning and motivation*, 32:385–418.

Klenner, M., Amsler, M., and Hollenstein, N. (2014). Verb polarity frames: a new resource and its application in target-specific polarity classification. In *KONVENS*, pages 106–115.

Kochkina, E., Liakata, M., and Augenstein, I. (2017). Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.

Kubin, E. and von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, 45(3):188–206.

Küçük, D. and Can, F. (2020). Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Kucher, K., Paradis, C., and Kerren, A. (2018). Visual analysis of sentiment and stance in social media texts. In *Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Posters*, EuroVis '18, page 49–51, Goslar, DEU. Eurographics Association.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., and Karahalios, K. (2017). Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 417–432.

Lai, M., Farías, D. I. H., Patti, V., and Rosso, P. (2017). Friends and enemies of clinton and trump: Using context for detecting stance in political tweets. *CoRR*, abs/1702.08021.

Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, 80(S1):392–410.

Li, C., Porco, A., and Goldwasser, D. (2018). Structured representation learning for online debate stance prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3728–3739, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Li, Y. and Caragea, C. (2019). Multi-task stance detection with sentiment and stance lexicons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6299–6305, Hong Kong, China. Association for Computational Linguistics.

Lima, L., Reis, J. C., Melo, P., Murai, F., Araujo, L., Vikatos, P., and Benevenuto, F. (2018). Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 515–522. IEEE.

Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Liu, C., Li, W., Demarest, B., Chen, Y., Couture, S., Dakota, D., Haduong, N., Kaufman, N., Lamont, A., Pancholi, M., Steimel, K., and Kübler, S. (2016). IUCL at SemEval-2016 task 6: An ensemble model for stance detection in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 394–400, San Diego, California. Association for Computational Linguistics.

Lu, H., Caverlee, J., and Niu, W. (2015). Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 213–222.

Matakos, A., Terzi, E., and Tsaparas, P. (2017). Measuring and moderating opinion polarization in social networks. *Data Mining and Knowledge Discovery*, 31(5):1480–1505.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444.

Mejova, Y., Zhang, A. X., Diakopoulos, N., and Castillo, C. (2014). Controversy and sentiment in online news. *arXiv preprint arXiv:1409.8152*.

Milroy, L. and Llamas, C. (2013). Social networks. *The handbook of language variation and change*, pages 407–427.

Mitchell, J. C. (1974). Social networks. *Annual review of anthropology*, 3:279–299.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3).

Morales, A. J., Borondo, J., Losada, J. C., and Benito, R. M. (2015). Measuring political polarization: Twitter shows the two sides of venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3):033114.

Moreira, R. C., Vaz-de Melo, P. O., and Pappa, G. L. (2020). Elite versus mass polarization on the brazilian impeachment proceedings of 2016. *Social Network Analysis and Mining*.

Mutlu, E. C., Oghaz, T., Jasser, J., Tutunculer, E., Rajabi, A., Tayebi, A., Ozmen, O., and Garibay, I. (2020). A stance data set on polarized conversations on twitter about the efficacy of hydroxychloroquine as a treatment for covid-19. *Data in brief*, 33:106401.

Newman, M. E. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104.

Pannucci, C. J. and Wilkins, E. G. (2010). Identifying and avoiding bias in research. *Plastic and reconstructive surgery*, 126(2):619.

Pergola, G., Gui, L., and He, Y. (2020). A disentangled adversarial neural topic model for separating opinions from plots in user reviews. *arXiv preprint arXiv:2010.11384*.

Popescu, A.-M. and Pennacchiotti, M. (2010). Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1873–1876.

Rajendran, G., Chitturi, B., and Poornachandran, P. (2018). Stance-in-depth deep neural approach to stance classification. *Procedia Computer Science*, 132:1646–1653. International Conference on Computational Intelligence and Data Science.

Rashed, A., Kutlu, M., Darwish, K., Elsayed, T., and Bayrak, C. (2020). Embeddings-based clustering for target specific stances: The case of a polarized turkey. *CoRR*, abs/2005.09649.

Rathje, S., Van Bavel, J. J., and Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26):e2024292118.

Ravi, K. and Ravi, V. (2015). A survey on opinion mining and sentiment analysis. *Know.-Based Syst.*, 89(C):14–46.

Rettore, P. H., Santos, B. P., Campolina, A. B., Villas, L. A., and Loureiro, A. A. (2016). Towards intra-vehicular sensor data fusion. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 126–131. IEEE.

Ribeiro, F. N., Saha, K., Babaei, M., Henrique, L., Messias, J., Benevenuto, F., Goga, O., Gummadi, K. P., and Redmiles, E. M. (2019). On microtargeting socially divisive ads: A case study of russia-linked ad campaigns on facebook. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 140–149.

Roy, S. and Goldwasser, D. (2020). Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716, Online. Association for Computational Linguistics.

Sailunaz, K., Dhaliwal, M., Rokne, J., and Alhajj, R. (2018). Emotion detection from text and speech - a survey. *Social Network Analysis and Mining (SNAM), Springer*, 8.

Santos, B. P., Silva, L. A., Celes, C. S., Borges Neto, J. B., Peres, B. S., Vieira, M. A. M., Vieira, L. F. M., Goussevskaia, O. N., and Loureiro, A. A. (2016). Internet das coisas: da teoria à prática. *Minicursos SBRC-Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuıdos*.

Sen, I., Flöck, F., and Wagner, C. (2020). On the reliability and validity of detecting approval of political actors in tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1413–1426, Online. Association for Computational Linguistics.

Shahrezaye, M., Papakyriakopoulos, O., Serrano, J. C. M., and Hegelich, S. (2019). Measuring the ease of communication in bipartite social endorsement networks: A proxy to study the dynamics of political polarization. *ACM International Conference Proceeding Series*, pages 158–165.

Siddiqua, U. A., Chy, A. N., and Aono, M. (2019). Tweet stance detection using an attention based neural ensemble model. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long and Short Papers), pages 1868–1873, Minneapolis, Minnesota. Association for Computational Linguistics.

Sunstein, C. R. (1999). The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper*.

Sunstein, C. R. (2018). *# Republic: Divided democracy in the age of social media*. Princeton University Press.

Swami, S., Khandelwal, A., Singh, V., Akhtar, S. S., and Shrivastava, M. (2018). An english-hindi code-mixed corpus: Stance annotation and baseline system. *CoRR*, abs/1805.11868.

Tachaiya, J., Irani, A., Esterling, K. M., and Faloutsos, M. (2021). Sentistance: Quantifying the intertwined changes of sentiment and stance in response to an event in online forums. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, page 361–368, New York, NY, USA. Association for Computing Machinery.

Thomas, M. and Joy, A. T. (2006). *Elements of information theory*. Wiley-Interscience.

Tokita, C. K., Guess, A. M., and Tarnita, C. E. (2021). Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences of the United States of America*, 118(50).

Tsakalidis, A., Aletras, N., Cristea, A. I., and Liakata, M. (2018). Nowcasting the stance of social media users in a sudden vote: The case of the greek referendum. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 367–376, New York, NY, USA. Association for Computing Machinery.

Tsytsarau, M., Palpanas, T., and Denecke, K. (2011). Scalable detection of sentiment-based contradictions. *DiversiWeb, WWW*, 1:9–16.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., and Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.

Valensise, C. M., Cinelli, M., and Quattrociocchi, W. (2022). The dynamics of online polarization. *arXiv preprint arXiv:2205.15958*.

Vicario, M. D., Quattrociocchi, W., Scala, A., and Zollo, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2):1–22.

Waller, I. and Anderson, A. (2021). Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888):264–268.

Watts, D. J., Rothschild, D. M., and Mobius, M. (2021). Measuring the news and its impact on democracy. *Proceedings of the National Academy of Sciences*, 118(15).

Weld, G., Glenski, M., and Althoff, T. (2021). Political bias and factualness in news sharing across more than 100,000 online communities. *arXiv preprint arXiv:2102.08537*.

Woodrum, E. and Davison, B. L. (1992). Reexamination of religious influences on abortion attitudes. *Review of religious research*, pages 229–243.

Yang, M., Wen, X., Lin, Y.-R., and Deng, L. (2017). Quantifying content polarization on twitter. In *2017 IEEE 3rd international conference on collaboration and internet computing (CIC)*, pages 299–308. IEEE.

Zhang, B., Yang, M., Li, X., Ye, Y., Xu, X., and Dai, K. (2020). Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197, Online. Association for Computational Linguistics.