Uso de URLs para caracterização de comunidades em Redes Sociais *Online*

Carlos M. G. Barbosa[1], **Lucas G. da S. Félix**[1], **Antônio Pedro S. Alves**[1], **Carolina Ribeiro Xavier**[1], **Vinícius da Fonseca Vieira**[1]

[1]Departamento de Ciência da Computação
Universidade Federal de São João del Rei (UFSJ)
São João del-Rei - MG- Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte- MG - Brasil

vinicius@ufsj.edu.br

Resumo. *Para uma melhor compreensão da organização complexa dos indivíduos em redes sociais online, é essencial a investigação dos grupos de usuários e das discussões em torno de determinados assuntos. Este trabalho apresenta uma metodologia para a caracterização de grupos em redes sociais com base nas fontes externas utilizadas pelos seus usuários para a construção da argumentação conduzida sobre determinados assuntos. Uma análise do Twitter em torno da discussão sobre a vacinação da COVID-19 no Brasil mostra que a metodologia é capaz de avançar no entendimento da forma como as informações são produzidas e propagadas, e nas diferenças entre a forma como diversos grupos de usuários utilizam fontes externas de informação.*

Abstract. *For a better understanding of the complex organization of individuals in online social networks, it is essential to investigate user groups and discussions around certain subjects. This work presents a methodology for the characterization of groups in social networks based on external sources used by their users to construct the arguments conducted on discussion. An analysis around the discussion in Twitter about COVID-19 vaccination in Brazil shows that the methodology is able to advance the understanding of the way information is produced and propagated, and the differences that the distinct groups of users use external sources of information.*

1. Introdução

Redes Sociais *Online* (RSO), como Twitter, WhatsApp, Instagram e Facebook, se tornaram eficientes canais para comunicação, acesso à informação, entretenimento e relacionamentos de diferentes formas, resultando em um organismo complexo, formado por milhões de usuários e interações, dispersos ao redor do mundo, cuja capacidade e influência pode estar além do ambiente da própria RSO. Abordagens que possibilitam a sua caracterização se tornam indispensáveis para entender a dinamicidade deste ambiente e nortear a tomada de decisão de governos e outras instituições.

Muitos trabalhos podem ser encontrados na literatura com o objetivo de caracterizar as redes sociais *online* e o comportamento de seus usuários em diferentes ambientes e sob diferentes perspectivas ([Christhie et al. 2018, Resende et al. 2018,

Martins et al. 2019, Cossard et al. 2020, Garimella et al. 2018]). Além do próprio conteúdo produzido e compartilhado por indivíduos dentro das redes sociais, é também muito importante estudar as fontes utilizadas pelas pessoas para sustentar a argumentação veiculada. Assim, é possível enriquecer substancialmente o entendimento da forma como os indivíduos se organizam e se posicionam dentro de seus grupos. Tentando aprofundar a compreensão de debates ocorridos nas redes sociais sob esse ponto de vista, esse trabalho propõe uma metodologia para a caracterização da discussão de usuários no *Twitter* a partir da análise das *Uniform Resource Locators* (URLs), por eles compartilhadas, endereços que apontam para *websites* externos à plataforma *Twitter*.

O estudo conduzido neste trabalho busca caracterizar indivíduos com comportamentos similares de acordo com o tipo de conteúdo que compartilham. Para isso, *tweets* de um determinado assunto são coletados, a partir dos quais é gerada uma rede social, que tem seus indivíduos agrupados topologicamente através de um algoritmo de detecção de comunidades. Em cada uma das comunidades, são identificadas as principais URLs compartilhadas, que são caracterizadas através da classificação apresentada por Guimarães *et al.* [Guimarães et al. 2020], que permite identificar o tipo da mídia externa para o qual apontam (*mainstream media*, mídia alternativa ou plataforma) e seu viés político unidimensional (esquerda, direita ou centro). De forma a melhor contextualizar a análise de URLs proposta neste trabalho, as comunidades são também caracterizadas sob outras perspectivas: dos usuários representantes e dos tópicos discutidos. Os usuários mais centrais, tratados como representantes principais das comunidades, podem indicar a linha de posicionamentos nelas adotadas. Os principais tópicos e termos discutidos em cada comunidade melhoram a compreensão sobre a forma como os assuntos investigados são tratados em cada grupo de usuários.

Este trabalho tem como objetivo investigar o impacto da análise das URLs compartilhadas para as comunidades que se organizam em torno de um tema em uma rede social. Para isso, duas Questões de Pesquisa (QP) são levantadas: QP1) A análise das URLs compartilhadas adiciona complexidade à compreensão das comunidades de usuários no *Twitter* feita através da análise da estrutura topológica e dos conteúdos dos *tweets*? QP2) É possível observar um comportamento distinto entre as comunidades em relação às URLs por elas compartilhadas?

Os resultados obtidos mostram que as comunidades identificadas, além de apresentarem uma organização topológica clara, apresentam também padrões de produção e compartilhamento de conteúdo muito próprios e coerentes com os seus usuários mais centrais. Tomando como base uma discussão sobre a vacinação da COVID-19 no Brasil, observa-se que algumas comunidades, apesar de se assemelharem em alguns aspectos, nitidamente têm visões diferentes – muitas vezes divergentes – sobre o assunto, utilizando o *Twitter* de forma muito própria.

2. Trabalhos relacionados

Muitos trabalhos encontrados na literatura dedicam-se à caracterização de RSO sob diferentes aspectos e, por isso, podem ser relacionados ao presente estudo. Christie *et al.* [Christhie et al. 2018] identificam posicionamentos de usuários do Twitter favoráveis e contrários a candidatos na corrida eleitoral de 2018 no Brasil. Considerando também um cenário de eleição, mas na Alemanha, Morstatter *et al*. [Morstatter et al. 2018] apresen-

tam uma caracterização da forma como se organizam grupos de direita e extrema direita no Twitter. Resende *et al.* [Resende et al. 2018] apresentam um modelo de monitoração e caracterização de dados propagados no WhatsApp em que são monitoradas as opiniões de 127 grupos públicos brasileiros relacionados a discussões políticas e de notícias em geral. Alguns autores investigam como câmaras de eco podem ser formadas em torno de grupos de usuários e qual seu impacto na propagação de informações em redes. Garimella *et al.* [Garimella et al. 2018] definem que este fenômeno pode ser apresentado em dois componentes distintos. Cossard *et al.* [Cossard et al. 2020] apresentam uma avaliação deste fenômeno considerando o debate em torno da vacinação na Itália, encontrando três grupos distintos em torno da discussão.

Neste trabalho, um dos principais aspectos explorados para a análise da discussão em redes sociais é a caracterização das fontes de informação externas à rede. Sabe-se que a descentralização de fontes de informação possibilitada pelas redes sociais *online* permitiu um amplo desenvolvimento de mídias alternativas, em comparação com a mídia conhecida como *mainstream*, formada por jornais, revistas e redes de televisão responsáveis pela formação e divulgação de notícias. Guimarães *et al.* [Guimarães et al. 2020] apresentam um estudo cujo principal objetivo é realizar uma classificação do alinhamento político (*political bias*) de páginas e figuras públicas no Facebook no Brasil e fornecem um *score* que os classifica entre -1 (para um alinhamento editorial dito de esquerda) a +1 (para um alinhamento editorial dito de direita). A base de dados gerada por Guimarães é utilizada aqui para uma classificação das URLs compartilhadas nas comunidades de maior modularidade e usuários mais representativos nessas comunidades.

3. Metodologia

A metodologia apresentada neste trabalho permite análises semiautomáticas em larga escala de discussões no Twitter focada em comunidades e seus usuários, considerando *retweets* ou menções. Primeiramente, é feita a coleta dos *tweets*, que são armazenados em um banco de dados para que possam ser identificados os usuários e os *posts* com menções e *retweets*. A partir disso, a rede é gerada e alguns passos para sua análise topológica são seguidos, que incluem a identificação das comunidades de usuários e dos usuários mais importantes de cada uma delas. Considerando o contexto de cada comunidade, são realizadas as análises do conteúdo dos *tweets*, que incluem o pré-processamento dos textos, a modelagem dos tópicos e a análise das URLs.

3.1. Coleta de dados e pré-processamento

Para a coleta de dados, é utilizada a API do Twitter, considerando *tweets* e *retweets* de acordo com determinadas *hashtags* (ex.: #vacina) ou termos convenientes.A limpeza dos *tweets* coletados para realização de análises de conteúdo é feita com as bibliotecas de processamento de linguagem natural, NLTK[1], Spacy[2]. Alguns passos tradicionalmente adotados em processos de mineração de texto são adotados para o pré-processamento do conteúdo: tokenização, remoção de *stopwords* e geração de bigramas.

3.2. Geração da rede

Para este trabalho, optou-se pelo desenvolvimento de uma rede baseada em *retweets*, embora outros tipos de redes (como redes de menções) pudessem ser facilmente incorpo-

[1] https://www.nltk.org/
[2] https://spacy.io/

rados. Os nós representam os usuários do *Twitter* envolvidos na discussão e as arestas representam uma relação de *retweet*, ou seja, indicam que um usuário compartilhou um *tweet* de outro. São geradas uma versão direcionada e uma não-direcionada da rede. Na versão direcionada da rede, uma aresta (A, B) indica que um usuário A reproduziu um *tweet* de um usuário B. As arestas são ponderadas pela frequência de *retweets* entre os pares de usuários.

3.3. Detecção de comunidades

Uma das hipóteses que sustentam a metodologia apresentada neste trabalho é que a análise das discussões no Twitter pode ser muito mais rica e reveladora quando se considera não apenas o conteúdo dos *tweets* gerados e compartilhados por um indivíduo, mas também o grupo de seus interlocutores, o que pode ajudar a compreender questões relacionadas ao alinhamento ideológico das pessoas que interagem com um determinado tipo de conteúdo e como isso afeta a própria maneira como esse conteúdo é produzido e difundido. Nesse sentido, toma-se como ideia de que a estrutura topológica da rede pode permitir que se identifique os grupos em que os indivíduos se organizam sob a perspectiva de comunidades.Tomando como base a versão não-direcionada da rede, neste trabalho, é utilizado o algoritmo de Louvain *et al.* [Blondel et al. 2008], frequentemente utilizado com sucesso por diversos trabalhos na literatura para o particionamento de grandes redes em comunidades em um intervalo de tempo reduzido, o que é bastante adequado ao presente trabalho.

3.4. Análise de conteúdo

Além da estrutura das redes de comunicação, a metodologia permite que sejam analisados os conteúdos dos *tweets*, sob a perspectiva de análise de tópicos, utilizando abordagem de modelagem de tópicos através do *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003]. Cada tópico é composto por um conjunto de termos, e a probabilidade desses termos aparecerem nesse tópico. Essa análise é realizada isoladamente em cada uma das comunidades encontradas, permitindo uma melhor compreensão das distintas visões ocorridas no debate considerando o assunto analisado.

3.5. Análise e classificação de fontes externas de informação

Este trabalho apresenta um processo de classificação quanto ao tipo de mídia e viés das URLs compartilhadas. Para isso, considerando cada uma das comunidades investigadas, são extraídas as URLs dos *tweets* coletados e armazenados. Expressões regulares são utilizadas para identificar URLs internas da plataforma e as URLs encurtadas (de serviços como `bit.ly` e `buff.ly`) são expandidas, permitindo a obtenção dos endereços originais. Após, as URLs são classificadas quanto ao tipo (*mainstream media*, mídia alternativa, ou plataforma) e viés político (esquerda, centro ou direita).

A classificação da URL quanto ao seu tipo é realizada através da adaptação da abordagem apresentada por Guimarães *et al.* [Guimarães et al. 2020]. Nessa abordagem, URLs de *websites* que não estão registrados em uma organização oficial de imprensa como Associação Nacional de Jornais (ANJ), ANER Associação Nacional de Editores de Revista (ANER) ou Agência Nacional de Telecomunicações (ANATEL) são classificados como mídia alternativa, sendo que mídias registradas em alguns destes órgãos são classificadas como *mainstream*. Essa abordagem foi ajustada para incluir duas novas classificações, uma referente a plataformas de distribuição de conteúdo como Spotify,

YouTube e redes sociais online que recebeu a nomenclatura de "plataforma". Mídias digitais alternativas, compreendendo *websites* que estavam registrados na Associação de Jornalismo Digital (AJOR), receberam a nomenclatura "mídia alternativa AJOR" – *websites* como nexo jornal estão registrados nesta categoria.

A classificação quanto ao viés (*bias*) é realizada utilizando uma base de dados disponibilizada pelo estudo de Guimarães *et al.* em que foram previamente classificados o alinhamento político de diversas páginas no Facebook, utilizando a API de publicidade do Facebook, combinada com uma estratégia de sobreposição de aprendizagem semi-supervisionada em grafos. Ampliando a classificação realizada por Guimarães *et al.*, foi realizada uma expansão dessa classificação utilizando como base a sobreposição de audiência dos *websites*, disponibilizada pela plataforma Alexa[3].

4. Resultados e discussão

Além de uma breve descrição dos experimentos realizados, esta seção apresenta os resultados encontrados e uma discussão quanto às suas aplicações e limitações considerando a metodologia apresentada na Seção 3, aplicada para análise da discussão em torno da vacinação da COVID-19. Além da clara motivação da escolha deste assunto devido à sua enorme relevância no contexto da pandemia da COVID-19 em que se encontra, esse assunto é também interessante por engajar indivíduos que se agrupam em visões ideológicas nitidamente distintas.

Foram coletados $1.779.024$ *tweets/retweets*, entre o período de janeiro a fevereiro de 2021, utilizando para coleta os termos relacionados à vacinação contra COVID-19[4]. Após a coleta dos *tweets*, foi construída uma rede de *retweets* com 502690 vértices e 1067358 arestas, com grau médio de $4,24$. A divisão da rede social em comunidades é um ponto fundamental da metodologia proposta, já que permite identificar usuários com um mesmo alinhamento de posicionamento/opiniões referentes a um determinado assunto. Tomando como base a rede social construída, é possível identificar a estrutura de comunidades de indivíduos que estabelecem um volume de interações maior do que o que se poderia esperar, indicando a organização de grupos com maior alinhamento de opinião considerando o assunto analisado. Utilizando o método de Louvain [Blondel et al. 2008], foi obtida uma partição com modularidade $Q = 0,6571$, o que pode ser considerada uma modularidade razoavelmente capaz de separar os usuários em grupos bem definidos.

4.1. Contextualização e caracterização das comunidades

Para simplificar a condução da discussão, serão apresentados os resultados das quatro comunidades de maior modularidade, que representam grupos de maior densidade topológica e, supostamente, envolvem indivíduos com maior alinhamento. Em cada uma das comunidades, os vértices foram ordenados de acordo com a centralidade de PageRank, gerando, assim, um *rank* que permite identificar os usuários mais significativos, que podem atuar como representantes de suas respectivas comunidades e auxiliar sua contextualização. A Tabela 1 apresenta algumas características básicas das comunidades analisadas, assim como seus 10 usuários melhor classificados de acordo com o PageRank.

[3] `https://www.alexa.com/`

[4] Os *tweets* coletados estão disponíveis em `https://datastudio.google.com/s/sTc7JJwEhpE`

	Comunidade 1	Comunidade 2	Comunidade 3	Comunidade 4
Modularidade	0.1435	0.0763	0.0527	0.0470
Núm. usuários	43736	43700	32868	41267
Núm. *tweets*	196687	138190	115639	100743
Razão *retweets*	0.7316	0.7101	0.6218	0.6739
Top usuários	CarlaZambelli38	costa_rui	g1	oatila
	lcoutinho	HaddadDebochado	CNNBrasil	lolaferreira
	jairbolsonaro	cartacapital	exilado	RandonadmM
	taoquei1	padilhando	GloboNews	luizacaires3
	leandroruschel	Debora_D_Diniz	RodrigoMaia	elsonh
	BrazilFight	cidadaoprimata	CNNBrBusiness	EthelMaciel
	LaurinhaIronic	LulaOficial	gugachacra	sailorthrash
	conexaopolitica	dadourado	gusthpa	ThomasVConti
	kimpaim	reinaldoazevedo	FMouraBrasil	brgenovez
	carteiroreaca	GuilhermeBoulos	folha	RenanPeixoto_

Tabela 1. Características básicas das comunidades encontradas.

Primeiramente, é preciso observar que nas comunidades investigadas, uma grande razão do conteúdo é propagado através de *retweets*, indicando que o critério definido para a construção da rede pode, de fato, levar a estruturas que condizem com a forma como a informação alcança os leitores. Além disso, uma análise imediata dos usuários que compõem as primeiras posições dos *ranks* de cada comunidade permite observar distinções claras em seus posicionamentos e linhas editoriais. A comunidade 1 é formada por indivíduos em muitas ocasiões alinhados com o Governo Federal e com as políticas por ele conduzidas no enfrentamento da pandemia da COVID-19 (como por exemplo os usuários `CarlaZambelli38`, `lcoutinho` e `jairbolsonaro`). A comunidade 2, por outro lado, tem suas primeiras posições formadas por usuários abertamente contrários às políticas adotadas pelo Governo Federal no enfrentamento da pandemia (como os usuários `costa_rui`, `HaddadDebochado`, `LulaOficial` e `GuilhermeBoulos`). As primeiras posições da comunidade 3 são formadas, majoritariamente, por jornalistas e veículos de mídia (como `g1`, `CNNBrasil`, `GloboNews` e `gugachacra`). Na comunidade 4 destacam-se epidemiologistas e divulgadores científicos (como os usuários `oatila`, `luizacaires3` e `EthelMaciel`).

Com o objetivo de melhorar a caracterização das comunidades na rede de *retweet*, foi realizado um estudo dos seus principais tópicos discutidos, utilizando o método *Latent Dirichlet Allocation* (LDA) para modelagem de tópicos. Para ilustrar a distinção de conteúdos discutidos nas comunidades, a Figura 1 apresenta os 10 principais termos discutidos nos três principais tópicos nas comunidades 1 e 2.

Os tópicos destacados na Figura 1 refletem a polarização observada na rede social gerada para a discussão de usuários sobre "vacinação", que descreve diretamente o contexto histórico da época da coleta dos dados, durante a pandemia do COVID-19 e logo após o anúncio da aprovação das primeiras vacinas e início das campanhas de imunização em diversos países do mundo, incluindo o Brasil. Analisando a comunidade 1 é possível identificar tópicos de claro ataque à vacinação em geral, mas, mais enfaticamente, à vacina `coronaVac`[5], desenvolvida pela farmacêutica chinesa Sinovac em parceria com instituto Butantan. Dentre os termos utilizados para criticar a vacina coronaVac podemos identificar os termos "vacina chinesa". Também é possível identificar

[5]`https://www.bbc.com/portuguese/brasil-54609665`

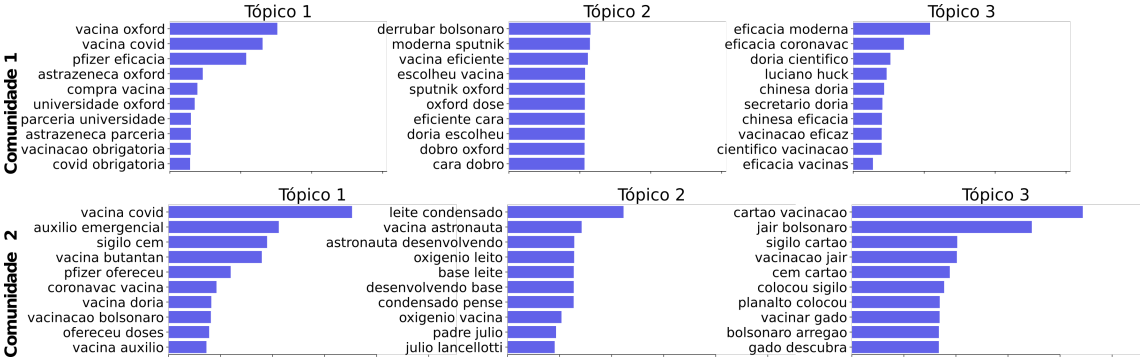

Figura 1. Termos dos três principais tópicos discutidos nas comunidades 1 e 2.

termos xenofóbicos como "virus chines". Outros termos presentes na comunidade 1 envolvem uma discussão quanto à condução de tratamento precoce utilizando cloroquina, comprovadamente ineficaz para o combate ao coronavírus, sendo seu uso não recomendado pela OMS (Organização Mundial da Saúde)[6]. Analisando a comunidade 2, podemos identificar termos relacionados a um pedido de impeachment do Presidente da República ("impeachment bolsonaro"), defesa da ciência, incentivo à vacinação, críticas ao sigilo colocado no cartão de vacinação do Presidente da República, além de uma defesa do SUS (Sistema Único de Saúde). Assim, é possível novamente identificar comunidades com posicionamentos nitidamente antagônicos, reforçando a distância entre esses dois grupos analisados.

Os resultados apresentados nesta seção mostram que há uma distinção clara do comportamento de diferentes comunidades. Entretanto, buscando responder às Questões de Pesquisa colocadas na Seção 1, precisamos investigar se as fontes utilizadas pelos grupos são uma potencial fonte para essas distinções de posicionamento observadas, o que será explorado na Seção 4.2.

4.2. Análise das fontes das informações

A construção da metodologia apresentada neste trabalho parte da hipótese que a análise das discussões ocorridas no Twitter trazem conclusões muito mais ricas e reveladoras quando são analisadas as fontes externas à plataforma com as quais os usuários contam para sustentar as argumentações. Nesse sentido, pode-se afirmar que a linha editorial e as direções ideológicas e políticas assumidas ou não pelas fontes consideradas pelos indivíduos têm uma forte relação com suas próprias condutas e seus próprios alinhamentos ideológicos. Na maior parte das vezes, a forma com a qual os usuários trazem conteúdos de fontes externas ao Twitter é através de URLs e, por isso, temos uma importante justificativa para o uso desse tipo de ligação externa. As URLs contidas nos *tweets* são classificadas quanto ao tipo de mídia para onde apontam (*mainstream*, alternativa e plataforma) e a um viés político unidirecional (esquerda, centro e direita), também chamado pelo termo em inglês *Political Bias*, através da abordagem apresentada na Seção 3.5.

Após uma filtragem para remoção de links internos do Twitter, foram selecionadas e expandidas 150.901 URLs, dos quais foram selecionadas as 80 URLs mais compartilha-

[6] https://www.who.int/news-room/q-a-detail/coronavirus-disease-(covid-19)-hydroxychloroquine

das nos assuntos analisados para serem classificadas seguindo os passos metodológicos apresentados na Seção 3.

A Figura 2 apresenta, para cada uma das comunidades investigadas, a frequência das 15 URLs dos *websites* mais compartilhados considerando os *tweets* estudados, assim como a classificação do tipo de mídia para o qual apontam. Com essa visualização, espera-se que seja possível identificar os padrões adotados por cada um dos grupos para a construção de argumentações que sustentem os discursos adotados nos *posts* compartilhados por seus usuários.

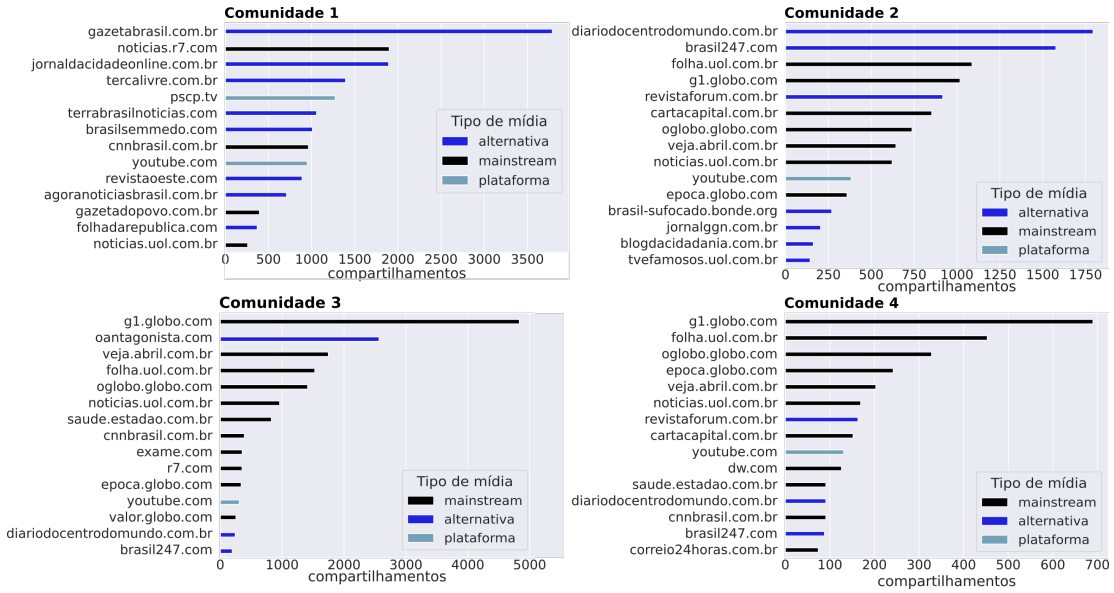

Figura 2. Frequência de ocorrência das URLs mais compartilhadas nas comunidades 1, 2, 3 e 4.

A partir da Figura 2, é possível perceber que mídias consideradas *mainstream* são predominantemente utilizadas nas comunidades 3 e 4, representando aproximadamente 73% das fontes apresentadas. Considerando os usuários observados na Tabela 1, é possível aprofundar o entendimento sobre a propagação de conteúdo ocorrida no *Twitter* a respeito do assunto investigado. Percebe-se que a comunidade 3 é formada por veículos de mídia tradicionais e jornalistas a eles associados, sendo assim, é coerente que esse tipo de mídia seja preferido nessa comunidade. A comunidade 4, que tem entre seus usuários mais importantes diversos cientistas e divulgadores científicos, também prioriza a utilização de mídias mais tradicionais, possivelmente devido à credibilidade conquistada por esses veículos diante de um maior público que as consome. Na comunidade 2, as mídias *mainstream* são utilizadas como fonte em menos da metade (aproximadamente 47%) das URLs apresentadas na Figura 2 e um número ainda menor é observado para a comunidade 1 (aproximadamente 28%).

Na comunidade 1 onde observa-se uma maior utilização de fontes do tipo alternativa, correspondendo a cerca de 57% do total de fontes apresentadas. Um número um pouco menor, mas ainda alto, de fontes alternativas (aproximadamente 47%) pode ser observado para a comunidade 2. Quando consideramos os usuários das comunidades 1 e 2, formados por figuras com posicionamentos ideológicos mais claros,

é possível encontrar uma justificativa para a grande utilização de mídias alternativas, que têm linhas editoriais mais ideologicamente definidas e menos pretensamente isentas que veículos tradicionais. Como exemplo notório, é possível apontar os veículos `jornalcidadeonline`, de viés declaradamente conservador, utilizado pela comunidade 1 e o `diariodocentrodomundo`, de viés declaradamente progressista, utilizado pela comunidade 2. Já as comunidades 3 e 4 apresentam 20% de fontes, entre as apresentadas, classificadas como mídia alternativa.

Entre as mídias classificadas como plataforma, é interessante notar que todas as comunidades estudadas apresentam o YouTube (`youtube.com`) entre as URLs mais utilizadas, o que aponta a importância de um estudo mais detalhado sobre os canais compartilhados nesse serviço para viabilizar uma discussão mais aprofundada. O Periscope (`pscp.tv`), serviço de *streaming* atualmente descontinuado, também aparece entre as plataformas utilizadas na comunidade 1. Com isso, é possível afirmar que os resultados apresentados pela Figura 2 aprofundam significativamente a compreensão das comunidades topológicas descritas na Tabela 1 e que, por isso, ajudam a responder à QP1.

Uma comparação das URLs mais compartilhadas entre dois pares de comunidades (1×2; 3×4) é apresentada pela Figura 3, com o objetivo de melhor distinguir o conteúdo compartilhado por cada par. Cada ponto representa uma URL (identificada pelo seu rótulo) e cada eixo representa a porcentagem de compartilhamento da URL em uma comunidade (também identificada pelo seu rótulo). A proximidade de um ponto a um eixo indica o quão exclusiva é aquela fonte para a respectiva comunidade (em relação à outra). Pontos próximos à linha diagonal indicam que a URL tem importância similar nas duas comunidades comparadas.

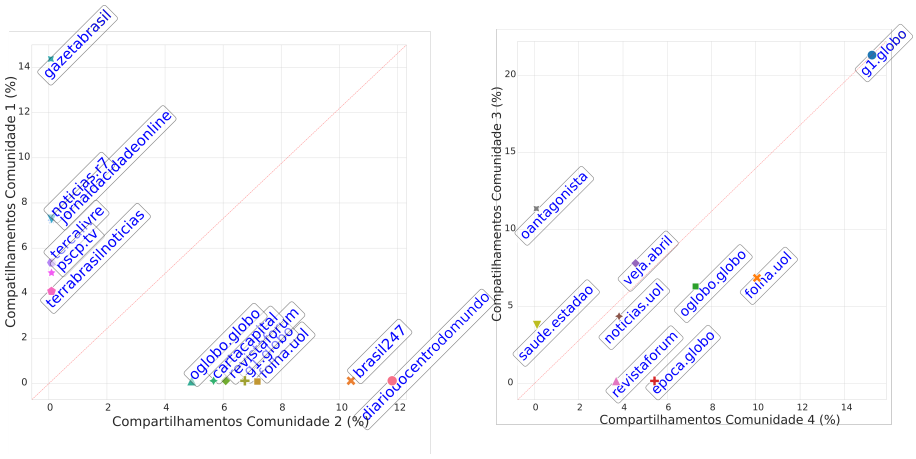

Figura 3. Comparação de compartilhamento URLs: comunidade 1×2 e 3×4.

Observando a comparação entre as comunidades 1 e 2 (painel esquerdo da Figura 3), é possível perceber que, entre as fontes apresentadas, não há nenhuma que seja usada por ambas as comunidades. Essa baixa sobreposição de fontes externas utilizadas é um forte indicativo da polarização do debate ocorrido nessas comunidades e da formação de câmaras de eco. Um resultado bastante diferente pode ser observado na comparação entre as comunidades 3 e 4 (painel direito da Figura 3). É possível observar que o *website* `g1` representa a fonte mais relevante para ambas as comunidades. Mas mesmo outros *websites* podem ser observados próximos à linha diagonal, como

`veja`, `noticias.uol` e `oglobo`. Alguns outros *websites* podem ser observados como fontes exclusivas de uma única comunidade, como é o caso de `oantagonista` e `estadao` para a comunidade 3; e `revistaforum` e `epoca.globo` para a comunidade 4. De qualquer forma, pode-se observar uma polarização muito menos acentuada na comparação entre as comunidades 3 e 4 do que entre as comunidades 1 e 2. Os resultados apresentados pela Figura 3 permitem que se avance nas respostas às QP1 e QP2, já que, além de aprofundar a compreensão da forma como os usuários de diferentes comunidades se comportam em torno de um determinado assunto, ainda possibilita a distinção da forma como se dá a utilização de fontes externas pelas diferentes comunidades. Especialmente, o painel esquerdo da Figura 3 permite caracterizar de maneira clara a polarização do debate *online*, não só no discurso, mas em elementos que sustentam esse debate.

4.3. Classificação do viés das URLs

Uma das formas mais esclarecedoras de compreender as atuações de comunidades de usuários de redes sociais é através da análise do viés político e ideológico adotado pelo corpo editorial das fontes por eles utilizadas em suas discussões. As análises dos vieses das URLs têm como principal intuito compreender melhor o ecossistema de fontes de notícias externas ao Twitter presentes no assunto analisado. A classificação das URLs externas é baseada principalmente no resultado disponibilizado por Guimarães [Guimarães et al. 2020], gerado a partir da classificação de páginas do Facebook.

A Figura 4 apresenta a distribuição da classificação obtida como resultado da avaliação das 80 URLs mais compartilhadas, classificadas entre "centro", "esquerda", "direita" e "plataforma". Há, ainda, algumas URLs que não puderam ser classificadas.

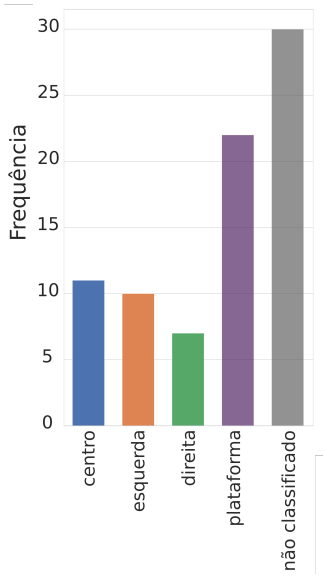

Figura 4. Distribuição da classificação das URLs quanto ao viés.

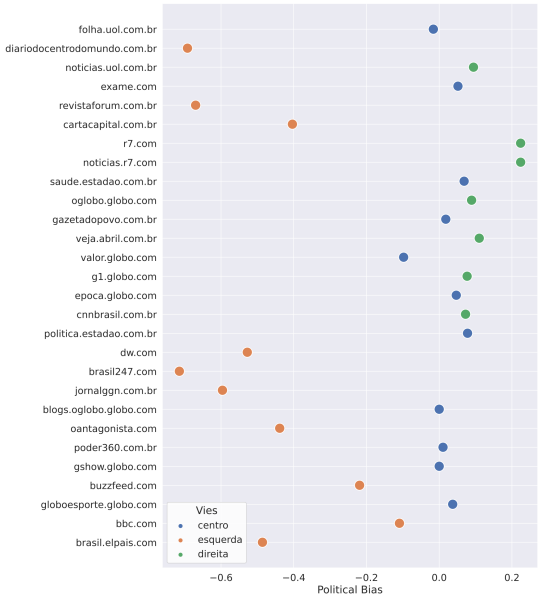

Figura 5. Classificação quanto o Viés (*Political Bias*) das URLs mais compartilhadas. (-1: esquerda, 0: centro e 1: direita).

Podemos identificar que grande parte das URLs ficaram classificadas como "plataforma", como é o caso de URLs que apontam para o Facebook e para o YouTube. Dentre

os outros *websites*, é possível identificar uma maior associação ao espectro de centro, seguida da esquerda e direita. No entanto, é importante destacar que uma grande parte das URLs não puderam ser classificadas utilizando a metodologia proposta.

A Figura 5 apresenta uma amostra com 28 páginas que foram classificadas quanto ao seu viés (*Political Bias*) e ao seu *score*. Podemos observar a classificação e a presença de veículos de diferentes correntes ideológicas indicando uma diversidade de visões e posicionamentos que podem influenciar as discussões nos assuntos analisados. No entanto, é importante destacar que grande parte das URLs não puderam ser classificadas. Quando observamos a classificação de viés político (Figura 5) em conjunto com a frequência de URLs em cada uma das comunidades (Figura 5) podemos perceber que a comunidade 1 apresenta um maior alinhamento com veículos de mídia mais classificados no espectro da direita, como o `noticias.r7`, `cnnbrasil` e `gazetadopovo`. É importante mencionar que uma série de veículos amplamente utilizados como fonte na comunidade 1 não tem classificação, como é o caso da `gazetabrasil`, do `jornaldacidadeonline` e `tercalivre`. Por outro lado, é possível perceber que os veículos utilizados pela comunidade 2 têm um alinhamento mais claro com o espectro da esquerda, segundo o *Political Bias*, como o `diariodocentrodomundo` e o `brasil247`. Uma distinção menos clara é observada entre as comunidades 3 e 4, que apresentam como fontes veículos classificados como de esquerda e de direita. Entretanto, observa-se que em ambas comunidades, mas especialmente na comunidade 3, há uma predominância de veículos de direita e que veículos de esquerda são fontes menos frequentes. Algumas distinções de fontes podem ser observadas entre a comunidades 3 (que apresenta entre suas principais fontes os veículos `valor.globo` e `cnnbrasil`) e a comunidade 4 (que apresenta entre suas principais fontes os veículos `revistaforum` e `cartacapital`). Esses resultados nos ajudam a responder a QP2, já que é possível observar uma distinção bastante clara na forma que os usuários de cada comunidade utilizam fontes externas para amparar as argumentações por eles sustentadas em seus debates. Além disso, esses resultados nos ajudam a completar a resposta à QP1, já que apenas pela análise da topologia da rede e pela modelagem dos tópicos utilizados nos conteúdos dos *tweets*, esse tipo de conhecimento sobre as comunidades de usuários não seria possível de ser obtido, mais uma vez reforçando a importância da metodologia apresentada neste trabalho.

5. Conclusões e trabalhos futuros

Este trabalho apresenta uma metodologia que permite a investigação do impacto da análise do compartilhamento das URLs para a compreensão da forma como diferentes grupos de usuários de comportamento com alinhamento similar se organizam e conduzem o debate em torno de um determinado assunto em uma rede social. Duas Questões de Pesquisa (QP) foram levantadas: QP1) A análise das URLs compartilhadas adiciona complexidade à compreensão das comunidades de usuários no *Twitter* feita através da análise da estrutura topológica e dos conteúdos dos *tweets*? QP2) É possível observar um comportamento distinto entre as comunidades em relação às URLs por elas compartilhadas? Tomando como base um estudo de caso com termos relacionados à vacinação da COVID-19 no Brasil ocorrido no *Twitter*, às duas questões levantadas puderam ser melhor compreendidas através da metodologia proposta. Em relação à QP1, é possível dizer que a análise de URLs é capaz de complementar a análise da discussão conduzida pelos usuários das comunidades, já que além de identificar grupos de usuários, que taci-

tamente sabe-se que pertencem a alinhamentos ideológicos distintos, foi possível identificar padrões na maneira como utilizam fontes externas de informação e em seus vieses políticos. Além disso, foi possível avançar no entendimento sobre a QP2, já que é possível observar que grupos mais abertamente posicionados em relação a alinhamentos políticos utilizam com mais frequência mídias alternativas de linha editorial mais claramente posicionada no espectro político, enquanto grupos compostos por usuários de posicionamento político menos claro utilizam como fontes veículos mais pretensamente ou declaradamente isentos. Em trabalhos futuros, é preciso avançar na caracterização de veículos que não apresentaram classificação por Guimarães *et al.* [Guimarães et al. 2020], além de aplicar a metodologia aqui apresentada em outros contextos. Além disso, outros métodos para identificação de comunidades e modelagem de tópicos podem ser explorados, tirando proveito da flexibilidade fornecida pela metodologia apresentada neste trabalho.

Referências

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Christhie, W., Reis, J. C. S., Moro, F. B. M. M., and Almeida, V. (2018). Detecção de posicionamento em tweets sobre política no contexto brasileiro. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*, Porto Alegre, RS, Brasil. SBC.

Cossard, A., Morales, G. D. F., Kalimeri, K., Mejova, Y., Paolotti, D., and Starnini, M. (2020). Falling into the echo chamber: the italian vaccination debate on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 130–140.

Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, pages 913–922.

Guimarães, S. S., Reis, J. C., Lima, L., Ribeiro, F. N., Vasconcelos, M., An, J., Kwak, H., and Benevenuto, F. (2020). Identifying and characterizing alternative news media on facebook. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 448–452. IEEE.

Martins, E., Gonçalves, K., and Filho, R. M. (2019). Caracterizando a campanha presidencial brasileira em 2018 usando dados do twitter. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*, pages 131–142, Porto Alegre, RS, Brasil. SBC.

Morstatter, F., Shao, Y., Galstyan, A., and Karunasekera, S. (2018). From alt-right to alt-rechts: Twitter analysis of the 2017 german federal election. In *Companion Proceedings of the The Web Conference 2018*, pages 621–628.

Resende, G., Messias, J., Silva, M., Almeida, J., Vasconcelos, M., and Benevenuto, F. (2018). A system for monitoring public political groups in whatsapp. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pages 387–390. ACM.