Submission date: 11/11/2022
Camera ready submission date: 07/08/2024
1st round notification: 28/05/2023
Available online: 15/08/2024
Publication date: 15/08/2024

Section: Extended version
Extended version of selected papers from the XI Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2022)

Investigação de fontes de informação para a caracterização de usuários em Redes Sociais Online

Investigation of information sources for the characterization of users in Online Social Networks

Carlos M. G. Barbosa[1], **Lucas G. da S. Félix**[2], **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

Abstract. *Individuals interact in a complex way for discussions in online social networks and the characterization of the way in which their are organized is essential for understanding the debate that take place around specific issues. This work presents a methodology for the characterization of discussions in online social networks and the groups of users that promove them based on the addressed topics and the external sources used by their users for the construction of the argumentation conducted on specific issues. An analysis of Twitter around discussions about the Federal Supreme Court (STF) and the COVID-19 vaccination in Brazil shows that the methodology is able to advance in the understanding of the way in which information is produced and propagated, and in the differences between the way various groups of individuals use external sources of information.*

Keywords. *Social networks; information sources; Twitter; communities; topic modeling.*

Resumo. *Indivíduos interagem de maneira complexa para promover discussões em redes sociais online e a caracterização da forma como se dá sua organização é essencial para a compreensão do debate ocorrido em torno de determinados assuntos. Este trabalho apresenta uma metodologia para a caracterização de discussões em redes sociais e os grupos de usuários que as promovem com base nos tópicos abordados e nas fontes externas utilizadas pelos seus usuários para a construção da argumentação conduzida sobre determinados assuntos. Uma análise do Twitter em torno de discussões sobre o Supremo Tribunal Federal*

Cite as: Barbosa, C. M. G., Félix, L. G. S., Alves, A. P. S., Xavier, C. R. & Vieira, V. F. (2024). Investigation of information sources for the characterization of users in Online Social Networks (Investigação de fontes de informação para a caracterização de usuários em Redes Sociais Online). iSys: Revista Brasileira de Sistemas de Informação (Brazilian Journal of Information Systems), Vol 17(1), 9:1-9:28. DOI: 10.5753/isys.2024.3019

(STF) e 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.

Palavras-Chave. *Redes sociais; fontes de informação; Twitter; comunidades; modelagem de tópicos.*

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 caracterização das redes sociais 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 este objetivo em diferentes ambientes e sob diferentes perspectivas. Por exemplo, é possível citar o trabalho de Morstatter *et al*. [Morstatter et al. 2018], que apresenta a caracterização da forma como grupos de direita se organizam em torno das eleições na Alemanha e o trabalho de Resende *et al.* [Resende et al. 2018], que tem como objetivo monitorar grupos de WhatsApp para investigar a propagação de informação sobre as eleições no Brasil.

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 URLs (*Uniform Resource Locators*), por eles compartilhadas, endereços que apontam para *websites* externos à plataforma *Twitter* e que muitas vezes contêm informações que são utilizadas pelos usuários para compor os conteúdos compartilhados.

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 em [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.

Neste contexto, este trabalho tem como objetivo responder três Questões de Pesquisa (QPs): **QP1) É possível observar uma distinção do conteúdo discutido dentro de cada uma das comunidades em relação à discussão no conjunto completo de usuários?** Para responder à QP1, os tópicos discutidos serão modelados considerando o contexto global de todos os usuários envolvidos na discussão em torno de determinados temas e dentro das comunidades isoladamente, visando identificar similaridades e distinções entre o conteúdo abordado em cada contexto. **QP2) 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*?** A QP2 busca investigar até que ponto a metodologia proposta para análise das URLs como fonte de conteúdo pode beneficiar a tarefa de análise de redes sociais online, visando identificar potencialidades e limitações dessa estratégia. **QP3) É possível observar um comportamento distinto entre as comunidades em relação às URLs por elas compartilhadas?** Para responder à QP3, será feita uma avaliação das diferenças observadas nas URLs encontradas na discussão em cada uma das comunidades e na base de dados considerando todos os usuários, visando a identificação de padrões de comportamento e a ocorrência de uma dominância de alguma comunidade específica sobre a discussão de maneira mais geral.

O presente trabalho é uma extensão de um trabalho anterior ([Barbosa et al. 2022]), publicado nos Anais do XI Brazilian Workshop on Social Network Analysis and Mining. Diferentemente do trabalho original, neste trabalho são exploradas as similaridades e distinções entre o conteúdo discutido e as URLs que sustentam essa discussão na base de dados completa, permitindo uma visualização mais clara da aplicabilidade da metodologia proposta para caracterização e análise do conteúdo discutido em redes sociais *online*.

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. Considerando duas bases de dados, que tratam de discussões no Twitter sobre dois assuntos distintos, a vacinação da COVID-19 no Brasil (abordada também no trabalho do qual este se estende [Barbosa et al. 2022]) e o Supremo Tribunal Federal (tratada originalmente no presente trabalho), observa-se que algumas comunidades, apesar de se assemelharem em alguns aspectos, nitidamente têm visões diferentes – muitas vezes divergentes – sobre o assunto, utilizando a rede social 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. [Caetano et al. 2018] apresentam um estudo do sentimento dos usuários engajados durante a eleição presidencial de 2016 nos Estados Unidos no Twitter, no qual foram analisados 23 milhões de *tweets* publicados por 115 mil usuários entre janeiro

e novembro de 2016. [Christhie et al. 2018] identificam posicionamentos de usuários do Twitter favoráveis e contrários a candidatos na corrida eleitoral de 2018 no Brasil. [Morstatter et al. 2018], por exemplo, apresentam uma caracterização de grupos de extrema-direita no Twitter considerando a eleição na Alemanha no ano de 2017. [Resende et al. 2018] apresentam um modelo de monitoramento e caracterização de dados propagados no WhatsApp em que são monitoradas as mensagens 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. 2018] definem que este fenômeno pode ser apresentado em dois componentes distintos. Primeiro, a câmara é a rede formada em torno de usuário que irá receber a opinião apresentada por outros usuários e o Eco é um retorno confirmatório desta opinião, condizente com a opinião compartilhada pelo usuário receptor, empobrecendo a construção de um pensamento crítico, para o qual é essencial a existência do contraditório. [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, classificados como apoiadores do processo de vacinação, contrários às vacinas, e um terceiro grupo de contas interessadas no debate da vacinação de animais.

Um aspecto interessante a ser explorado por trabalhos que tomam como base as RSO para compreender fenômenos sociais de propagação de informações e ideias é caracterizar quais fontes de informações externas às redes sociais são utilizadas para sustentar a argumentação entre os indivíduos. Nesse sentido, é possível investigar a presença de mídias hiper-partidárias, que divulgam somente o conteúdo alinhando a determinadas figuras políticas, partido ou espectros ideológicos [Bhatt et al. 2018, Recuero et al. 2020]. Estes veículos apresentam notícias utilizando uma linguagem agressiva quando comparado com os veículos tradicionais, sendo amplamente propagadas pelas redes sociais [Rae 2021]. Veículos hiper-partidários estão presentes no Twitter, recebendo uma maior importância em períodos eleitorais, como as eleições de 2018 no Brasil. Nesse período, esses veículos foram amplamente difundidos, muitas vezes impulsionados por equipes de *marketing* direcionadas a campanhas políticas, resultando em uma grande polarização e em ambiente favorável a propagação de notícias falsas ou com uma significativa manipulação [Recuero et al. 2020].

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. 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. A Figura 1 apresenta uma visão geral da metodologia, dividida pelos seguintes passos. 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* (painel 1). A partir disso, a rede é gerada e alguns passos para sua análise topológica são seguidos (painel 2), 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* (painel 3), que incluem o pré-processamento dos textos, a modelagem dos tópicos e a análise das URLs.

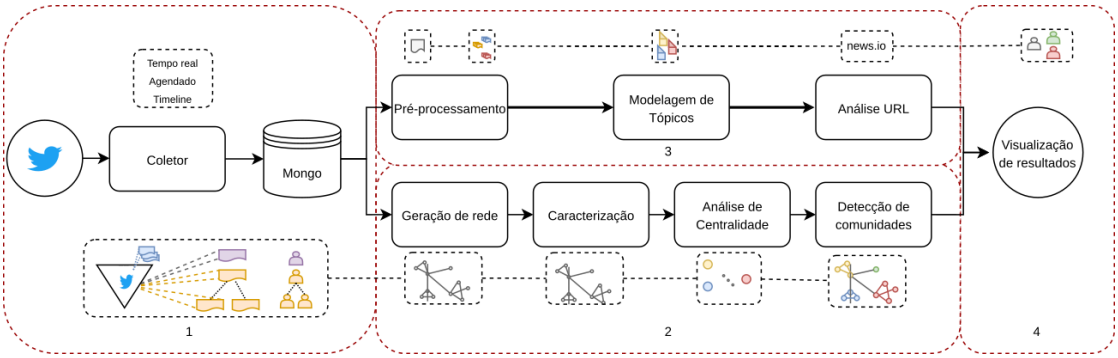


Figura 1: Fluxo básico da metodologia: **1)** Coleta e armazenamento de dados; **2)** Análise das redes; **3)** Análise do conteúdo; **4)** Visualização dos resultados

Os passos apresentados na Figura 1 são discutidos com mais detalhes nas próximas seções.

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. Todas as etapas do processamento são focadas para limpeza de conteúdo escritos em português.

3.2. Geração da rede

Para este trabalho, optou-se pelo desenvolvimento de uma rede baseada em *retweets*, embora outros tipos de redes pudessem ser facilmente incorporados. Os nós representam os usuários do *Twitter* envolvidos na discussão e as arestas representam uma relação de

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

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 compartilhou 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 [Ferreira et al. 2019, Nobre et al. 2020]. Tomando como base a versão não-direcionada da rede, neste trabalho, é utilizado o algoritmo de [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 dado o potencial tamanho da rede gerada.

3.4. Análise de conteúdo

Além da estrutura das redes, a metodologia permite que sejam analisados os conteúdos dos *tweets*, sob a perspectiva de análise de tópicos, utilizando a abordagem do *Latent Dirichlet Allocation* (LDA) [Blei et al. 2003]. Por padrão são extraídos 6 tópicos, em que os tópicos e termos encontrados podem ser apresentados em uma nuvem de palavras, ou em uma lista de relevância dos termos. 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

É possível afirmar que, além do próprio conteúdo produzido e compartilhado por indivíduos em redes sociais, é importante estudar as fontes utilizadas pelas pessoas para sustentar a argumentação para que seja possível compreender a forma como os indivíduos se organizam e se posicionam em redes sociais *online*.

Este trabalho apresenta um processo de classificação quanto ao tipo de mídia e viés das URLs compartilhadas, considerando os *tweets* e as comunidades de interesse. Para isso, considerando cada uma das comunidades investigadas, são extraídas as URLs dos *tweets* coletados. 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. Em seguida, as URLs são classificadas quanto ao tipo (*mainstream media*, mídia alternativa, ou plataforma) e viés político (esquerda, centro ou direita). Uma visão do fluxo das análises também pode ser observada na Figura 2.

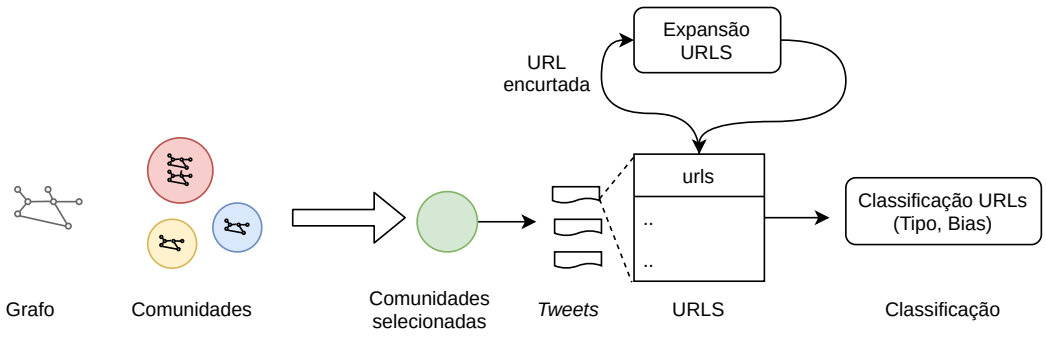


Figura 2: Fluxo da extração, expansão e classificação de URLs de comunidades selecionadas.

Primeiramente as URLs de *websites* que não estão registrados em uma organização oficial de imprensa como Associação Nacional de Jornais (ANJ), 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*, como apresentado por [Guimarães et al. 2020]. A abordagem foi ajustada para incluir duas novas classificações, uma referente a plataformas de distribuição e a outra referente a mídias digitais alternativas. O primeiro caso inclui Spotify, YouTube e redes sociais online, e receberam a nomenclatura de "plataforma". Já o segundo caso contempla os *websites* que estavam registrados na Associação de Jornalismo Digital (AJOR), e 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 esta classificação, foi realizada uma expansão dessa classificação utilizando como base a sobreposição de audiência dos *websites*, disponibilizada pela plataforma Alexa[3]. Essa classificação foi realizada utilizando os *websites* de maior sobreposição de audiência, como apresentado na Figura 3, sendo que caso se encontre uma classificação majoritária a partir dos quatro *websites*, essa atribuição é replicada para o site sem classificação.

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

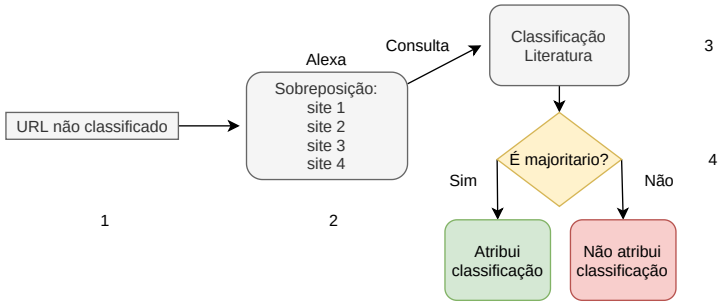


Figura 3: Classificação dos *websites*; 1) URL não classificada; 2) URLs com maior sobreposição; 3) Obtém da literatura a classificação dos *websites* sobrepostos; 4) Atribui a classificação majoritária para a URL não classificada.

4. Resultados e discussão

Esta seção apresenta, além de uma descrição dos experimentos realizados, uma discussão quanto às suas aplicações e limitações quando aplicada a metodologia apresentada na Seção 3 considerando as discussões em dois contextos distintos: o Supremo Tribunal Federal (STF), o órgão legislativo máximo no Brasil, que será tratado como `STF` e a vacinação da COVID-19, que será tratado como `vacinação`. A escolha do assunto `STF` teve como motivação os intensos debates ocorridos sobre esse órgão, gerados por uma desconfiança de sua atuação promovida, especialmente, pelo então governo federal e outros agentes ligados a ele [Cruz 2020]. A escolha do assunto `vacinação` teve como motivação, além de sua enorme relevância no contexto da pandemia da COVID-19 naquele momento de coleta, o significativo engajamento de indivíduos que se agrupam em intenções e visões ideológicas nitidamente distintas [4]. Ambos assuntos escolhidos têm grande potencial para a geração de polarização no debate, já que apresentam figuras públicas bastante conhecidas com posicionamentos antagônicos amplamente defendidos.

4.1. Análise dos tweets sob uma perspectiva global

Utilizando o termo `STF`, foram coletados 914914 *tweets/retweets* sobre o assunto entre os meses de maio a agosto de 2020. Em relação ao assunto `vacinação`, foram coletados 1779024 *tweets/retweets* entre os meses de janeiro a fevereiro de 2021. Foram utilizados para coleta os termos frequentemente relacionados ao debate sobre a vacinação da COVID-19 no Brasil na época da coleta, como `vacinaja, vacina, coronavac, vemvacina, vacinaparatodos`[5].

4.1.1. Modelagem de tópicos global

A metodologia aplicada neste trabalho permite combinar a investigação da estrutura topológica das redes geradas a partir dos *retweets* com uma análise dos conteúdos pro-

[4] `https://www.bbc.com/portuguese/brasil-53993365`

[5] O conjunto completo de termos utilizados, assim como a base de dados coletada estão disponíveis em `https://datastudio.google.com/reporting/17221eaf-b214-4a40-b36c-1572708e9071`

duzidos pelos usuários envolvidos na discussão de assuntos específicos baseada na modelagem de tópicos, como apresentado na Seção 3. De forma a permitir uma melhor contextualização dos resultados apresentados, a presente seção realiza uma análise dos assuntos de uma forma global, ou seja, considerando todos os usuários identificados na discussão de um determinado assunto.

O principal objetivo da aplicação de um algoritmo de modelagem de tópicos é auxiliar a compreensão dos principais tópicos levantados em torno dos assuntos analisados e seus termos, para que seja possível identificar correntes ideológicas distintas que se organizem em torno dos assuntos. Para isso, foi utilizado o algoritmo *Latent Dirichlet Allocation* (LDA), um algoritmo probabilístico de modelagem de tópicos que permite resumir uma grande quantidade de texto em tópicos. É importante destacar que outros algoritmos para modelagem de tópicos foram experimentados, como *Non-Negative Matrix Factorization* (NMF), não oferecendo vantagem em relação ao LDA.

Os resultados encontrados para a modelagem de tópicos para os assuntos `STF` e `vacinação` são apresentados nas Figuras 4 e 5, respectivamente, que apresentam os 6 tópicos mais importantes sobre cada assunto e os 10 principais termos de cada tópico.

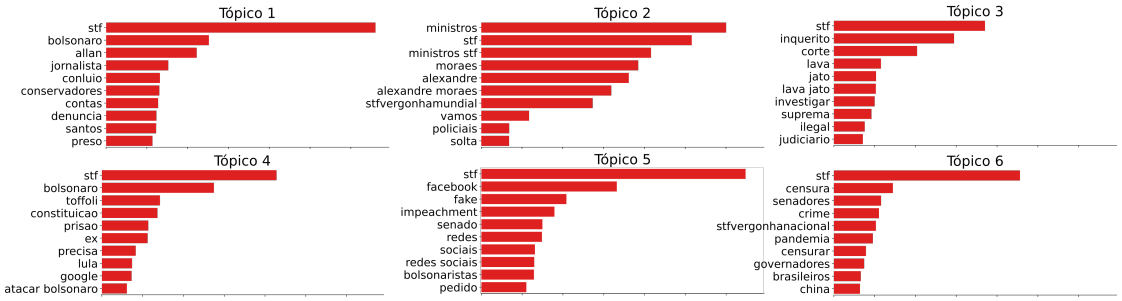


Figura 4: Modelagem de tópicos utilizando LDA assunto `STF`.

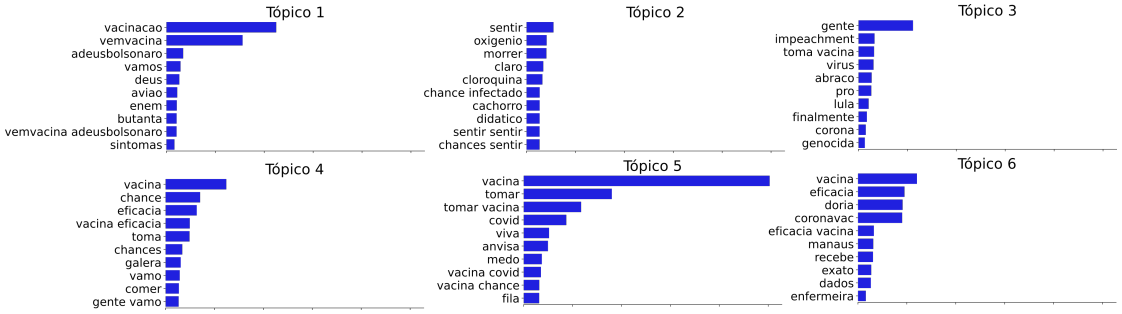


Figura 5: Modelagem de tópicos utilizando LDA assunto `vacinação`.

Os tópicos encontrados pelo LDA em relação ao assunto `STF` (Fig. 4) permitem a identificação de diversos tópicos que determinam posicionamentos contrários ao Supremo Tribunal Federal do Brasil que remetem a sua atuação na época da coleta dos dados[6].

[6]`https://www.bbc.com/portuguese/brasil-52827790`

O Tópico 1 faz referência ao blogueiro conservador Allan dos Santos (termos `allan` e `santos`), que teve sua prisão determinada pelo ministro Alexandre de Moraes em decorrência de investigações que apontaram sua atuação na organização de movimentos antidemocráticos, dos quais obteve vantagem econômica. Outros tópicos possuem termos ainda mais explícitos em relação a ataques ao STF, como o `stfvergonhamundial` e variações como `vergonha`, presentes em diversos tópicos. Há ainda uma "evocação" por liberdade de expressão presente no discurso observada em termos como `censura`, `censurar` e `democracia` e uma "denúncia" a uma suposta "ditadura", como pode ser observado em termos como `ditadura stf` e `imprensa`. O Tópico 6 nos ajuda a sumarizar uma teoria de usuários apoiadores do governo federal de que há um crime de censura – termos `crime censura` – combinado a um boicote ao governo federal provocado pelos governadores e os senadores – termos `governadores` e `senadores` – no combate à pandemia – termo `pandemia` – financiado pela China – termo `china`.

Em relação ao assunto `vacinação` (Fig. 5), é também importante enfatizar o recorte histórico no qual se encontra a coleta dos *tweets* realizada, nos meses de Janeiro e Fevereiro de 2021. Primeiramente, destaca-se que essa foi a época de um colapso do sistema de saúde na cidade de Manaus, no norte do Brasil, na ocasião de um súbito aumento de casos graves de COVID-19 que ocasionou uma grave escassez de cilindros de oxigênio e equipes médicas capazes de atender aos pacientes, provocando muitas mortes [7]. Nesse contexto, é possível observar os termos `oxigenio`, `morrer` e `chance infectado`. Há também a presença do termo `cachorro` que faz referência a uma fala do então presidente Jair Bolsonaro que, ao ser questionado sobre uma doação de cilindros de oxigênio da Venezuela ao Brasil, insinuou que na Venezuela não há cachorro pois a população comeu todos e que o seu governo não deveria se preocupar com o Brasil. A época da coleta dos dados foi também marcada pela aprovação da vacina "Coronavac"[8], fabricada pelo laboratório Chinês Sinovac em parceria com o instituto público Butantã, sediado na cidade de São Paulo, que provocou uma clara polarização no debate no Twitter, como é possível observar nos termos identificados. Alguns usuários mostravam entusiasmo pela vacina em termos como `vemvacina`, `finalmente`, `vacina aprovada` e `tomar vacina`. Por outro lado, outros usuários propagavam discursos anti-vacina e apoiando tratamentos alternativos comprovadamente ineficazes através de termos como `cloroquina`, `tratamento precoce` e `tratamentoprecocesalvavidas`. A época da aprovação da vacina foi marcada por discussões técnicas, embora rasas, sobre as vacinas, o que é bastante incomum entre pessoas que não estão diretamente envolvidas em seu processo de elaboração e viabilização, mas que pode ser visto em termos como `eficácia` e anvisa (a agência brasileira que regula, entre outras coisas, fármacos e vacinas no país). É impossível isolar a discussão em torno da vacinação dos elementos políticos que a envolvem, o que fica claro na Fig. 5. A polarização em torno da aceitação e do entusiasmo em relação à vacina se reflete como uma polarização entre os apoiadores do então presidente Jair Bolsonaro, abertamente defensor de tratamentos ineficazes – com termos como `tratamento precoce`, `medico`, `pazuello` (em referência ao então

[7] `https://www.bbc.com/portuguese/brasil-55674229`

[8] `https://agenciabrasil.ebc.com.br/saude/noticia/2021-01/vacinacao-contra-covid-19-come%C3%A7a-em-todo-o-pais`

ministro da saúde) e `india` (país onde havia uma negociação do Governo Federal para compra de vacinas, posteriormente fracassada) – e os apoiadores de João Dória, governador do Estado de São Paulo, responsável pela gestão do Instituto Butantã, onde é produzida a vacina Coronavac – com termos como `coronavac`, `doria` e `enfermeira`, em referência à primeira pessoa a receber a vacina no Brasil. Há também uma série de termos ridicularizando uma fala do presidente à época Jair Bolsonaro que afirmou que ninguém se responsabilizaria caso alguém que tomasse a vacina virasse um jacaré – com termos como `jacare`, `virar jacare`. Outros termos fazem críticas e referem-se a pedidos para impeachment do presidente, como `genocida`, `adeusbolsonaro` e `impeachment` e mencionam o então ex-presidente Luiz Inácio Lula da Silva, como o termo `lula`.

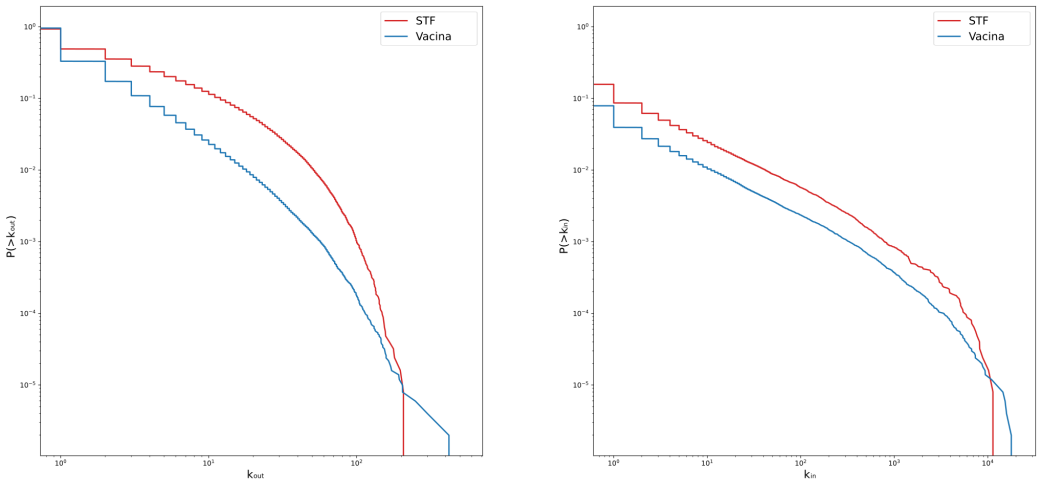
4.1.2. Caracterização básica das redes de *retweets*

A partir da coleta das bases de dados, foi possível gerar as redes, seguindo a metodologia apresentada na Seção 3.2, das quais as características básicas das componentes gigantes são apresentadas na Tabela 1. O valor alto no grau médio da rede relacionada ao assunto

	STF	**Vacinação**
Número de nós	124636	502690
Número de arestas	608741	1067358
Número de *tweets*	914914	1779024
Grau médio	9.7650	4.2466

Tabela 1: Características básicas da rede direcionada ponderada

`STF` é um indicador de que esse é um assunto que gerou um maior engajamento dos usuários no período analisado e reforça a importância de se estudar esse assunto. Com o objetivo de aprofundar a investigação sobre os graus das redes, as Figuras 6a e 6b apresentam as distribuições dos graus para as redes de ambos assuntos analisados.



(a) Grau de saída. (b) Grau de entrada.

Figura 6: Distribuição de graus.

Observando a Figura 6a é possível identificar uma distinção nas curvas dos graus por ordens de grandeza, principalmente na sua região mais central. Os gráficos de distribuição de graus apresentado nas Figuras 6a e 6b permitem identificar que a rede direcionada ponderada possui um pequeno conjunto de usuários com grau muito superior à média dos usuários. A distribuição dos graus de saída mostra que, apesar da rede `STF` apresentar uma maior frequência de graus na parte central da distribuição, a rede `vacinação` apresenta uma cauda mais longa.

Os resultados apresentados nas Seções 4.1.1 e 4.1.2 permitem que seja feita uma análise da forma como os usuários se organizam e do conteúdo discutido em torno de assuntos específicos no Twitter sob uma perspectiva global. Entretanto, uma visão muito mais rica pode ser obtida quando a investigação é feita sob uma perspectiva mais local, o que será apresentado a seguir.

4.2. Análise dos tweets sob uma perspectiva de comunidades

Além da investigação de assuntos específicos sob uma perspectiva global, a metodologia proposta neste trabalho permite uma exploração em uma perspectiva mais local, enriquecendo o conhecimento da forma como usuários com visões semelhantes se organizam em torno desses assuntos. Neste trabalho, são considerados usuários de visões semelhantes aqueles que sejam identificados na mesma comunidade na rede de retweets.

Considerando o método de Louvain para detecção de comunidades baseado na otimização da modularidade, foram encontradas partições com modularidades $Q = 0,3251$ e $Q = 0,6571$ para as redes dos assuntos `STF` e `vacinação`, respectivamente. À primeira vista, pode-se considerar que, embora a partição da rede `vacinação` apresente uma modularidade razoavelmente alta, a modularidade da partição da rede `STF` é bastante baixa, o que pode ser um fato desencorajador para o estudo das redes sob uma perspectiva de comunidades. Entretanto, a observação da Figura 7, que apresenta a

distribuição acumulada complementar das modularidades obtidas para cada comunidade das partições isoladamente permite um entendimento mais aprofundado sobre as redes.

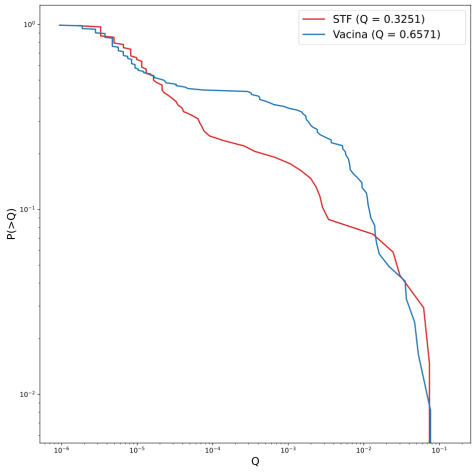


Figura 7: Distribuição da modularidade para cada comunidade observada nas partições identificadas para as redes de *retweets*.

Ambas curvas da Fig. 7 apresentam intervalos de distribuição muito similares, com comunidades com modularidades desde na ordem de 10^{-6} até quase 10^{-1}. Entretanto, é possível perceber uma queda mais acentuada na rede `STF`, indicando que há um número consideravelmente menor de comunidades com modularidade mais alta quando comparada à rede `vacinação`. Assim, é possível que a diferença na parte central das distribuições reflita diretamente no valor observado para a modularidade da partição como um todo. De qualquer modo, independentemente do valor de modularidade global observado para as partições das redes, é possível afirmar que em ambos assuntos estudados, as partições apresentam suas principais comunidades como bastante modulares, indicando um forte alinhamento de opiniões, segundo a suposição feita neste trabalho, o que representa uma forte justificativa para seu uso como base para a análise dos conteúdos dos tweets de maneira mais local. Nesse sentido, podemos argumentar que é razoável concentrar a análise de cada assunto nos usuários das comunidades mais modulares, que representam os usuários com posicionamento mais coeso sobre um determinado tema, o que é apresentado nas próximas seções.

4.2.1. Caracterização topológica das comunidades

Como discutido na Seção 4.2, utilizando o método de Louvain, as estruturas de comunidades para as redes de *retweets* `STF` e `vacinação` foram identificadas e, nesta seção, será conduzida uma análise com base apenas em algumas comunidades de maior interesse. Especificamente, as comunidades analisadas são as de maior modularidade para cada uma das redes (cinco para a rede `STF` e quatro para a rede `vacinação`), sobre as quais é possível argumentar que representam grupos de maior densidade e envolvem indivíduos com maior alinhamento.

As Tabelas 2 e 3 apresentam algumas características básicas das comunidades selecionadas para as redes: modularidade, número de usuários, número de tweets e razão de retweets. São também apresentados os dez usuários mais centrais considerando a centralidade de PageRank, que são chamados como top usuários e permitem compreender melhor o direcionamento das opiniões nas suas respectivas comunidades.

	#1	**#2**	**#3**	**#4**	**#5**
Id	0	2	1	3	4
Modularidade	0.09746	0.0738	0.0619	0.0299	0.02437
Núm. usuários	34639	17922	25961	15358	13248
Núm. *tweets*	90159	204407	228910	143661	84799
Razão *retweets*	0.8458	0.7834	0.8874	0.8888	0.8978
Top usuários					
	MarceloFreixo	MarcosQuezado1	leandroruschel	blogdojefferson	Rconstantino
	felipeneto	mitags	lpbragancabr	Biakicis	AnaPaulaVolei
	jovensreacinhas	LBonoro2	JornalBSM	oswaldojor	GFiuza_Oficial
	juliaduailibi	FabioTalhari	oiIuiz	antoniorayol	marcelvanhattem
	Adrieli_S	Jouberth19	carlosjordy	BolsonaroSP	jrguzzofatos
	GeorgMarques	nelsonpaffi	filipebarrost	PastorMalafaia	revistaoeste
	JotaInfo	RaquelStasiaki	secomvc	CarlaZambelli38	viniciuscfp82
	GuilhermeBoulos	xfischer	flaviogordon	taoquei1	paulomathias
	DanielaLima_	Ro_Moller	allantercalivre	BrazilFight	renataagostini
	RevistaCrusoe	AlanLopesRio	kimpaim	JornalDaCidadeO	roxmo

Tabela 2: Análise da rede e dos *tweets*: STF.

	#1	**#2**	**#3**	**#4**	**#5**
Id	0	1	5	2	3
Modularidade	0.1435	0.0763	0.0527	0.0470	0.0365
Núm. usuários	43736	43700	32868	41267	37521
Núm. *tweets*	196687	138190	115639	100743	48180
Razão *retweets*	0.7316	0.7101	0.6218	0.6739	0.7826
Top usuários					
	CarlaZambelli38	costa_rui	g1	oatila	eudeborabrasil
	lcoutinho	HaddadDebochado	CNNBrasil	lolaferreira	danielvvsantos
	jairbolsonaro	cartacapital	exilado	RandonadmM	AssimFalouLucas
	taoquei1	padilhando	GloboNews	luizacaires3	joaoluizsb
	leandroruschel	Debora_D_Diniz	RodrigoMaia	elsonh	MariliaMoscou
	BrazilFight	cidadaoprimata	CNNBrBusiness	EthelMaciel	motadouglas
	LaurinhaIronic	LulaOficial	gugachacra	sailorthrash	deckerjdoe
	conexaopolitica	dadourado	gusthpa	ThomasVConti	acellency
	kimpaim	reinaldoazevedo	FMouraBrasil	brgenovez	_putindesaias
	carteiroreaca	GuilhermeBoulos	folha	RenanPeixoto_	renatofrei_

Tabela 3: Análise da rede e dos *tweets*: `vacinação`.

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.

Considerando apenas informações autodeclaradas pelos usuários em sua bio no Twitter (pequeno texto de apresentação que aparece no topo de cada perfil), é possível perceber uma distinção de posicionamentos políticos em relação ao assunto STF da comunidade 1 para as demais. Os usuários classificados nas primeiras posições da comunidade 1 são os únicos dos quais não é possível apontar um claro alinhamento com o então governo federal. Alguns top-usuários da comunidade 1 são figuras políticas que constantemente defendem posicionamento contrário ao governo Bolsonaro, como MarceloFreixo (então deputado federal pelo PSOL, partido de oposição ao governo Bolsonaro) e GuilhermeBoulos (então coordenador do MTST e declarado opositor do então governo Bolsonaro). Mas também é possível encontrar na comunidade 1 jornalistas de postura notadamente mais progressista, como DanielaLima_, juliaduailibi e GeorgMarques, além de influenciadores, como felipeneto (de posicionamento declaradamente contrário ao governo Bolsonaro) e jovensreacinhas (perfil satírico de críticas ao então governo). Nas outras comunidades é fácil identificar usuários com claro alinhamento ideológico com o governo Bolsonaro. Por exemplo, a comunidade 2 é basicamente formada por usuários autoidentificados como conservadores, religiosos e defensores da família tradicional (AlanLopesRio, MarcosQuezado1, mitags, RaquelStasiaki). A comunidade 3 é formada predominantemente por políticos declaradamente conservadores e alinhados ao então governo federal, como é o caso de filipebarrost, lpbragancabr, carlosjordy; mas também possui jornalistas ligados a blogs independentes, como é o caso de leandroruschek, allantercalivre e kimpaim. Perfis semelhantes ao da comunidade 3 são observados nas comunidades 4 e 5, onde observa-se figuras políticas conservadoras, como BiaKicis, BolsonaroSP, CarlaZambelli38 e marcelvanhattem e secomvc (a própria Secretaria de Comunicação do governo federal); mas também jornalistas e influenciadores ligados a blogs e veículos independentes, como blogdojefferson, taoquei1 e BrazilFight; e jornalistas ligados a veículos da grande mídia, como Rconstantino, AnaPaulaVolei e paulomathias. Assim, de maneira resumida, é possível afirmar que há um posicionamento contrário ao então governo federal adotado pelos usuários da comunidade 1, mas um posicionamento favorável pelos usuários das comunidades 2, 3, 4 e 5, embora os usuários apresentem papéis bastante distintos.

Em relação à rede vacinação, é também possível perceber uma clara distinção de posicionamentos editoriais e políticos quando as comunidades são comparadas. A comunidade 1 é formada por indivíduos em muitas ocasiões alinhados com o então Governo Federal e com as políticas por ele conduzidas no enfrentamento da pandemia da COVID-19, como por exemplo os usuários CarlaZambelli38 (Deputada Federal fortemente alinhada ao ex-presidente Jair Bolsonaro), lcoutinho (autor com posicionamento declaradamente de direita) e jairbolsonaro (ex-presidente do Brasil). A comunidade 2, por outro lado, tem suas primeiras posições formadas por usuários abertamente contrários às políticas adotadas pelo então Governo Federal no enfrentamento da pandemia, como os usuários costa_rui (então governador do estado da Bahia, filiado e co-fundador do Partido dos Trabalhadores), HaddadDebochado (influenciador de posicionamento declaradamente contrário ao então governo federal), LulaOficial (eleito presidente em

2022 e co-fundador do Partido dos Trabalhadores) e GuilhermeBoulos. As primeiras posições da comunidade 3 são formadas, majoritariamente, por jornalistas e veículos de mídia, como g1 (um dos maiores veículos de mídia digital do Brasil), CNNBrasil (a versão brasileira da mundialmente conhecida CNN), GloboNews (um dos maiores canais de notícias do Brasil) e gugachacra (jornalista da GloboNews). Na comunidade 4 destacam-se epidemiologistas e divulgadores científicos, como os usuários oatila (divulgador científico que ganhou muita notoriedade durante a pandemia), luizacaires3 (divulgadora científica, editora da Universidade de São Paulo) e EthelMaciel (epidemiologista e Professora da Universidade Federal do Espírito Santo). De maneira resumida, e utilizando apenas informações autodeclaradas pelos usuários em sua bio no Twitter, é possível observar uma distinção das comunidades da seguinte maneira: comunidade 1 - usuários alinhados ao governo Bolsonaro; comunidade 2 - usuários contrários ao governo Bolsonaro; comunidade 3 - jornalistas e veículos de mídia declaradamente neutros; comunidade 4 - divulgadores científicos e intelectuais.

4.2.2. Modelagem de tópicos por comunidade

Com o objetivo de melhorar a caracterização das comunidades na rede de *retweets*, foi realizado um estudo dos seus principais tópicos discutidos, utilizando o método *Latent Dirichlet Allocation* (LDA) para modelagem de tópicos. Os resultados apresentados nesta seção complementam o que foi observado na Seção 4.1 para a rede como um todo, mas enfatizando as distinções entre os termos utilizados nas diferentes comunidades. Para manter a concisão dos resultados e facilitar a condução das discussões, foram selecionadas as duas comunidades de maior modularidade para as redes STF (Figura 8) e vacinação (Figura 9), das quais são apresentados os 6 tópicos principais e seus 10 termos principais.

A polarização em torno do assunto STF, especialmente no período de observação devido à operação das Fake News[9], pode ser observada na forma como se dá a distribuição dos tópicos e termos discutidos em cada comunidade. A comunidade #1, cujos usuários mais representativos são perfis relacionados a políticos e influenciadores abertamente contrários ao então governo federal no Brasil utiliza frequentemente termos que fazem referência a ameaças contra ministros do STF feita por figuras de extrema-direita, como os termos alexandre e moraes (em referência ao ministro do STF Alexandre de Moraes). Há ainda o uso de termos comotoffoli, que remete ao então presidente do STF Dias Toffoli e a ação que determinou a abertura do inquérito conta *fake news* (através dos termos acao stf, inquerito e combate). Analisando a comunidade #2, que possui mais perfis de usuários empenhados em promover ataques ao STF, que se auto identificam como conservadores é possível observar tópicos mais extremos como pedidos de destituição de ministros do STF, como destacado nos tópicos 1 e 6. Também é possível identificar termos associados a críticas ao inquérito das *Fake News*, como apresentado nos tópicos 1 e 7. Assim, é fácil perceber que essas comunidades se colocam com posicionamentos bastante antagônicos, representando usuários com visões distintas, muitas vezes

[9] https://www.bbc.com/portuguese/brasil-52827790

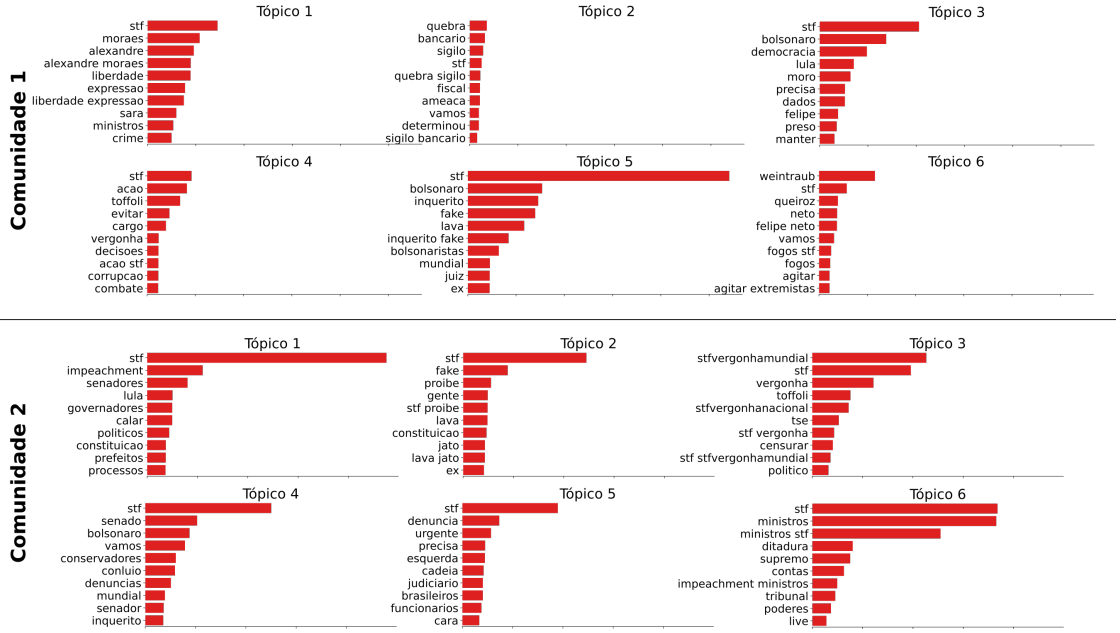


Figura 8: Termos dos três principais tópicos discutidos nas comunidades 1 e 2 considerando o assunto STF.

opostas, quanto à atuação do STF, em especial em relação à operação de investigação de propagação de notícias falsas. É possível observar que os tópicos globais estão presentes nestas comunidades, onde são reforçados.

Quando levamos em consideração os tópicos e termos observados para a rede de forma global (Figura 4), podemos perceber que a análise das comunidades isoladamente fornece uma visão complementar, que permite uma visão mais clara, não apenas da discussão que se dá na rede como um todo, mas dentro de cada grupo. A Tabela 2 mostra que, entre as principais comunidades da rede, a maior parte (comunidades 2, 3, 4 e 5) é formada por usuários de posicionamento declaradamente contrário ao STF, fazendo com que os principais tópicos discutidos de maneira global sejam quase completamente dominados por termos de ataque à instituição. Por outro lado, ao observar a Figura 8, fica mais fácil perceber a distinção entre o discurso promovido por usuários com alinhamentos ideológicos distintos.

Os tópicos destacados na Figura 9 refletem a polarização também 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`[10], 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 termos

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

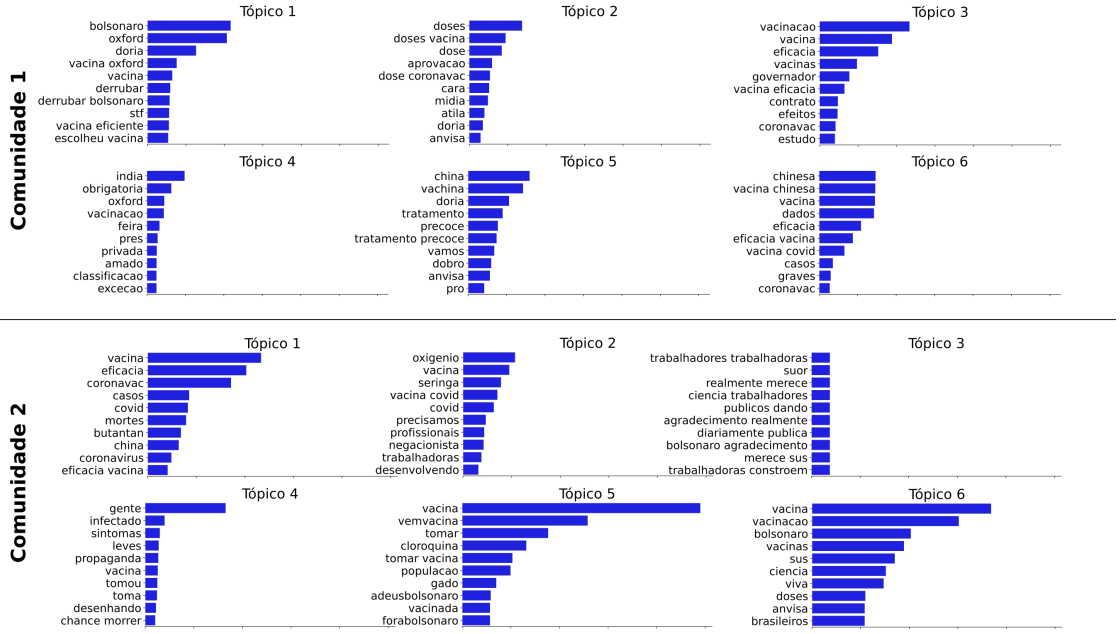


Figura 9: Termos dos três principais tópicos discutidos nas comunidades 1 e 2 considerando o assunto vacinação.

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)[11]. Analisando a comunidade 2, podemos identificar termos relacionados a um pedido de impeachment do então 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 da análise global da estrutura topológica da rede, além dos principais tópicos e termos – apresentados na Seção 4.1 – permitem que se tenha uma visão bastante clara dos assuntos discutidos sob diferentes perspectivas. Porém, quando consideramos a análise dentro de cada uma das comunidades, investigando os principais atores envolvidos nas discussões, além dos tópicos e termos especificamente abordados por grupos com alinhamentos ideológicos distintos, é possível perceber que há uma clara heterogeneidade na forma como os discursos são sustentados pelos usuários, como é apresentado nesta seção. Dessa forma, é possível avançar na resposta à QP1 (É possível observar uma distinção do conteúdo discutido dentro de cada uma das comunidades em relação à discussão no conjunto completo de usuários?). Buscando ainda avançar nas outras Questões de Pesquisa levantadas 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

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

na Seção 4.3.

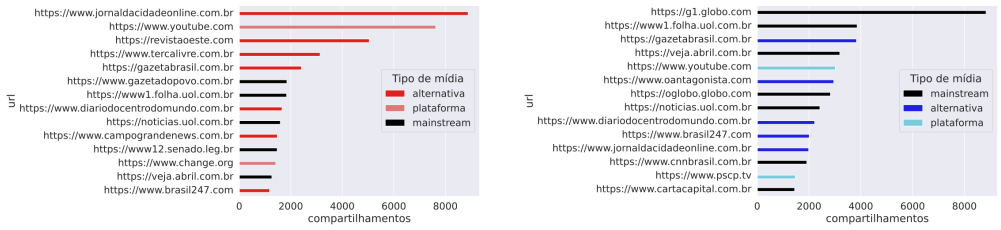
4.3. Análise das fontes das informações

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 e a um viés político unidirecional.

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

4.3.1. Fontes de informação sob uma perspectiva global

As Figuras 10a e 10b apresentam o resultado em função das bases de dados completa, sem a separação por comunidades. É possível identificar a presença de URLs que são fonte de informação externa ao Twitter, com a presença de plataformas como YouTube, mídias *mainstream* e mídias alternativas.



(a) STF. (b) Vacinação.

Figura 10: Frequência de URLs para os assuntos investigados.

Entre as plataformas compartilhadas nos diferentes assuntos analisados, como apresentado nas Figuras 10a e 10b, o YouTube possui uma importância significativa, estando presente nos dois assuntos explorados. Podemos identificar que o assunto "vacinação" tem como principais URLs, uma grande quantidade de veículos classificados como *mainstream*, quando comparado com "STF"(Figura 10a).

O assunto "STF" (Figura 10a), seguido do assunto "vacinação" (Figura 10b) concentra veículos classificados como mídia alternativa, de espectros políticos distin-

tos, como `jornalcidadeonline` e `diariodocentrodomundo`. A partir de uma inspeção manual foi identificado que alguns sites classificados como mídia alternativa apresentam conteúdo hiper-partidário, apresentando as informações de maneira desbalanceada, apresentando somente um ponto de vista e com uma grande quantidade de conteúdo publicitário em detrimento a notícias, sendo um forte indício que estes sites possuem como foco uma monetização com publicidade onde as notícias acabam ficando em segundo plano. Outro padrão identificado foi a reescrita de notícias originadas de veículos tradicionais com uma linguagem mais agressiva e tendenciosa podendo ser eficaz para atrair leitores mais polarizados. Em alguns destes veículos também estão ausentes informações sobre os editores responsáveis e os princípios editorais.

4.3.2. Fontes de informação sob uma perspectiva de comunidades

Aprofundando as análises das fontes de informação nas comunidades mais modulares, as Figuras 11 e 12 apresentam as frequências das 15 URLs dos *websites* mais compartilhados considerando os *tweets* estudados, assim como a classificação do tipo de mídia para o qual apontam para os assuntos `STF` e `vacinação`, respectivamente. 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.

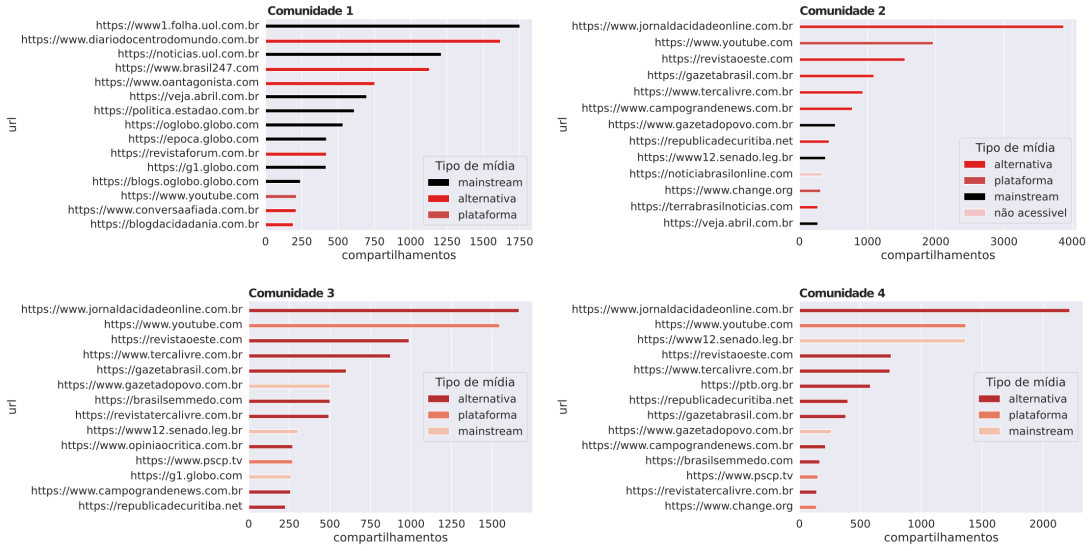


Figura 11: Frequência de ocorrência das URLs mais compartilhadas nas comunidades 1, 2, 3 e 4 para o assunto `STF`.

A Figura 11 permite observar que, para o assunto `STF`, a comunidade 1 utiliza como fonte, prioritariamente, veículos de mídia do tipo *mainstream*, enquanto as comunidades 2, 3 e 4 priorizam o uso de veículos de mídia do tipo alternativo (presentes, respectivamente, em 54%, 64% e 64% das URLs mais utilizadas). Assim, quando combinamos a análise das URLs mais frequentes à observação dos usuários mais centrais das principais

comunidades (Tabela 2), podemos notar que os indivíduos que frequentemente se declaram em posição de oposição ao STF, apesar de terem sido divididos entre as comunidades 2, 3 e 4 apresentam um padrão de comportamento discursivo muito alinhado, quando leva-se em conta as fontes de informação. Os veículos `jornaldacidadeonline` e `youtube` são os mais utilizados por todas essas comunidades e, além disso, outros veículos como `tercalivre`, `gazetabrasil` e `gazetadopovo` estão entre os mais utilizados, indicando uma forte convergência argumentativa entre os usuários. Considerando que a observação de tal padrão de alinhamento só foi possível através da análise das URLs utilizadas, é possível avançar na resposta à QP2 ("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*?"), já que, de fato, a análise das fontes de informação permite uma compreensão mais aprofundada dos usuários do *Twitter* envolvidos na discussão. Um padrão distinto pode ser observado na comunidade 1, que engloba usuários declaradamente contra o governo Bolsonaro, que prioriza o uso de veículos de mídia *mainstream* como fonte de informação (presentes em 53% das URLs mais utilizadas).

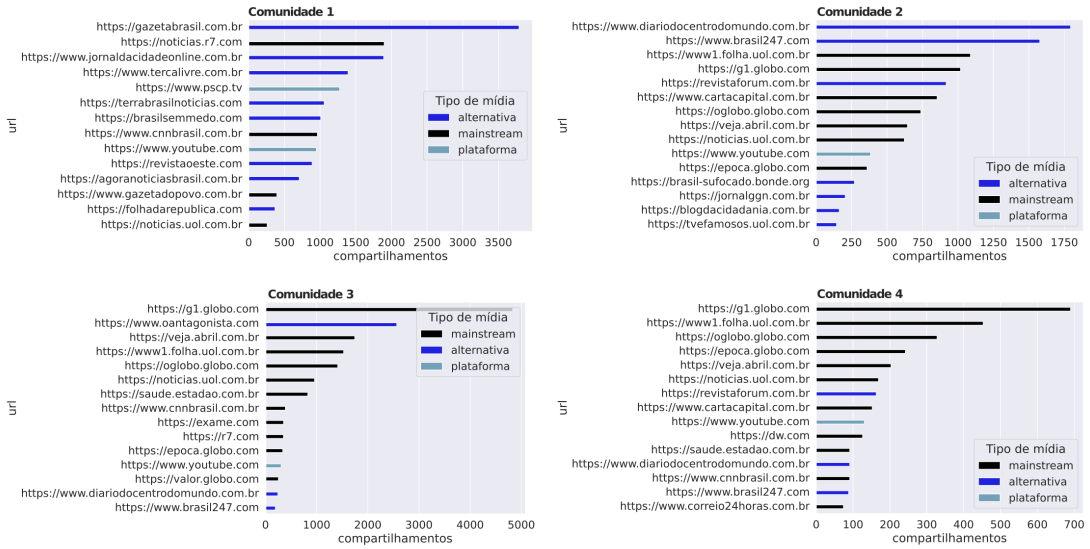


Figura 12: Frequência de ocorrência das URLs mais compartilhadas nas comunidades 1, 2, 3 e 4 para o assunto `vacinação`.

A partir da Figura 12, é possível perceber que, para o assunto `vacinação`, mídias consideradas *mainstream* são predominantemente utilizadas nas comunidades 3 e 4, representando aproximadamente 73% das fontes apresentadas. Considerando os usuários observados na Tabela 3, é possível aprofundar o entendimento sobre a propagação de conteúdo ocorrida no *Twitter* a respeito do assunto Vacinação. 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 12 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.

Dessa maneira, é possível afirmar que os resultados das Figuras 11 e 12 permitem um aprofundamento significativo da compreensão das comunidades descritas nas Tabelas 2 e 3 que, por isso, ajudam a responder à QP2 (“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*?”).

Comparações das URLs mais compartilhadas entre dois pares de comunidades para cada um dos assuntos STF e Vacinação são apresentadas pelas Figuras 13 e 14. 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.

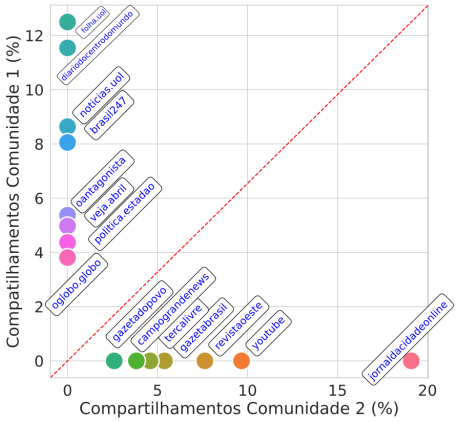
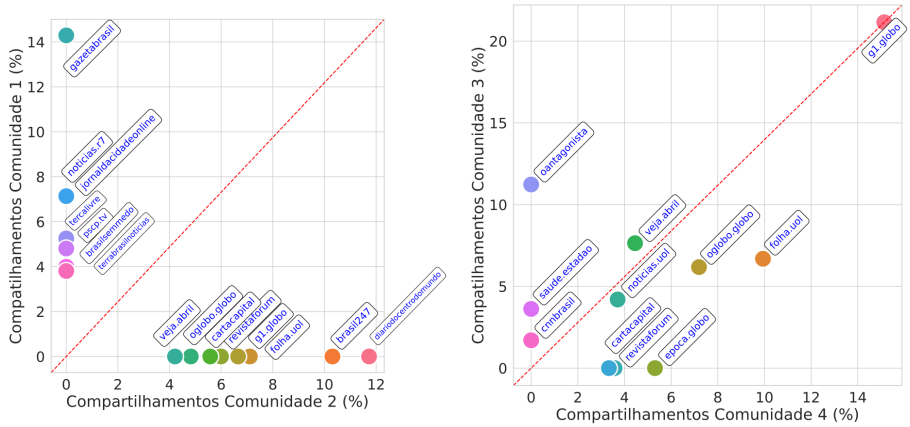


Figura 13: Comparação de compartilhamento URLs no assunto STF: comunidades 1 × 2.



(a) Comunidades 1 × 2.

(b) Comunidades 3 × 4.

Figura 14: Comparação de compartilhamento URLs no assunto Vacinação.

Considerando o assunto `STF`, a comparação entre as comunidades 1 e 2 (Figura 13) permite observar que, entre as fontes mais usadas, não há nenhuma que seja usada em ambas comunidades, mais uma vez enfatizando a distinção da forma, não só como os usuários interagem em cada um desses grupos, mas na forma como as discussões são por eles conduzidas. Assim, os resultados apresentados pela Figura 13 permitem que se avance nas respostas às QP2 e QP3, 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, caracterizando de maneira clara a polarização do debate *online*.

Observando a comparação entre as comunidades 1 e 2 para o assunto `vacinação` (painel esquerdo da Figura 14), 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 14). 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. Novamente, os resultados apresentados pela Figura 14 permitem que se avance nas respostas às QP2 e QP3, auxiliando a compreensão da forma como os usuários de comunidades distintas conduzem um assunto. Especialmente, o painel esquerdo da Figura 14 permite mais uma vez a caracterização da polarização do debate, não só no discurso, mas em elementos que sustentam esse debate.

4.4. 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. Neste trabalho, a classificação das URLs externas é baseada principalmente no resultado disponibilizado por [Guimarães et al. 2020].

A Figura 15 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.

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 16 apresenta uma amostra com 28 páginas que foram classificadas quanto ao seu viés e o 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. A análise da classificação de viés político (Figura 16) em conjunto com a frequência de URLs em cada uma das comunidades para os assuntos `STF` e `vacinação` (Figuras 11 e 12, respectivamente) nos ajuda a aprofundar a compreensão da forma como os indivíduos engajam nas diferentes discussões.

Para o assunto `STF`, podemos observar que os indivíduos da comunidade 1, embora tenham seus usuários mais centrais declaradamente opositores ao governo Bolsonaro (Tabela 2), utilizam como fonte veículos classificados com diferentes vieses, desde veículos de esquerda (como é o caso do `diariodocentrodomundo` e `brasil247`),

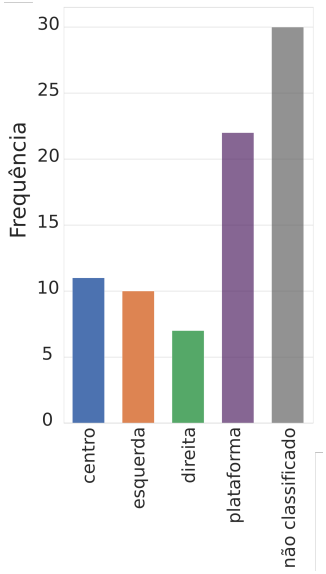


Figura 15: Distribuição da classificação das URLs quanto ao viés.

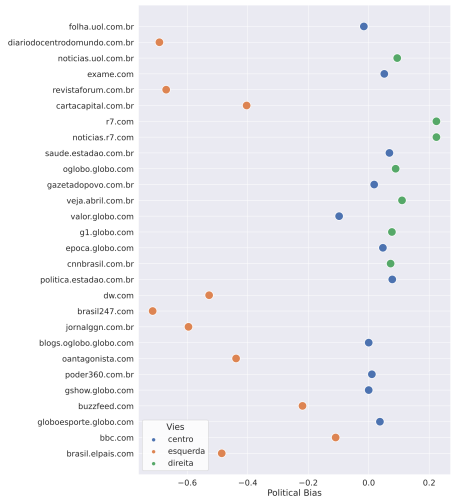


Figura 16: Classificação quanto o Viés (*Political Bias*) das URLs mais compartilhadas. (-1: esquerda, 0: centro e 1: direita).

passando por veículos de centro (como `folha.uol` e `politica.estadao`) até veículos de direita (como `noticias.uol` e `veja.abril`). Essa observação pode indicar que indivíduos da comunidade 1 abrem espaço para o debate do contraditório em seu discurso, mas também pode indicar que os indivíduos da comunidade 1 têm mais dificuldade em pautar o debate, reagindo a fontes de outros vieses ideológicos. . Por outro lado, nas comunidades 2, 3 e 4, que apresentam como usuários mais centrais indivíduos alinhados ao governo Bolsonaro, há uma maior tendência pela adoção de fontes do tipo plataforma e sem classificação quanto ao viés político, como é o caso do `jornaldacidadeonline` e da `revistaoeste`. A ausência de mídias do tipo *mainstream* como fonte nas comunidades 2, 3, 4 e 5 dificulta a análise do viés político adotado por seus indivíduos. Entretanto, o alinhamento editorial desses veículos permitem uma clara distinção em relação à natureza do debate conduzido pelos indivíduos na comunidade 1.

Já em relação ao assunto `vacinação`, 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, 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`).

Todos esses resultados nos ajudam a responder a QP3, 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 QP2, 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. Além disso, a metodologia apresentada permite uma análise comparativa da forma como os principais tópicos e termos relacionados a um assunto são discutidos na rede de forma global e nas comunidades de usuários isoladamente. Três Questões de Pesquisa (QP) foram levantadas: QP1) É possível observar uma distinção do conteúdo discutido dentro de cada uma das comunidades em relação à discussão no conjunto completo de usuários? QP2) 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*? QP3) É possível observar um comportamento distinto entre as comunidades em relação às URLs por elas compartilhadas? Tomando como base os dois estudos de caso com termos relacionados a discussões sobre o Supremo Tribunal Federal (STF) e sobre a vacinação da COVID-19 no Brasil, ocorridas no *Twitter*, as três questões levantadas puderam ser melhor compreendidas através da metodologia proposta. A análise comparativa dos experimentos realizados na rede completa e em cada uma das comunidades permitiu a observação da heterogeneidade do discurso em cada grupo de usuários. Assim, o assunto discutido na rede, dominado por um subconjunto de usuários, pôde ser melhor compreendido quando eram analisadas comunidades compostas por usuários com visões mais alinhadas sobre os assuntos. Dessa forma, foi possível responder, de maneira eficaz à QP1 levantada neste trabalho. Em relação à QP2, 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 tacitamente 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 QP3, 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 foram classificados por [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

[Barbosa et al. 2022] Barbosa, C., Félix, L., Alves, A., Xavier, C., and Vieira, V. (2022). Uso de urls para caracterização de comunidades em redes sociais online. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining*, pages 25–36, Porto Alegre, RS, Brasil. SBC.

[Bhatt et al. 2018] Bhatt, S., Joglekar, S., Bano, S., and Sastry, N. (2018). Illuminating an ecosystem of partisan websites. In *Companion Proceedings of the The Web Conference 2018*, pages 545–554.

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

[Blondel et al. 2008] 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.

[Caetano et al. 2018] Caetano, J. A., Almeida, J., and Marques-Neto, H. T. (2018). Characterizing politically engaged users' behavior during the 2016 us presidential campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 523–530. IEEE.

[Christhie et al. 2018] 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 et al. 2020] 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.

[Cruz 2020] Cruz, I. (2020). Quais os rumos do inquérito das fake news no supremo. Acessado em: 12 dez. de 2020.

[Ferreira et al. 2019] Ferreira, C., Murai, F., Matos, B., and Almeida, J. (2019). Modeling dynamic ideological behavior in political networks. pages 1–14.

[Garimella et al. 2018] 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 et al. 2020] 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.

[Morstatter et al. 2018] 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.

[Nobre et al. 2020] Nobre, G., Ferreira, C., and Almeida, J. (2020). Beyond groups: Uncovering dynamic communities on the whatsapp network of information dissemination.

[Rae 2021] Rae, M. (2021). Hyperpartisan news: Rethinking the media for populist politics. *New Media & Society*, 23(5):1117–1132.

[Recuero et al. 2020] Recuero, R., Soares, F. B., and Gruzd, A. (2020). Hyperpartisanship, disinformation and political conversations on twitter: The brazilian presidential election of 2018. In *Proceedings of the international AAAI conference on Web and social media*, volume 14, pages 569–578.

[Resende et al. 2018] 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.