The Shapes of the Fourth Estate During the Pandemic: Profiling COVID-19 News Consumption in Eight Countries

CAI YANG, Australian National University, Australia
LEXING XIE, Australian National University, Australia
SIQI WU*, University of Michigan, USA

News media is often referred to as the Fourth Estate, a recognition of its political power. New understandings of how media shape political beliefs and influence collective behaviors are urgently needed in an era when public opinion polls do not necessarily reflect election results and users influence each other in real-time under algorithm-mediated content personalization. In this work, we measure not only the average but also the distribution of audience political leanings for different media across different countries. The methodological components of these new measurements include a high-fidelity COVID-19 tweet dataset; high-precision user geolocation extraction; and user political leaning estimated from the within-country retweet networks involving local politicians. We focus on geolocated users from eight countries, profile user leaning distribution for each country, and analyze bridging users who have interactions across multiple countries. Except for France and Turkey, we observe consistent bi-modal user leaning distributions in the other six countries, and find that cross-country retweeting behaviors do not oscillate across the partisan divide. More importantly, this study contributes a new set of media bias estimates by averaging the leaning scores of users who share the URLs from media domains. Through two validations, we find that the new average audience leaning scores strongly correlate with existing media bias scores. Lastly, we profile the COVID-19 news consumption by examining the audience leaning distribution for top media in each country, and for selected media across all countries. Those analyses help answer questions such as: Does center media *Reuters* have a more balanced audience base than partisan media *CNN* and *Fox News* in the US? Does far-right media *Breitbart* attract any left-leaning readers in any countries? Does *CNN* reach a more balanced audience base in the US than in UK and Spain? In sum, our data-driven methods allow us to study media that are not often collected in editor-curated media bias reporting, especially in non-English-speaking countries. We hope that such cross-country research would inform media outlets of their effectiveness and audience bases in different countries, inform non-government and research organizations about the country-specific media audience profiles, and inform individuals to reflect on our day-to-day media diet.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: news consumption; media bias; cross-country analysis; Twitter; COVID-19

ACM Reference Format:
Cai Yang, Lexing Xie, and Siqi Wu. 2023. The Shapes of the Fourth Estate During the Pandemic: Profiling COVID-19 News Consumption in Eight Countries. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 317 (October 2023), 29 pages. https://doi.org/10.1145/3610108

*Corresponding author

Authors' addresses: Cai Yang, cai.yang@anu.edu.au, Australian National University, Canberra, ACT, Australia; Lexing Xie, lexing.xie@anu.edu.au, Australian National University, Canberra, ACT, Australia; Siqi Wu, siqiwu@umich.edu, University of Michigan, Ann Arbor, Michigan, USA.

2573-0142/2023/10-ART317 $15.00
https://doi.org/10.1145/3610108

1 INTRODUCTION

News media is an important political force, as it is sometimes referred to as the Fourth Estate[1] or fourth power, after the three estates – the clergy, the nobility, and the commoners – to reflect its power in advocacy and framing of political issues. Conventional knowledge about this fourth estate has been challenged in recent years, by unexpected election results [8], by alleged interference from foreign powers faraway [5], by the widespread use of proprietary, algorithm-mediated content personalization [19], and by the fact that social media users influencing each other in real time [4]. New understandings of how media reach and influence an inter-connected audience base around the world are very much called for, especially since much research has focused on the United States (e.g., its media, users, and political systems) [6, 9, 64] and a small number of Western countries [26, 47].

This paper aims to address three prominent gaps on measuring contemporary media consumption. First, recent work on quantifying media bias – via either editorial ratings [2, 48], surveys [9, 26], or data-driven measurements [6, 56] – all focuses on generating one single numerical score or ordinal label for each media, without explicitly taking into account the breadth of audience base or the diversity of published content. However, characterizing the distribution of audience on their political leaning or other attributes is important to reflect the diversity in both content providers and media consumers [40], hence the metaphorical "Shape of the Fourth Estate" in the paper title. The second aspect is in measuring media outlets beyond the US and most-studied Western countries [10, 45, 61], and designing automated methods and data-driven metrics that can generalize to a larger set of understudied media and countries. Quantitative profiles of media consumption in a diverse collection of countries can catalyze conceptual advances in characterizing media systems [20, 33], especially for societies outside the Western democratic system and culture. The third is the analysis of cross-country differences in audience profiles. Existing studies on media consumption either focus on in-depth analysis within a single country [6, 45, 56], or a number of countries by separately accounting for their respective political systems [37]. To the best of our knowledge, there has not been comprehensive political profiling of media audiences for the outlets across different countries. Overall, our work seeks to answer two questions:

RQ1: For a given country, in terms of political leaning, what is the breadth of readership for the media outlets there?

RQ2: For a given media, in terms of political leaning, how does its audience profile vary across different countries?

To answer those questions, we collect one of the largest COVID-19 tweet datasets from March to November 2020. We refer to this one billion tweet dataset as COVID2020 (Section 3). In the data processing pipeline, we first extract high-precision Twitter user locations from geotagged tweets and self-reported location strings (Section 4.1). We select eight countries where we can identify a sufficient number of politicians on the Twittersphere. We use the politicians' political position labels as seeds to estimate the user political leanings by applying a label spreading algorithm [66] over the user-user retweet networks (Section 4.2). We profile the overall user leaning distributions in these eight countries, and link the observations to existing theories in comparative media literature [33] (Section 5.1). Next, we profile the group of bridging users who have cross-country retweets (Section 5.2). We find consistent behavioral patterns – if users mostly retweet one political side in a country, it is unlikely that they would turn to mostly retweeting the opposite side in another country, indicating that the political leaning scores estimated from the within-country retweet networks are comparable with each other.

[1] https://en.wikipedia.org/wiki/Fourth_Estate

We further derive a new media bias estimate by taking the average of user leaning scores from those who have shared URLs from this media (Section 5.3). We find these new estimates are highly correlated with existing media bias ratings by conducting two validations: (a) against six US-centric media bias scores [2, 6, 9, 12, 56]; (b) against a recent survey study for 12 different countries [26]. Using URL sharing as a signal of media consumption, this work contributes a series of nuanced profiles of audience leaning distributions for a specific media in a specific country (Section 5.4). From a country-centric view, we show the country-level top consumed media and their audience leaning distributions. From a media-centric view, we show the different audience bases of one media across different countries. Those two views offer answers to many questions. Revisiting the questions we pose in the abstract, we find center media (e.g., *reuters*) tends to have a more balanced leaning distribution than partisan media (e.g., *cnn* and *foxnews*) in the US. As the partisan attitude of a media becomes stronger, or even extreme, fewer and fewer users from the opposite side would choose to consume its news (e.g., *breitbart*). Moreover, even if a media has an imbalanced audience base in one country (e.g., *cnn* in the US), it may still have a more balanced audience base in another country (e.g., *cnn* in Spain). All of those questions cannot be readily answered if one opts to disregard the audience leaning distribution in different countries.

Overall, our work presents a data-driven profiling of country-specific audience leaning distributions, which provides new insights to media outlets, activist groups, and media watchdog organizations, as well as to individuals for reflecting on and choosing our day-to-day media diet. We hope our methodology and observations serve as a basis for further inquiry, and they will shed light on understanding political behavior and media landscape in countries that do not often appear in computational social science writings, such as the Global South countries.

In sum, the main contributions of this work include:

- A new, large, and longitudinal tweet dataset COVID2020.[2]
- A new set of media bias estimates (surrogated by the average audience leaning score from users who have shared URLs from this media) that are shown to be highly correlated with existing editorial and survey-based media bias scoring.
- A new analysis of bridging users across countries, showing that cross-country retweeting behaviors do not oscillate across the partisan divide.
- A fine-grained profiling of audience political leaning distribution for a given media in a given country, which offers new insights to many comparative questions.

2 RELATED WORK

2.1 Measuring Media Biases

Having a reliable estimate of political media bias is important to scholars in political science, communication, and journalism. There have been three lines of approaches to measuring media bias. The first is via expert rating on media content. Several independent websites, such as AdFontes [1], AllSides [2], MBFC [48], all rely on a small editorial team to judge the content published by media and produce their own media bias reports. While the bias manifested in media content is seemingly a reasonable measure, the operation of expert rating makes the annotation process non-transparent, hard to scale, and subject to the biases from the human annotators. Second, there is a rich literature on automated identification of media bias, to name a few, text analysis based on the published news articles [22], user analysis based on the interactions over multiple communities [63], network flow analysis based on media's corporate contribution dollars to political parties [29], and content analysis based on the exposure of political actors [39] or biased news coverage [62]. For more discussion on this topic, we refer the reader to a recent survey [34]. The third line is using the

[2]Code and data are publicly available at https://github.com/computationalmedia/media_landscape

leaning of a media's audience as a surrogate of media bias. This is under the general observation of selective exposure [60]. Many studies have chosen this approach to estimate media bias [6, 9, 12, 56]. One challenge here is how to obtain user leaning. In previous research, Robertson et al. [56] linked Twitter users to US voter registration records, Bakshy et al. [6] used self-reported political affiliation from Facebook US users, Pew Research Center [12] surveyed partisan US internet users about their media habits. Once obtaining the user leaning, for one media, they usually computed the relative difference of Democrats and Republicans (or liberals and conservatives) who had interacted with this media as a metric of media slant.

In this work, we opt for the last approach. Differing from the aforementioned research [6, 12, 56] that leverages offline signals (e.g., voting records, demographics surveys), we use a computational method to infer user political leaning. We first identify a set of real-world politicians on Twitter, and use their political position labels as seeds to run a label spreading algorithm [66] on the within-country user-user interaction network. This method has proved to work remarkably well in both our experiments and prior studies [43, 59]. Furthermore, unlike most research [6, 9, 56] that only reports the average of audience leaning scores, we profile the distributions of audience leaning. This allows us to measure the variance of media audience. Even if two media have the same political bias rating, they might have a huge difference in whom reads news from them. Many examples are given in Figure 9. For instance, while *politico* and *cbsnews* are both rated center-left media, the audience of *politico* are mostly left-leaning, but the audience of *cbsnews* consist of a diverse set of left-leaning and right-leaning users. See the detailed analysis in Section 5.4.

2.2 Cross-Country Comparative Media Studies

The research of media consumption receives unequal attention geographically with many studies only focusing on the US [3, 6, 9, 12, 24, 64]. This is in part because US users have shown high degree of polarization in many online and offline behaviors [21], and in part because the two-party political system operating in US, which naturally fits into the left-right or liberal-conservative political spectrum. For a comprehensive picture of media consumption in multiple countries, we refer to the Reuters Institute's yearly digital news reports [51].

Cross-country comparative media research is important yet challenging. As Livingstone [46] highlights, "folk wisdom cautions against comparing apples and oranges" (p. 480). The problem is rooted in the gaps in data collected following different procedures, measures estimated over different populations, and concepts drifted across different countries. One canonical example is the contrasting attitudes toward abortion rights over the world. In the US, abortion attitudes are highly polarized by political ideology and abortion policies vary by state; while in many European countries, abortion is generally permitted. This example illustrates that abortion attitudes may be an effective signal to identify discordant groups in the US, but will become ineffective if one wants to apply it in European countries.

There is a rich line of comparative research; however, we argue that some are not really cross-country analyses because the within-country observations should be calibrated for cross-country comparisons. For example, Huszár et al. [37] audited the algorithmic amplification of political content on Twitter in seven countries. They measured the attitudes toward major political parties in each country – two parties in the US, four in the UK, five in Japan, six in Germany, and seven in France. While the measures in each individual country are valid, there is no warrant for the comparability across countries. In order to conduct rigorous comparative research, one must construct a common measurement axis [47]. For example, Fletcher et al. [27] used surveys to directly ask panelists from different countries to label their political leaning on a left-right spectrum. While a survey is an effective way to obtain self-reported subjective labels, it greatly limits the number of users one can study and generally has a much higher cost.

In this work, we scrutinize the cross-country measures before analysis. We identify groups of bridging users who have interactions in multiple countries. Except for France, we do not find any systematic shifts in the political leaning users estimated from the within-country retweet networks. This assures the comparability of the user political leaning estimates in different countries. See the detailed analysis in Section 5.2.

3 A NEW COVID2020 TWEET DATASET

We use the Twitter filtered streaming API to construct a new, high-volume COVID-19 tweet dataset. Our initial investigation in early 2020 shows that the daily volume of COVID-19 tweets exceeds Twitter API limit by a large margin. For example, about 25M COVID-19 tweets were posted each day in March 2020. This is significantly over the Twitter API limit of 1% of total tweets, which is roughly 4M for a day.[3] If we just use the Twitter API naively, more than 80% of relevant tweets would be missing. A better data collection strategy is needed to increase the data coverage and recall. We thus adopt a crawling strategy proposed by Wu et al. [65], which consists of three steps: (a) partitioning the whole data stream into several sub-streams by the tracked keywords and languages; (b) estimating the sampling rate for each sub-stream using the rate-limit messages provided by Twitter; (c) taking the union of all sub-streams. This strategy is shown to significantly reduce data loss, and thus can result in a high-fidelity dataset for high-volume, real-time tweet stream under Twitter API's limits [65].

We ran the data crawler from March 2020 to November 2020, covering eight months in the first year of the pandemic. The list of tracked COVID-19 keywords is obtained from Chen et al. [17] (see Appendix A). Those keywords include not only generic terms such as "corona virus", "covid", but also non-pharmaceutical interventions such as "lockdown", "n95", and "social distancing". Note that we do not periodically update the tracked keyword list to account for emerging COVID-19 topics such as "vaccine". However, vaccination had not been rolled out in the US until December 2020, and in most other countries, not until 2021.[4] We thus believe our keyword list reflects the public interest in COVID-19 during the data collection period.

Our sub-stream configurations are shown in Appendix A. The estimated sampling rates of different sub-streams range from 95% to 100%, meaning that we have a very high recall for all COVID-19 posts on the Twittersphere. To reduce the computational load, we processed one week's data in every two weeks. The resulting dataset contains 18 calendar weeks over an eight-month period (Mar-Nov, or week 13 to week 47 in 2020). We experienced several server glitches during data collection, and lost data for two entire weeks (week 27, 29) and another four days (in week 17, 31). In total, we obtained 999,040,035 COVID-19 tweets posted by 62,687,121 users. We refer to this dataset as COVID2020.[5]

The temporal data volume of the COVID2020 dataset is shown in Figure 1 (green line). The x-axis is the week number. The left y-axis shows the average number of daily tweets in the selected week. The volume of COVID2020 dataset starts from more than 25M tweets per day in March, and then drops to about 5M tweets per day in November. The decline may be attributed to user fatigue. According to a survey from Pew Research Center [15], 71% of Americans expressed willingness to take breaks from COVID-19 news. Another possible reason is the surfacing of new and uncollected topics. For instance, the discussions about COVID-19 shift from lockdown to vaccine over time.

[3]Researchers have found that the Twitter filter streaming API provides no more than 50 tweets per second [65], though the Twitter API limit is more commonly known as no more than 1% of the entire tweet volume for any time intervals.
[4]https://ourworldindata.org/covid-vaccinations
[5]To comply with Twitter's content redistribution policy, we only release the tweet IDs of the collected tweets. We cannot directly release our processed data because user-level information (e.g., extracted geolocation and estimated political leaning) is considered sensitive personal information by Twitter and by GDPR [28].

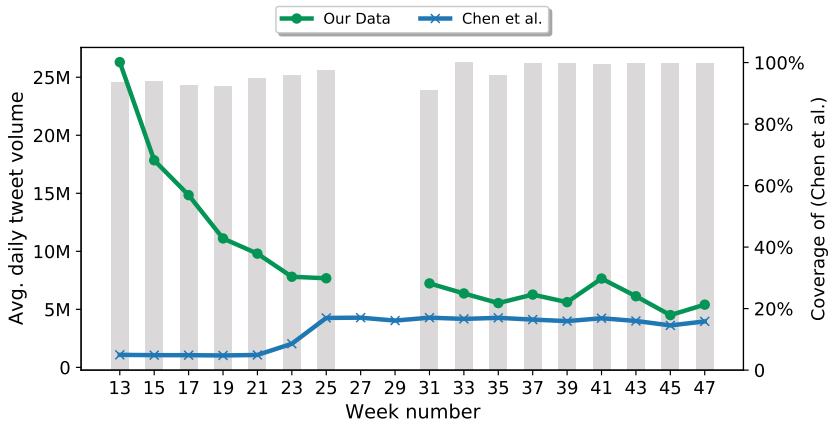


Fig. 1. Comparison between our COVID2020 dataset and Chen et al. [17]. Left y-axis: COVID2020 (green line) is much larger than Chen et al. [17] (blue line) throughout time. Right y-axis: most tweets (on average 96.6%) from Chen et al. [17] are also included in COVID2020 (gray bar).

Significance and bias of the COVID2020 dataset. On April 29, 2020, Twitter announced its own COVID-19 data stream endpoint and restricted access to those who applied before October 2020[6]. Unfortunately, despite that our data collection began in March 2020, we only became aware of this data source after the application deadline. Thus we could not provide a direct comparison with Twitter's COVID-19 data stream. From the documentation page archived by the Wayback Machine,[7] we find that Twitter's tracked COVID-19 keyword list is much larger than ours because Twitter includes many non-English terms from countries where people usually refer to "corona virus" in their corresponding native languages, for instance, China, Japan, Korea, India, Indonesia, and the Arab world. This indicates that our COVID2020 dataset is not a representative sample for public COVID-19 discourse and Twitter users in non-English-speaking countries and we should exclude those countries from our analysis. For countries where people still primarily use "corona virus", "covid", and/or their stems to refer to COVID-19, we expect the COVID2020 dataset to have a high coverage for COVID-19 tweets posted there. For instance, people say "le coronavirus", or "le virus corona", or "le (virus) Covid dix-neuf" in French. All of those keyword variants are captured by our data collector.

We also compare the COVID2020 dataset with a widely cited COVID-19 tweet dataset constructed by Chen et al. [17] (Figure 1 blue line). We find that COVID2020 is significantly larger throughout the data collection period, though the volume difference decreases as time passes. For example, in week 13 (03/23 - 03/29), Chen et al. [17] had only 7.6M tweets while we collected 184M tweets (24 times larger). Note that the observation of decreasing volume of COVID-19 tweets is only made possible by our data collection strategy with a high-sampling rate (green line). Reading the daily volumes from a single Twitter data stream would lead to the opposite and incorrect conclusion – i.e., rising COVID-19 tweet volume (blue line). On the right y-axis, we show the fraction of tweets from Chen et al. [17] that also appear in COVID2020 for every week (gray bar). On average, 96.6% of tweets from Chen et al. [17] are included in COVID2020. To the best of our knowledge, COVID2020 is the largest publicly available COVID-19 tweet dataset for the same period.

[6]https://twittercommunity.com/t/new-covid-19-stream-endpoint-available-in-twitter-developer-labs/135540/5
[7]https://web.archive.org/web/20221025172340/https://developer.twitter.com/en/docs/twitter-api/tweets/covid-19-stream/filtering-rules

4 EXTRACTING USER GEOLOCATION, POLITICAL LEANING, AND SHARED MEDIA DOMAINS

In this section, we extract attributes from tweets and Twitter users that are necessary for profiling COVID-19 news consumption in different countries, namely high-precision user geolocation (Section 4.1), user political leaning in the left-right spectrum (Section 4.2), shared URLs and media domains (Section 4.3).

4.1 Extracting Geolocation from Geotagged Tweets and User Profiles

On Twitter, we can extract the location information from two places: (a) geotag embedded in each tweet, which has high precision but is known to have low coverage and (b) location field in the user profile, which has higher coverage but can be very noisy.

- **Geotagged users.** 0.85% of all tweets in COVID2020 are attached with geotags, comparable to a prior study [52]. A geotagged tweet includes a (city, state, country) place tag. We use the country tag as the surrogate of user country. To aggregate geotagged tweets into geotagged users, we assign all the extracted locations to a user, which builds the user mobility traces during the COVID-19 pandemic. For simplicity, we filter out all users posting geotagged tweets from multiple countries – a reasonable criterion due to the significantly reduced global mobility in our data collection period. With this method, we identify 1,958,826 geotagged users.
- **Geoparsed users.** Because most users do not post geotagged tweets, we also parse the self-reported location strings from the user profiles. We use the *Simplemaps*[8] data to construct an exhaustive list of all possible combinations of cities, states, and countries (in both full names and abbreviations). We check the location strings against the curated list, and we only include users with one exact string-matching location. Restricting geoparsing to exact matches improves precision, and eliminates free-form imaginary locations such as *neverland* or *Mars* that are commonplace on Twitter. We identify 20,975,218 geoparsed users.

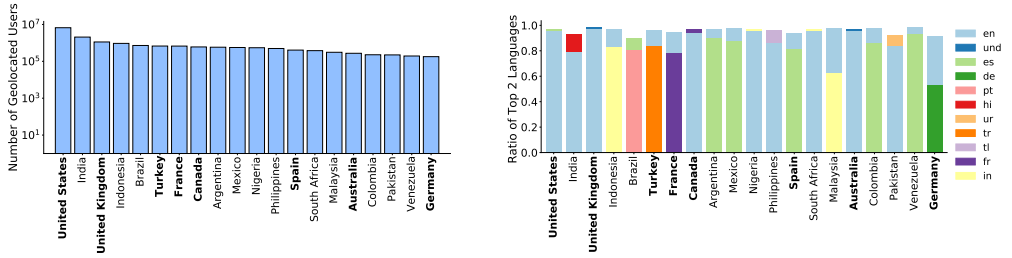
Evaluation. At the country level, there are 1,077,887 overlapping users from the two aforementioned methods. We consider the locations obtained by geotagged tweets as ground-truth and then evaluate the precision of parsing self-reported location strings on the overlapping users. We find that geoparsing correctly identifies the locations for 93% of overlapping users. The high precision shows that it is possible to take advantage of both methods to extract user locations on Twitter. We hence merge the results of two methods, with mismatched users assigned locations from geotagging. In total, there are 21,097,109 users with one unique geolocation, accounting for 33.6% users in COVID2020. They have posted 397,943,516 tweets, accounting for 39.8% tweets in COVID2020.

There are several third-party libraries for extracting geolocations from social media data [50], for example, OpenStreetMap and Google Geocoder API. We experimented with those libraries in our pilot test. However, we find that those libraries often operate based on a quota system and thus are unable to process millions of requests in a reasonable amount of time. We also evaluated several machine-learning-based geoparsing libraries (e.g., Mordecai) that trained on location texts, but found their performances not satisfying on the free-text Twitter profile location strings.

Geolocated users broken down by country. Figure 2a shows the top 20 countries with the most geolocated users. US has the most users (6,638,437) while Germany has the least (178,571). The country ranking of geolocated Twitter users is on par with previous work [42].

Top tweet languages in each country. We perform another sanity-check of geolocated tweets by examining the language distribution in the top 20 countries. Language information is extracted from the "`lang`" field in the tweet object. It can be one of the BCP 47 language codes or `und` if the

[8]https://simplemaps.com/data/world-cities



(a) (b)

Fig. 2. (a) Top 20 countries with the most geolocated Twitter users. (b) The fractions of the two most used languages in the top 20 countries. Bold texts mark the final set of eight selected countries.

language is not recognized by Twitter's language classifier. Figure 2b shows the stacked bar plot of the two most used languages in each country (to avoid the long tail of languages cluttering the legend). The y-axis shows the ratio of tweets written in the marked language. We find that a large fraction of tweets are posted using the native or official languages in non-English speaking countries. For instance, Spanish (`es`) is the most used language in Argentina, Mexico, Spain, Colombia and Venezuela; Portuguese (`pt`) is most used in Brazil; Turkish (`tr`) in Turkey; and French (`fr`) in France. This assures us that the COVID2020 dataset covers global Twitter users who are representative of their residence countries, and not obviously biased toward English-speaking users.

4.2 Estimating User Political Leaning from Within-Country User-User Retweet Networks

Ideally, there is a universal axis to measure the political polarization in different countries, but we find it an intrinsically difficult task due to the multifaceted nature of politics-diverse political systems. Commonly seen political facets include but are not limited to ideology (liberal vs. conservative), political leaning (left vs. right-leaning), party affiliation (Democratic vs. Republican), value (collectivism vs. individualism), and stance toward contested issues (pro vs. anti-abortion). After careful consideration, we find that some of those dimensions cannot be applied uniformly at the global level. Firstly, the religious and social values, and issue stances have huge variations and sometimes are even misaligned in ideological groups across countries [11]. For example, US liberals support abortion while US conservatives oppose it; however, abortion rights are generally supported in Canada and Europe regardless of political ideology [54]. If we use pro or anti-abortion stances to classify users, we would misclassify Canadian conservatives as liberals. Secondly, the political dividing issues vary significantly across countries. For example, in the US, contested issues include abortion, gun policy, racial justice, etc. [43]. But in most European countries, debates between political groups primarily focus on taxation, welfare, and immigration [10]. Issues in Asian and African countries are also less studied or documented.

For many countries, Wikipedia maintains a list of political parties and their political positions in a five-point Likert scale: left-wing, center-left, center, center-right, right-wing, which we collapse into three groups: left, center, right.[9] We thus decide to use the left or right-wing party labels provided by Wikipedia to estimate the user political leaning (left or right-leaning) from the within-country user-user retweet network involving local politicians. The term "within-country" means that a user's predicted leaning would only be affected by how s/he interacts with other users

[9]An example Wikipedia page of political parties in Canada: https://en.wikipedia.org/wiki/List_of_federal_political_parties_in_Canada#Current_parties

and politicians from the same country. Researchers have found highly segregated behaviors of politicians with opposing party affiliation [23, 30, 31], as well as deepening polarization in the mass public [25, 44]. One can expect what partisan information a user proactively chooses to share on social media indicates the alignment of the user's own political leaning with the shared partisan.

Note that the left and right-wing party labels are umbrella terms covering many social and economic issues. It is possible that two left-wing parties in different countries do not share a common belief over all issues [11]. Instead of dealing with nuanced differences directly, we rely on the crowd wisdom of Wikipedia contributors, who probably have more local political knowledge. It is also possible that assessments of left or right-wing party labels on Wikipedia do not follow the same criteria across countries. In other words, a left-wing party in one country may be perceived as right-wing by Wikipedia contributors in another country. To mitigate this concern, we empirically measure the cross-country retweet behavior on the set of bridging users in Section 5.2. As a preview, for users who retweet others in multiple countries, we do not find many people who mostly retweet one political side in one country, but turn to mostly retweeting the opposite side in another country.

Specifically, our procedure contains three steps (details in the following subsections). Firstly, we collect the list of political parties in each country. We search for the party members servicing in the year 2020 and their Twitter accounts. Next, we construct the user-user retweet network for each country. Finally, using local politicians as the seed nodes, we apply the label spreading algorithm [66] to estimate the user's political leaning in the left-right spectrum. We remark on three important points for this procedure. First, the use of label spreading algorithm is based on the assumption of network homophily. Prior research has shown that the retweet network is of high homophily [18, 45] and can be used to identify political groups on Twitter [7, 43, 45]. Second, the political leaning score that we estimate in this work does not quantify the extent of leaning left or right. It is possible that a user who disproportionately retweets one political side has only a mild view and a user who sporadically retweets both sides has a more extreme view. For the problem of detecting online extremism, we refer the reader to [32]. Third, the score neither represents probability because it is not yet calibrated. One can calibrate the predicted scores to real probability scores by obtaining a human-annotated subsample and then applying the Platt scaling technique [53]. The best way to interpret the estimated political leaning scores is that how confident we are in asserting a user is left or right-leaning by measuring their relative frequency of broadcasting information from the left and right-leaning groups.

4.2.1 Collecting politicians' Twitter accounts. Following [37], we use Wikidata's Query Service to identify politicians and their Twitter accounts. For a political party in a country, the Wikidata database lists its political position and party members, whose Twitter accounts are also indexed in Wikidata. For the top 20 countries with the most geolocated users, we collect the list of legislators and governors (or other equivalent roles). We require that they hold an active position in the year 2020. We also include all the election candidates from the 2020 United States presidential election.[10] Not all politicians actively advocate their views on Twitter and the usage of Twitter for political activism varies significantly by countries. We compute the coverage of politicians with identified Twitter accounts and filter out countries where the coverage is lower than 40%. This excludes 12 countries (India, Indonesia, Brazil, Argentina, Mexico, Nigeria, Philippines, South Africa, Malaysia, Colombia, Pakistan, and Venezuela) from our study. We are left with the eight countries shown in Table 1 (United States, United Kingdom, Canada, Australia, Spain, France, Germany, and Turkey).

[10]https://en.wikipedia.org/wiki/2020_Democratic_Party_presidential_candidates and https://en.wikipedia.org/wiki/2020_Republican_Party_presidential_primaries

	#politicians	#w/ Twitter accounts	%coverage	size of retweet network	%coverage of geo. users
United States	609	593	97.4%	946,436	14.3%
United Kingdom	1,456	636	43.7%	98,061	8.7%
Canada	454	192	42.3%	40,103	6.7%
Australia	235	176	74.9%	23,838	8.7%
Spain	635	530	83.5%	46,122	11.4%
France	943	550	58.3%	40,576	6.1%
Germany	752	574	76.3%	14,508	8.1%
Turkey	681	476	69.9%	27,809	4.1%

Table 1. The number of politicians, number and coverage of politicians with Twitter accounts, size of within-country user-user retweet network, and coverage of geolocated users (i.e., geolocated users in the retweet network divided by all geolocated users) in the eight selected countries.

4.2.2 Constructing the within-country user-user retweet network. A simple retweet (without comment) disseminates the original post to the retweeter's followers, showing support from retweeter to retweetee.[11] To this end, we construct the user-user retweet network for the geolocated users in each of the eight selected countries. The result is an undirected network. Each node is a geolocated user, and each edge indicates at least one retweet between the two users. We use the number of retweets between the two users as the edge weight. To remove insignificant edges in the network, we run the disparity filtering algorithm [58] to extract the network backbone. We set the significance threshold to be 0.05 as suggested [58]. We make a minor modification to the filtering process: if an edge is statistically insignificant but connects nodes that are both seed users, we do not discard it to ensure better connectivity to seed users (i.e., politicians in this study). This modification is shown to help the label spreading step described later. As shown in Table 1, US has the largest user-user retweet network and the highest coverage of geolocated users.

4.2.3 Propagating labels from politicians to general users. We apply the label spreading algorithm [66] on the extracted retweet network to infer the user's political leaning in each country. Local politicians are used as the seed nodes. An important hyperparameter is α, which controls the trade-off between preserving original information in the focal node and receiving new information from the neighboring nodes. We use 10-fold cross-validation to do a line search between 0 to 1 and find the optimal α in each country respectively. Upon convergence, label spreading returns a score between 0 and 1, which is then rescaled into $[-1, 1]$. We interpret the rescaled score as the predicted user political leaning. A score closer to -1 (1) means this user disproportionately retweets information from other left-leaning (right-leaning) users. A score closer to 0 means this user interacts with people from both partisans. We use 10-fold cross-validation to measure the prediction results on the seeded politicians. The accuracy is consistently high ($0.90 - 0.98$) over all countries, with the US having an accuracy of 0.96 and France having the lowest accuracy of 0.90.

Additional evaluation for the US users. One drawback of the above performance evaluation is that the sizes of seed nodes (#local politicians) are rather small, which may cause spuriously promising results. Because our COVID-19 data collection period is leading up to the 2020 US presidential election, we find that many tweets in COVID2020 also contain hashtags related to the US election. This enables us to adopt a hashtag-based method to predict the leaning of US users by measuring the stances toward election candidates from the two major US political parties.

Specifically, we collect political dividing hashtags from previous work [38, 64], for example, *#bidenharris2020*, *#voteblue* for left-leaning hashtags; *#maga*, *#trump2020* for right-leaning hashtags.

[11] We exclude retweets with comments (also known as "quote tweet" on Twitter) because the added comment may manifest disagreement or a negative attitude.

We obtain 314 left-leaning and 246 right-leaning hashtags. The full list is available in Appendix B. For each user, we calculate a score by $\frac{R-L}{R+L}$, where R and L are the numbers of right- and left-leaning hashtags this user has posted. To reduce uncertainty, we only consider users with a score less than -0.9 (higher than 0.9) as left-leaning (right-leaning) users. In total, we identify 195,719 users (83,382 left, 112,337 right). We then repeat the label spreading step on the US retweet work by using the 195,719 users as seeds. By this, we are able to estimate the political leaning for 1,023,093 US users, $\sim 76K$ (8%) more than using politicians as seeds.

Comparing the two versions of prediction results (politician seeds vs. hashtag seeds), there are 600,146 common users and 93.4% of them have the same predicted political leaning. Although we still do not know about the ground-truth labels for those users,[12] the high agreement rate between two different manners demonstrates the effectiveness of our overall data processing pipeline. For consistency, we use the politician-seeded predictions in the follow-up analysis.

4.3 Extracting Shared Media Domains from the Embedded URLs

URL sharing behavior is commonly used as a proxy to study media consumption patterns online [6]. To understand how Twitter users in different countries consume COVID-19 news, we extract URLs embedded in the tweets posted by users whom we can estimate their geolocation and political leaning. We use the "`expanded_url`" field in the tweet object.

We find that a non-trivial number of URLs are shortened URLs. If the URLs are shortened by Twitter's own shortener (i.e., *t.co*), the "`expanded_url`" field will contain the full URL address. If the URLs are shortened by other shortening services, we cross check the shortened URLs with a public shortener list.[13] We find 1,061,510 unique shortened URLs from 426 distinct shorteners. Among them, 331 shorteners (314,604 URLs) are platform-specific, meaning that the shortened URLs always redirect to a specific domain (e.g., *cnb.cx* to *CNBC*, *wapo.st* to *Washington Post*). The remaining 95 shorteners are general URL shortening services (e.g., *bit.ly*, *tinyurl.com*) and produce a total of 746,906 shortened URLs. We programmatically send web requests to those shortened URLs and successfully resolve 532,171 (71.3%) full URL addresses. The overall success rate for resolving shortened URLs is 79.8%, with failures due to HTTP timeouts, shorteners no longer functioning, and/or deprecated URLs.

We further extract the domain information from the shared URLs using a package called `tldextract`.[14] There are two benefits of studying news consumption at the level of the media domain. First, if a URL has only been shared by a handful of users, aggregating it into the domain level would give us more users to analyze. Second, even from the same media group, news from different domains may target different audience bases. For example, one would expect that *cnn* and *cnnespanol* have very different audiences because the former focuses on English-speaking users while the latter focuses on Spanish-speaking users, though they both belong to the CNN media group. The finer data granularity is particularly desirable in cross-country analysis. In total, we obtain 4,999,798 unique URLs from 155,302 distinct domains posted by 903,699 Twitter users.

5 RESULTS

We start with showing the distribution of user political leaning in the eight countries (Section 5.1). Next, we present an in-depth analysis of the bridging users who retweet from multiple countries (Section 5.2). With the shared media domains from geolocated users with estimated political leaning, we derive a new media bias score by computing the average of audience leaning estimates, and

[12]The ground-truth political leaning has to be obtained by directly asking the users, e.g., via survey or interview, but not by any means of inference.

[13]https://github.com/PeterDaveHello/url-shorteners

[14]https://github.com/john-kurkowski/tldextract

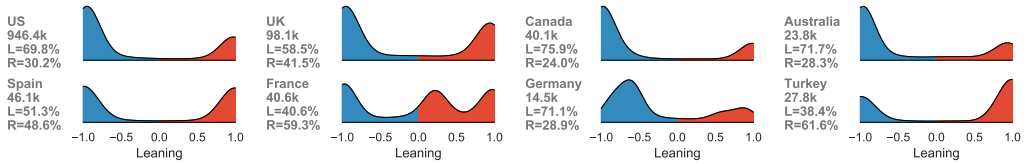


Fig. 3. The relative frequencies (density) of estimated user leaning in the eight countries. The numbers below a country name are the total number of geolocated users with estimated leaning scores, the fractions of left-leaning and right-leaning users found in the respective country from the COVID2020 dataset.

we show this score is correlated with existing media bias scores (Section 5.3). Lastly, we profile COVID-19 news consumption by generating the audience leaning distributions for a given country and for a given media domain (Section 5.4).

5.1 Distribution of User Political Leaning in Each Country

Figure 3 displays the density plots of the estimated user political leanings, one for each selected country. Users with scores in $[-1, 0)$ are considered leaning left within each country and colored in blue, while those in $(0, 1]$ are considered leaning right within each country and colored in red. Left-leaning users are the majority in most countries (ranging from 51.3% to 75.9%) except in France (%L=40.6%) and Turkey (%L=38.4%). For the US, the estimated fraction of left-leaning users is 69.8%, and this is in line with a COVID-19 era survey conducted by Pew Research Center finding that 69% of prolific US Twitter users are pro-Democratic [16]. Another overall observation is that the user distributions in seven out of eight countries have a bi-modal shape, meaning that only one mode (local maximum of a distribution) is in each of the left-leaning and right-leaning regimes. France is a notable exception with three modes in the user distribution.

Interpretation and limits of interpretations. Twitter is known to operate more like broadcast media than a social network [41]. We think that the literature on comparative media systems may help explain and conceptualize our observations. In the three models of Western media systems proposed by Hallin and Mancini [33], the US, UK and Canada are classified as the *liberal* model. We expect Australia to have a similar pattern due to the cultural resemblance and close alliance in the real world. The user distributions in these four countries indeed manifest high similarity (Figure 3 top row) – having a mode near the end of the left-leaning and right-leaning regimes. Yet there are also deviations from the comparative media system framework. For example, the media system in France is close to *polarized pluralist*, with media integrated into party politics and a strong role of the state.

The three-mode user distribution we observe for France concurs with the commonly-used populist/mixed/non-populist groupings [14] although there may be a formal discussion on the three groups that we are not aware of. User distributions in our study appear highly polarized in almost all countries without a prominent center/middle group except for France. The reason for this could be that social media inherently amplifies polarization [6], and that the original comparative media system framework pre-dates social media and only considers four dimensions: newspaper industry, political parallelism, media professionalized, and role of the state. This is confirmed by several recent literature, where Powers and Benson [55] find that online media increases external pluralism in media content in the US. Overall, research has shown the positive role of social media in increasing people's political participation [49], but in the meantime, social media attracts considerable attention to increasing polarization of political views [36], especially around COVID-19 issues and policies [35].

	United States	United Kingdom	Canada	Australia	Spain	France	Germany	Turkey
United States	940,154	2,797	3,924	2,086	755	1,076	1,306	370
United Kingdom	2,150	95,558	561	503	451	544	429	125
Canada	1,020	252	38,954	147	61	109	91	26
Australia	531	236	160	23,186	40	76	82	33
Spain	438	202	90	105	45,561	170	148	10
France	354	118	240	46	128	40,026	139	31
Germany	285	106	68	44	73	90	14,157	67
Turkey	173	68	44	55	0	115	154	27,504

Table 2. The number of geolocated users from the {row} country who have retweeted at least once geolocated users from the {column} country. The off-diagonal entries are the sets of bridging users between the {row} country and {column} country.

Note that the user leanings are estimated on whom are connected to local politicians via retweets about COVID-19 topics. While we do not claim that the studied user population is representative of each respective country or even its Twitter user base, in Section 5.3.2, we show that audience leaning scores estimated from this COVID2020 dataset have significant correlations with the media biases obtained via traditional survey methods in US, UK, Australia, France, Germany, and Spain [26]. We believe the geolocated users in COVID2020 dataset provide a novel and geographically diverse picture of political polarization around the world.

5.2 An Analysis of the Bridging Users

One key challenge in comparative research is the systematic variation of data collected from different places and/or different periods. Putting it into the context of cross-country political polarization, it is possible that left-leaning users in one country are more *right* than right-leaning users in another country. This misalignment would invalidate observations at face value. To make fair comparisons, a common axis is needed – a process we call "calibration". Traditionally, researchers rely on a set of users, whom they can collect answers along two dimensions and learn a mapping function between these two dimensions. They can then apply the mapping function to calibrate answers from one dimension to the other dimension or vice versa. For instance, Lo et al. [47] surveyed respondents about perceptual placements of multiple parties in Europe and calibrated the results to construct a common left-right scale to correct party positions across countries. Recall that the political leaning scores we estimate in Section 4.2 quantify how frequently a user retweets information from the left and right-leaning groups. The scores do not measure the degree of leaning left or right. We therefore cannot answer questions such as *are US left-leaning users more right than UK right-leaning users*. Instead, we should ask *do US left-leaning users retweet more from UK right-leaning users than from UK left-leaning users*. If the answer is yes, we should design a calibration procedure to align US left-leaning users with UK right-leaning users.

We identify a set of users who have interactions between multiple countries. We define *bridging users* as the geolocated users from one country and having retweeted geolocated users from another country. Table 2 summarizes the numbers of bridging users for every country pair. First, the diagonal entries show that most users' retweeting behaviors are limited to other users in the same country. Second, we focus on the bridging users between US and non-US countries due to the low number of bridging users between any pair of two non-US countries in COVID2020. For instance, we find 2,797 US users retweeting UK users and 2,150 UK users retweeting US users; hence a total of 4947 bridging users between US and UK. We generate two versions of political leaning scores for the bridging users. Specifically, for bridging users between the US and a country C, we first add them to

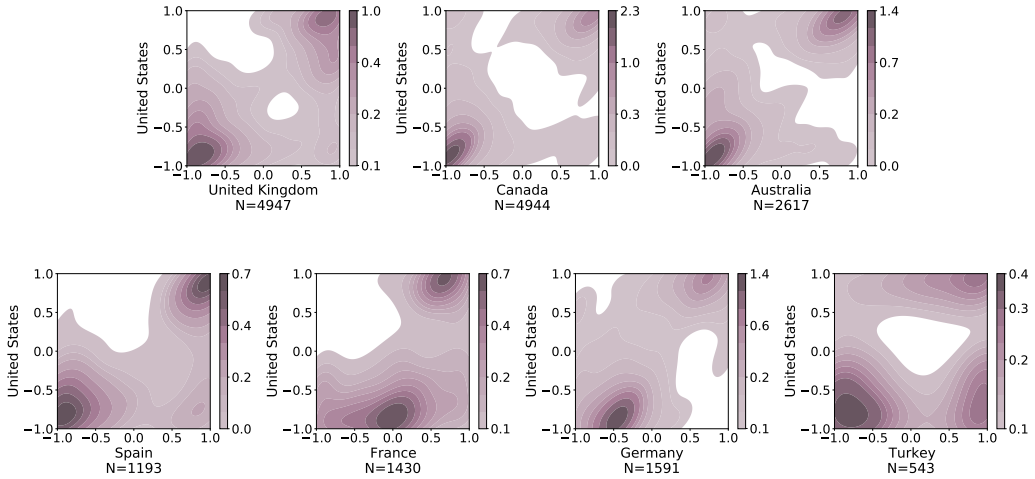


Fig. 4. 2D density plot of two versions of predicted political leaning scores for bridging users. x-axis: estimated by using the user-user retweet network in {x_label} country; y-axis: by using the retweet network in the US. N indicates the size of bridging users in each country pair. Except for France and Turkey, all other countries have the highest densities in the bottom left and top right regimes.

the within-country retweet networks from the US and from C. We then repeat the label spreading step to estimate their respective leaning scores by using the networks from the US and from C. We hence obtain two predictions for each bridging user.

Why do we NOT calibrate the predicted scores across countries? Figure 4 shows the 2D density plot of the two versions of predicted scores in each US vs. non-US country pair. We find that densities are higher in the bottom left and top right area in almost all pairs, indicating consistent retweeting behaviors across countries – if a group of users mostly retweet one political side in a country, they would not switch to mostly retweeting the opposite political side in another country. This observation warrants that no special calibration is needed to adjust the political leaning scores estimated from different within-country retweet networks. Even if there exists a latent common axis that we can project all predicted scores onto, we are unlikely to overturn the estimated left and right-leaning labels. Two exceptions in Figure 4 are France and Turkey. For the France-US pair, many bridging users interact with a politically diverse set of French users while mostly interacting with left-leaning US users. For the Turkey-US pair, although the highest density still occurs in the bottom left area, the bottom right area also has a moderate density, suggesting a modest number of users who mainly interact with right-leaning Turkish users and left-leaning US users. This may relate to two prominent groups of right-leaning French users and a large number of right-leaning Turkish users as presented in Figure 3. In order to understand the nuances there, one has to investigate the tweets they posted and the users they retweeted. We leave this case study as future work.

We also experimented with fitting a linear regression function for each country pair similar to Lo et al. [47]. We did not report the calibrated results for two reasons: (1) the 2D density plot suggests the likelihood of a non-linear relation; hence the linear fitting is not ideal; (2) the density in the middle area is low; hence we would have a high level of uncertainty (i.e., large confidence interval) for users around that area.

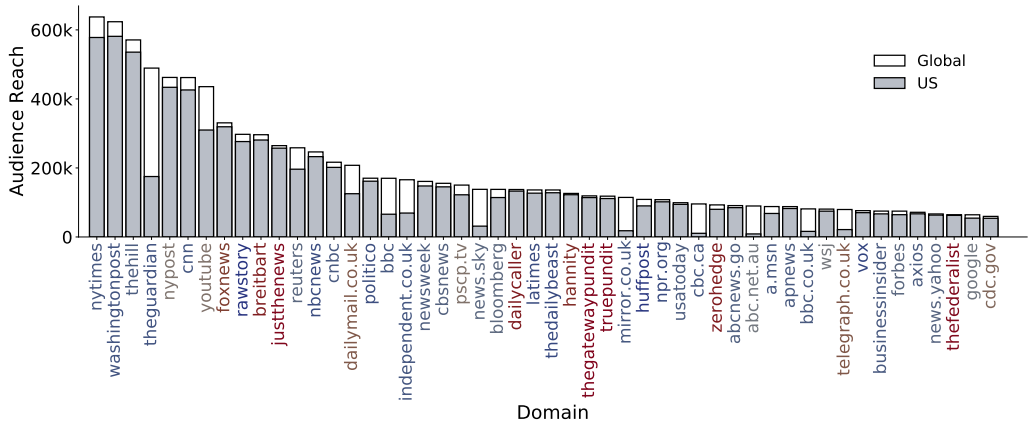


Fig. 5. Top 50 domains ranked by their global audience reach. The shaded area indicates the size of the US audience. The domain names are colored by their average audience leaning scores in the US. Blue (red) color indicates more left-leaning (right-leaning) audience.

5.3 A New Media Bias Score by Average Audience Leaning

5.3.1 Computing audience leaning distribution for each domain. In this work, each Twitter user can be represented as a tuple (u, l_u, p_u, D_u), where u is the user id, l_u is the extracted user location (Section 4.1), p_u is the estimated political leaning score (Section 4.2), and $D_u = \{d_{u1}, d_{u2} \ldots\}$ is the set of media domains that user u has shared (Section 4.3).

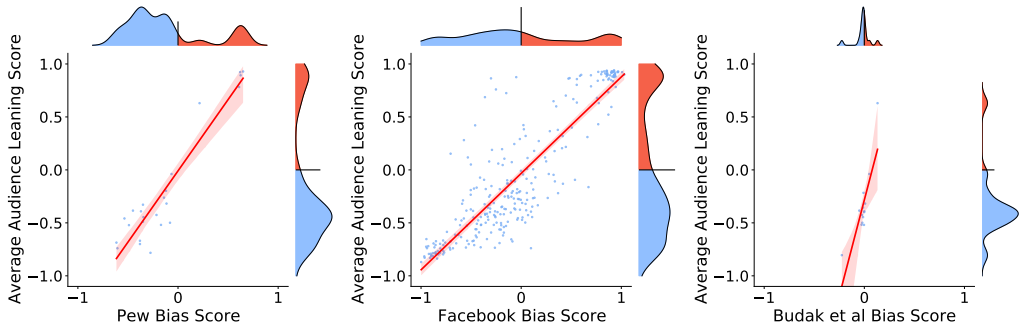
We define the global *audience reach* $\kappa(d)$ of a media domain d as the number of all unique Twitter users sharing URLs from this domain $\kappa(d) = |\{u|d \in D_u\}|$. Similarly, the audience reach of domain d in location l is $\kappa(d, l) = |\{u|d \in D_u, l_u = l\}|$, where $|\cdot|$ denotes set cardinality. Figure 5 displays the top 50 domains according to their global audience reach.[15] We observe not only mainstream media (e.g., *cnn*, *foxnews*), but also alternative, extreme media (e.g., *breitbart*, *thegatewaypundit*). The shaded area indicates the size of the US audience. We find many non-US-based media (e.g., *theguardian* from the UK, *cbc.ca* from Canada). The top domains span the full range of the left-right political spectrum (e.g., from left-leaning media *cnn* and *bbc*, to right-leaning media *breitbart* and *independent.co.uk*). The domain names are colored by their average audience leaning scores (detail follows), where blue (red) indicates a more homogeneous left-leaning (right-leaning) audience base.

We compute the *average audience leaning* $\bar{p}(d)$ of each domain d by averaging over all users who have shared d at least once. Similarly, we can compute the average leaning of domain d shared by users in location l, denoted as $\bar{p}(d, l)$. In the subsequent use of the average audience leaning score, we omit the notions of domains d and location l when they are clear from context.

$$\bar{p}(d) = \frac{1}{\kappa(d)} \sum_{u|d \in D_u} p_u; \quad \bar{p}(d, l) = \frac{1}{\kappa(d, l)} \sum_{u|d \in D_u, l_u = l} p_u$$

User-based aggregation was commonly used in recent work, but only to compute the mean. Bakshy et al. [6] and Robertson et al. [56] both described a domain by averaging over the leaning scores of users who had shared it. Information extracted from the COVID2020 dataset offers a rich

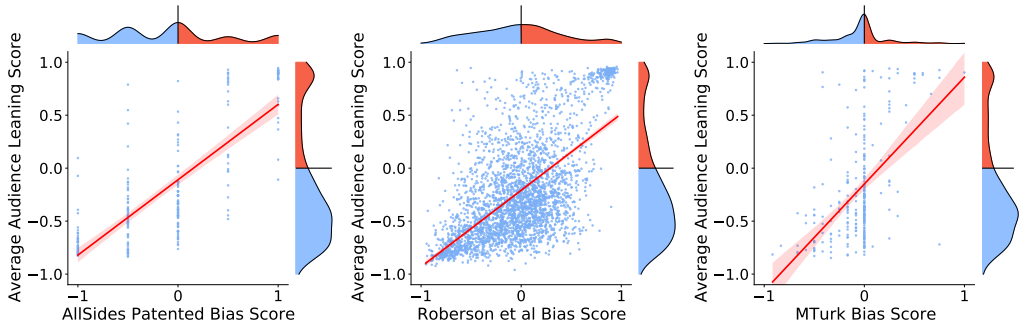
[15]Despite using the word "media", we notice that some platforms also appear in top positions (e.g., *youtube*, *pscp.tv*, *google*). They are content hosting services rather than producing news content by themselves. We opt to keep those platform domains since many media outlets also function as a platform for user-generated content (e.g., *New York Times Opinion*), and drawing a precise line is difficult.



(a) Pew Research Center [12], n=24 (86%), $r = 0.94^{***}$

(b) Bakshy et al. [6], n=281 (65%), $r = 0.88^{***}$

(c) Budak et al. [9], n=14 (93%), $r = 0.83^{***}$



(d) AllSides [2], n=237 (81%), $\rho = 0.75^{***}$

(e) Robertson et al. [56], n=3132 (16%), $r = 0.63^{***}$

(f) Robertson et al. [56], n=262 (68%), $\rho = 0.46^{***}$

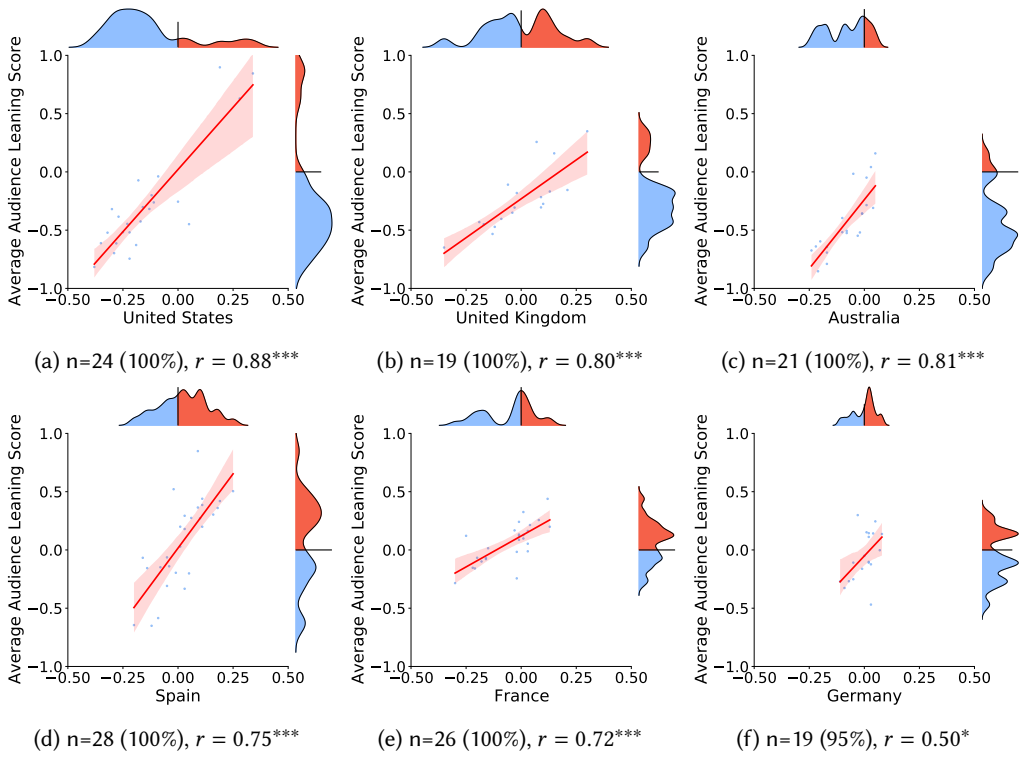
Fig. 6. Correlation between average audience leaning scores from our data (y-axis) vs. media bias scores from prior literature (x-axis). The subfigure title lists the reference, number of overlapped domains, coverage, and correlation coefficient (Pearson's r for numerical labels and Spearman's ρ for ordinal labels). $^{***}p < 0.001$. Side densities represent the distribution of domain leaning scores from either source.

set of statistics for each domain. The distribution of user leanings for domain d in location l can be represented as mean, median, and density plots (See plotting styles of Figure 3, 9 and 10).

5.3.2 Validating average audience leanings. As a validation, we compare the average audience leaning to two kinds of publicly available media bias reporting.

Comparing to existing media bias ratings. We compare the average audience leaning scores to six other estimates from recent literature [2, 6, 9, 12, 56]. Since these studies focus on the US users, we compare them to the average audience leaning for US users $\bar{p}(d, l = \text{'US'})$. To mitigate noise, we only include domains with a US audience reach of at least 50. This yields 7,924 media domains. We take the intersecting domains from our study and prior literature. Some work provides a numerical label (from -1 to 1), while some work provides an ordinal label (from extreme left to center to extreme right). We compute Pearson's correlation coefficient r for numerical labels and Spearman's rank correlation coefficient ρ for ordinal labels. The higher the coefficient is, the stronger the correlation between the two data sources. Figure 6 summarizes the results with scatter plots of the scores and correlation coefficients.

Pew Research's audience profile scores [12] were collected from 2,901 web respondents that were representative of the US Internet users in 2014. We reconstructed these scores from the interactive



(a) n=24 (100%), $r = 0.88^{***}$ (b) n=19 (100%), $r = 0.80^{***}$ (c) n=21 (100%), $r = 0.81^{***}$

(d) n=28 (100%), $r = 0.75^{***}$ (e) n=26 (100%), $r = 0.72^{***}$ (f) n=19 (95%), $r = 0.50^{*}$

Fig. 7. Correlation between our average audience leaning score (y-axis) vs. those from Fletcher et al. [26] (x-axis). The subfigure title lists the number of overlapped domains, coverage, and Pearson's r correlation test. $^{***}p < 0.001$,* $p < 0.05$. For domains shared in multiple countries, we calculate the average leaning scores from audience in each country independently.

webpage [13]. The remaining five sets of media bias scores are obtained from [2, 6, 9, 56]. Figures 6a to 6d show high correlations between $\bar{p}$ and survey-based [12], sharing-based [6], crowdsourcing-based [9], and expert-based [2] media bias scores. Notably, the correlation of $\bar{p}$ with the Pew scores and AllSides scores are much higher ($r = 0.94^{***}$ and $\rho = 0.75^{***}$ respectively) than the same comparisons conducted in [56] ($r = 0.78^{***}$ and $\rho = 0.64^{***}$ respectively). The lowest correlation is observed with the MTurk scoring ($\rho = 0.46^{***}$ in Figure 6f), but it is consistent with Robertson et al. [56]'s own observation ($r = 0.50^{***}$).

Comparing to international media surveys. A key contribution of our study is the cross-country analysis. To this end, we compare $\bar{p}$ with survey results conducted by Fletcher et al. [26], in which respondents are a stratified sample from 12 different countries, and were asked about their political leaning on a seven-point Likert scale (later coded into a scale from −0.5 to 0.5). Respondents were also asked about the news outlets they had read online and offline in the past weeks, from a candidate list of 30 popular outlets that varied from country to country. Fletcher et al. [26]'s political leaning score for each news outlet is the mean of the self-identified leaning scores of its audience. We find a strong correlation between the media bias scores from our estimation and that from [26] in the US, UK and Australia ($r \geq 0.8^{***}$, Figures 7a to 7c). We also notice a moderate correlation in Spain and France ($r \geq 0.7^{***}$, Figures 7d and 7e). The correlation with Germany is

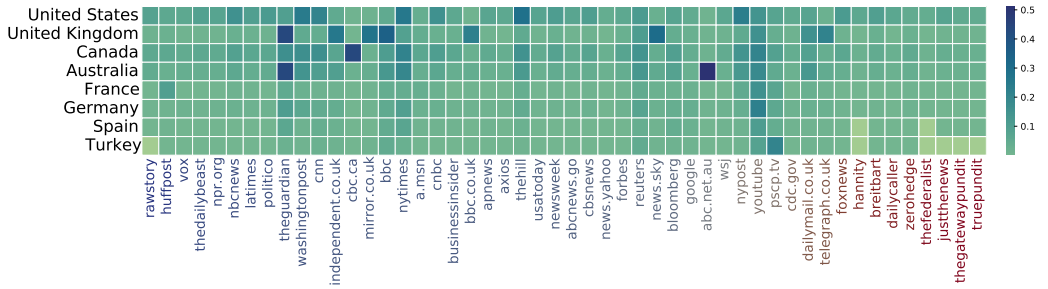


Fig. 8. Heatmap of within-country audience reach of the overall top 50 domains in the eight selected countries. x-axis: domain names ranked by their average audience leaning in the US. y-axis: country names. Cell color represents the fraction of audience reach domain d in country l ($\kappa(d, l)$), normalized by all users in that country.

lower ($r = 0.50^*$, Figure 7f). This is likely due to that survey outcomes for German media in [26] are mostly around the center.

Through two validation tasks, we find strong correlations between our computed average audience leaning score $\bar{p}$ and other estimates of US media domain biases and international surveys. This provides us great confidence in using the new scores to produce novel observations about the media consumption behavior internationally, though we caution that the reliability of $\bar{p}$ in countries other than those validated above may require further scrutiny.

5.4 Profiling Cross-Country COVID-19 News Consumption

Figure 8 provides an overview of the audience reach of the top 50 media domains (ordered by average audience leaning in the US) across the eight selected countries. These background statistics confirm the popularity of mainstream media in the four English-speaking countries, such as the *guardian, bbc, news.sky* for the UK, *cbc.ca* for Canada, and *guardian* and *abc.net.au* for Australia. Note that the top 50 domains are dominated by those in English, and that media consumption in the four non-English-speaking countries does not have a prominent peak – necessitating country-specific profiles that examine non-English domains that are important for each country. This motivates the next set of profile plots that visualize audience distributions in a way comparable across media or across countries.

A country-centric view. Figure 9 profiles the distribution of audience leaning scores for the top 15 media domains by within-country audience reach in the eight selected countries.[16] The media domains within each country are then reordered by their average leaning score $\bar{p}(d, l)$ (shown as a black vertical line). We also show media bias labels from AllSides [2] and MBFC [48] on the legend, wherever labels are available (L: left, CL: center-left, C: center, CR: center-right, R: right, ER: extreme right). Comparing the two editorially curated ratings, MBFC covers more news outlets, especially outside of the US. At a glance, the geolocated users in the COVID2020 dataset seem to heavily consume news media from the respective country. There is a modest level of news circulation among the four English-speaking countries (e.g., *theguardian, washingtonpost, nytimes* appear in the US, UK, Canada, and Australia). In the four non-English-speaking countries, the most

[16]Note that *bbc.co.uk* and *bbc* both appear in the UK profiles. The former is for browser requests sent from the UK, while the latter is for those sent from other countries. We treat them as two separate domains to be consistent with prior work such as Bakshy et al. [6], Budak et al. [9], Robertson et al. [56].

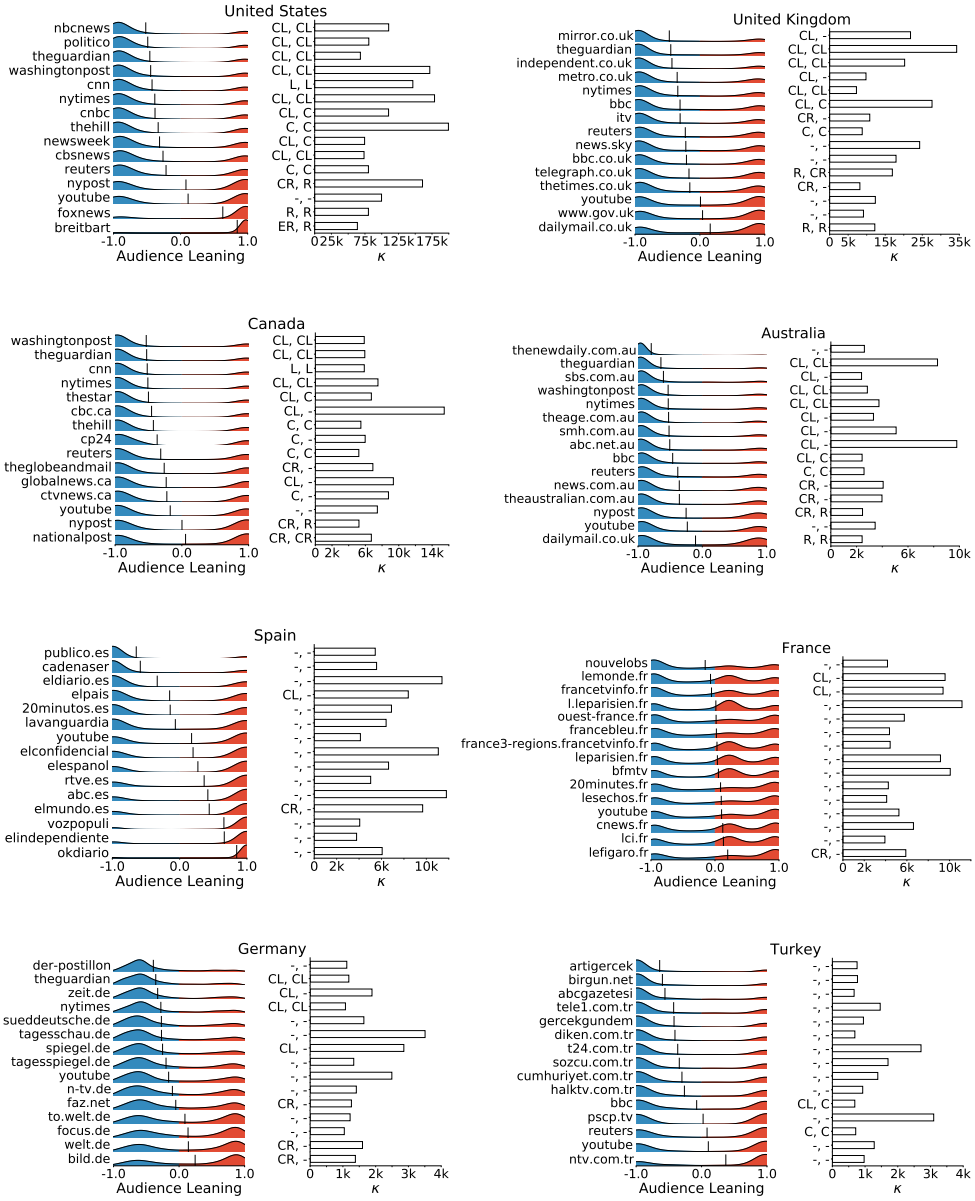


Fig. 9. Audience leaning distribution for the top 15 media domains ranked by audience reach in the eight selected countries. In each subfigure, left part: ridge plots of audience leaning distribution; black vertical line: mean. The domains are ranked by their average audience leaning scores $\bar{p}(d, l)$ (solid lines). Media at the top (bottom) have more left-leaning (right-leaning) audience. Right part: bar plots of audience reach $\kappa(d, l)$ of each domain in the respective country. x-axis: audience reach in thousands. y-axis: media bias ratings from MBFC [48] and AllSides [2], "-" means not rated by the corresponding source.

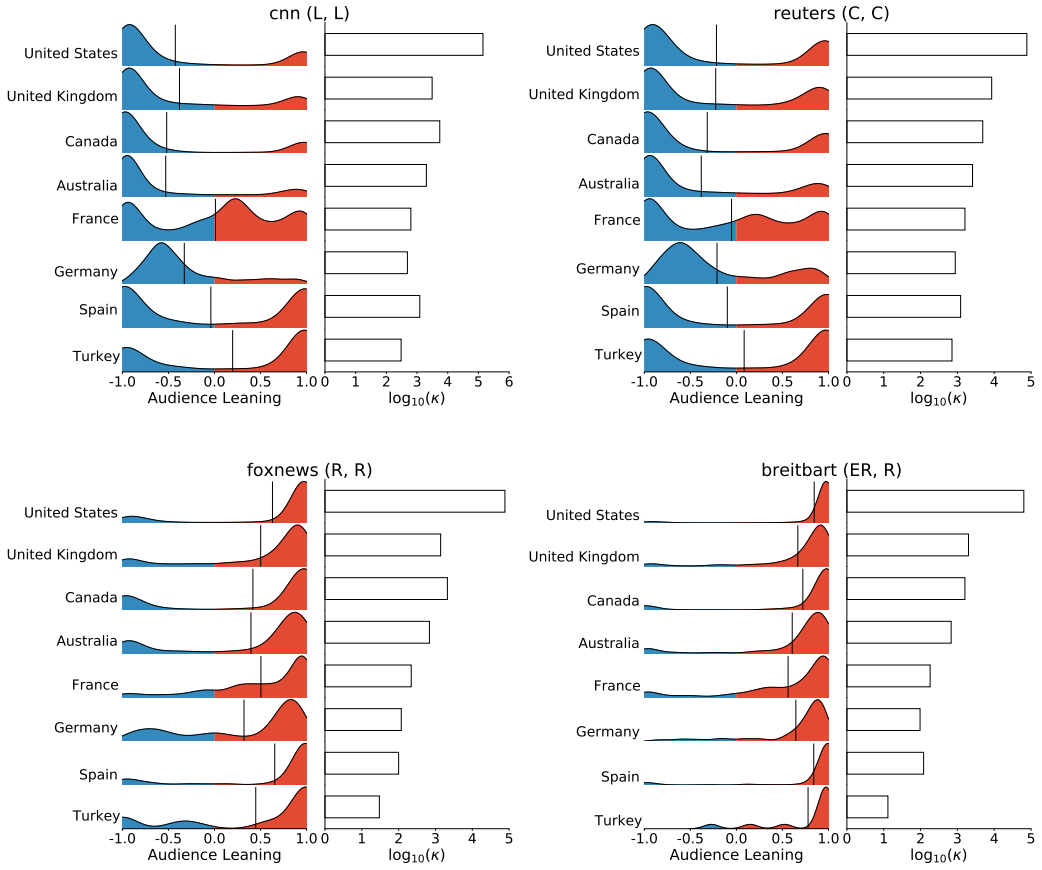


Fig. 10. Audience distributions across different countries for *cnn*, *reuters*, *foxnews*, and *breitbart*. In each subfigure, left part: ridge plots of audience leaning distribution in eight countries; black vertical line: mean. Right part: bar plots of audience reach in each country in log scale. x-axis: audience size in thousands. MBFC [48] and AllSides [2] labels are in brackets after each domain name.

popular media are strongly language-specific – almost all are from the respective country, except for *youtube* that appears in all countries.

From these distributions, we observe considerable variations in terms of both the mean and spread of audience leanings. This paints a unique picture of the media landscape. For example, firstly, 11 out of the top 15 domains in the US are either center or left-leaning if we only look at the labels from MBFC and AllSides. However, there exist large variances in their audience distribution – audiences of domains such as *nbcnews, politico, theguardian* consist of a large proportion of left-leaning users in the US, while domains like *newsweek, cbsnews, reuters* are shown to have a more balanced audience base. Secondly, as the media becomes more partisan, its audience base shrinks to a narrower range. For example, *reuters* in the UK consists of a balanced set of left-leaning and right-leaning users, but partisan media *mirror.co.uk* and *dailymail.co.uk* have few audience from the opposite political side. For far-right media such as *breitbart* in the US, *okdiario* in Spain, *ntv.com.tr* in Turkey, almost no left-leaning users consume it. All this information is not readily available without

profiling the distribution of audience leaning scores. Thirdly, the data-driven leaning estimates can enrich or contradict existing media bias labels For example, *cnn* is rated left-leaning by both sources, but the average audience leaning of *nbcnews* and *washingtonpost*, both rated center-left, are more left (smaller) than that of *cnn*. Fourth, in countries other than the US, the relative audience leaning for media can confirm or contradict default intuitions. In the UK, most media has audiences that are spread across the left-right spectrum, including *bbc* (a government-sponsored media) and *www.gov.uk* (the government website itself). In Canada, *theguardian* ranks more right than *washingtonpost*, but it's the opposite in the US. Lastly, this audience leaning distribution provides a new angle to estimate the biases for media not collected in sources like MBFC or AllSides, or from understudied non-English-speaking countries. For example, we can generate media bias estimates for 12 unrated domains in Spain (less youtube, same for the other countries), 11 in France, 7 in Germany, and 12 in Turkey. In Australia, all media outlets have an average audience leaning score less than 0 (lean left); this is certainly influenced by the background distribution of Australian Twitter users in COVID2020 (71% left). Such observationx indicate that the mean audience leaning score is more readily interpreted relative to each other rather than in an absolute sense, and that background statistics in the sample population are important.

A media-centric view. Figure 10 profiles the distribution of audience leaning scores for four example media across the eight countries. We choose *cnn*, *reuters*, *foxnews* and *breitbart*, which have been labeled as left, center, right, and extremely-right by MBFC [48]. For each media, the computed average audience leaning scores have a large variance across countries (i.e., audience bases). For instance, *cnn* is shown to have average audience leaning scores varying between -0.5 to 0.2. Despite being rated as left-leaning, the calculation over different countries reveals a more nuanced picture – it has an average audience leaning score close to zero in France and Spain and even a positive score in Turkey. It has a bimodal distribution in six countries, except for a unimodal distribution in Germany and a trimodal distribution in France. For center media *reuters*, its average audience leaning is close to 0 in France, Spain and Turkey, but most of its audience in Canada, Australia, and Germany are estimated left-leaning. Turkey shows an opposite observation, where about half of the audience are estimated to be right-leaning. In most countries, the distributions are bimodal except for France and Germany. For right-leaning media *foxnews*, the average leaning scores vary from 0.32 to 0.65 across the selected countries. Its audience has been shown to be mainly right-leaning users in the US, UK, France, and Spain (> 80%); while there is also a nontrivial number of audience occupying the left-leaning spectrum (> 25%) in Canada, Australia, Germany, and Turkey. The far-right media *breitbart* has average audience leaning scores ranging from 0.5 to 0.85 in the eight countries. Audience leaning distributions of *breitbart* are unimodal in all countries, suggesting that very few left-leaning users would consume *breitbart* news. Its audiences in the US, UK, Spain, and Turkey are over 90% right-leaning, whereas in Canada, Australia, France, and Germany, less than 90% but still more than 75% of its audiences are estimated right-leaning.

Our observations are significant due to two reasons. Firstly, while there are many editor-curated [2, 48] and data-driven [6, 9, 26, 56] media bias estimates, all of them focus on a notion of *average* leaning, whereas in this work we quantify the spread of the audience base. Secondly, our data-driven measure reveals significant cross-country variation in media domains, the average audience leanings, as well as the political diversity in the audience base.

6 CONCLUSION AND DISCUSSION

In this work, we compute a set of measurements to profile the audience of media outlets in eight countries. In particular, our profiling describes not just the average, but also the spread of user leanings for a given media outlet in a given country. Components of our quantitative method include: a new COVID-19 dataset with high coverage; a highly accurate set of geolocated users

based on their geotagged tweets and Twitter profiles; a set of comparatively estimated political leaning scores across countries; and a comprehensive validation of the audience leaning score against existing media bias reporting. We use the distributions of audience leaning to profile different media outlets within a specific country, and also profile the audience of a given outlet across different countries.

Broader implications. We posit that these new measures will provide scholars in political science and communications with a data-driven view of a diverse set of countries. The observations can inform media outlets of their comparative effectiveness within and across countries, and give activist groups and individuals insights with which to reflect on our day-to-day media diet. We posit that the methodology used here is transferrable to other large datasets from Twitter and other domains.

Ethical considerations. In this work, we report aggregate statistics and do not publish extracted geolocation, nor predicted political leaning of any individual Twitter users. We release datasets according to Twitter's terms of service and guidelines to academic researchers. We caution that interpretations of media leaning scores are influenced by audience sample, background distribution of left vs. right-leaning users in any country, and the importance of including non-English query terms in work that applies this method to other domains (see limitations below).

Limitations. It is important to acknowledge a number of limitations related to the data collection and filtering, user leaning estimation, and result interpretation.

- Despite COVID-19 being the dominant topic over the world in 2020, the media leanings and audience base are only estimated from COVID-19 related content in this work, rather than a representative set of all media content. An alternative data collection strategy would be extracting all news articles from the 1% Twitter data stream, and then replicating the geolocation extraction and political leaning estimation steps from this work. Much prior research studies the general news sharing on social platforms [3, 57]. However, one potential risk of using the general news articles is the topic misalignment of public discourse across different countries. Anchoring on a specific topic ensures the maximal comparability.
- The choices of English-only or English-majority query terms and hashtags introduce bias in the subsequent content analysis for non-English speaking countries. This is a gap that the current author team does not have the expertise to bridge. Countries such as Japan, South Korea, and Thailand have a significant Twitter presence but are not included due to the COVID-19 related queries being in the Latin alphabet and will not match their respective native languages.
- The profiled audience bases may still bias toward the subset of English-speaking Twitter users in the four non-English-speaking countries.
- This method only applies to countries with a significant Twitter presence. Users in many countries congregate on locally-preferred social media platforms (such as WhatsApp, Viber, Naver, among a large number of others). This means that our observations are limited to those who also engage on Twitter, despite that our proposed methodology could still be applied to data from other platforms, other countries, or other topics. Furthermore, user representation in countries with tight state censors is necessarily skewed towards those with the willingness and technical means to circumvent such barriers.
- While the retweet interactions with local politicians surface a collection of politically divided users, there is still an unknown number of politically modest users left out from this study.
- The COVID2020 tweet dataset focuses on COVID-19 related topics and is collected during a period when COVID-19 attracts an unprecedented amount of attention all over the world. We are therefore cautious about generalizing our findings to the general media consumption patterns or to other topics. Because if we replace the COVID-19 with another topic (e.g., abortion rights,

gun control, or Black Lives Matter activism [43]), we will obtain a new set of estimated user political leaning scores, thus a new set of media bias estimates. Some observations in the current manuscript may change under other topics.

The COVID-19 pandemic is a time when the whole world has come together to fight a global crisis. Given the COVID-19 news consumption can influence public health behaviors (e.g., social distancing, mask wearing, vaccine belief), it is of great importance to understand the news consumption patterns during the pandemic. This quantitative analysis presented in this work should be considered a supplement to, rather than a replacement for, in-depth examinations of media systems and dynamics in different countries. The analysis could be interpreted alongside data and analysis from other sources, such as politicians' Twitter presence and interactions [37]. We hope this work could lay the foundation for future analysis of media ecosystems in times of disaster.

Future work. It will be interesting to intersect media audience leanings with the public pandemic attitudes and actions in different countries. One could also quantify the robustness of media bias metrics with respect to different (and likely lower) sampling rates of Twitter data. We desire to expand Twitter queries and the scope of political leaning estimates, to potentially have multi-dimensional descriptions of media audiences over a larger number of countries. Another methodology challenge deserving more community efforts is the political leaning estimation on high-dimensional data, which can shift the research focus from modeling the oversimplified one-dimensional political axis to the two-dimensional Socio-Economic political compass.

ACKNOWLEDGMENTS

This work is supported in part by AFOSR Grant FA2386-20-1-4064. We are grateful for the computing and infrastructure support by Nectar Research Cloud, a collaborative Australian research platform supported by the NCRIS-funded Australian Research Data Commons (ARDC). We also thank the anonymous reviewers and area chairs for their valuable comments that helped shape our methodology and results.

REFERENCES

[1] AdFontes. 2023. Ad Fontes Media Bias Chart. https://adfontesmedia.com. (2023). [Online; accessed Apr-15-2023].
[2] AllSides. 2023. AllSides Media Bias. https://www.allsides.com/media-bias. (2023). [Online; accessed Apr-15-2023].
[3] Jisun An, Daniele Quercia, and Jon Crowcroft. 2014. Partisan Sharing: Facebook Evidence and Societal Consequences. In *Proceedings of the ACM Conference on Online Social Networks*.
[4] Sinan Aral. 2021. *The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health–and How We Must Adapt.* Penguin Random House.
[5] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.
[6] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to Ideologically Diverse News and Opinion on Facebook. *Science* (2015).
[7] Pablo Barberá. 2015. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis* (2015).
[8] Ceren Budak. 2019. What Happened? The Spread of Fake News Publisher Content during the 2016 US Presidential Election. In *Proceedings of the Web Conference*.
[9] Ceren Budak, Sharad Goel, and Justin M. Rao. 2016. Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. *Public Opinion Quarterly* (2016).
[10] Arthur Capozzi, Gianmarco De Francisci Morales, Yelena Mejova, Corrado Monti, André Panisson, and Daniela Paolotti. 2021. Clandestino or Rifugiato? Anti-Immigration Facebook Ad Targeting in Italy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
[11] Pew Research Center. 2011. The American-Western European Values Gap. https://www.pewresearch.org/global/2011/11/17/the-american-western-european-values-gap/. [Online; accessed Apr-15-2023].

[12] Pew Research Center. 2014. Political Polarization & Media Habits. https://www.pewresearch.org/journalism/2014/10/21/political-polarization-media-habits/. [Online; accessed Apr-15-2023].
[13] Pew Research Center. 2014. Where News Audiences Fit on the Political Spectrum. https://www.pewresearch.org/journalism/interactives/media-polarization/. [Online; accessed Apr-15-2023].
[14] Pew Research Center. 2018. News Media and Political Attitudes in France. https://www.pewresearch.org/global/fact-sheet/news-media-and-political-attitudes-in-france/. [Online; accessed Apr-15-2023].
[15] Pew Research Center. 2020. About Seven-in-Ten U.S. Adults Say They Need to Take Breaks From COVID-19 News. https://www.pewresearch.org/journalism/2020/04/29/about-seven-in-ten-u-s-adults-say-they-need-to-take-breaks-from-covid-19-news. [Online; accessed Apr-15-2023].
[16] Pew Research Center. 2020. Differences in How Democrats and Republicans Behave on Twitter. https://www.pewresearch.org/politics/2020/10/15/differences-in-how-democrats-and-republicans-behave-on-twitter/. [Online; accessed Apr-15-2023].
[17] Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Tracking Social Media Discourse about the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health and Surveillance* (2020).
[18] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political Polarization on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*.
[19] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the ACM Conference on Recommender Systems*.
[20] Boguslawa Dobek-Ostrowska, Michał Głowacki, Karol Jakubowicz, and Miklos Sükösd. 2010. *Comparative Media Systems: European and Global Perspectives*. Central European University Press.
[21] James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2021. How Affective Polarization Shapes Americans' Political Beliefs: A Study of Response to the COVID-19 Pandemic. *Journal of Experimental Political Science* (2021).
[22] Samantha D'Alonzo and Max Tegmark. 2022. Machine-Learning Media Bias. *Plos One* (2022).
[23] Gregory Eady, Richard Bonneau, Joshua A Tucker, and Jonathan Nagler. 2020. News Sharing on Social Media: Mapping the Ideology of News Media Content, Citizens, and Politicians. (2020).
[24] Gregory Eady, Jonathan Nagler, Andy Guess, Jan Zilinsky, and Joshua A Tucker. 2019. How Many People Live In Political Bubbles on Social Media? Evidence from Linked Survey and Twitter Data. *Sage Open* (2019).
[25] Eli J Finkel, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand, et al. 2020. Political Sectarianism in America. *Science* (2020).
[26] Richard Fletcher, Alessio Cornia, and Rasmus Kleis Nielsen. 2020. How Polarized Are Online and Offline News Audiences? A Comparative Analysis of Twelve Countries. *The International Journal of Press/Politics* (2020).
[27] Richard Fletcher, Craig T Robertson, and Rasmus Kleis Nielsen. 2021. How Many People Live In Politically Partisan Online News Echo Chambers in Different Countries? *Journal of Quantitative Description: Digital Media* (2021).
[28] GDPR. 2016. Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016. *Official Journal of the European Union* (2016).
[29] Matthew Gentzkow and Jesse M Shapiro. 2010. What Drives Media Slant? Evidence from US Daily Newspapers. *Econometrica* (2010).
[30] Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. 2019. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica* (2019).
[31] Yevgeniy Golovchenko, Cody Buntain, Gregory Eady, Megan A Brown, and Joshua A Tucker. 2020. Cross-Platform State Propaganda: Russian Trolls on Twitter and YouTube during the 2016 US Presidential Election. *The International Journal of Press/Politics* (2020).
[32] Jarod Govers, Philip Feldman, Aaron Dant, and Panos Patros. 2023. Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. *Comput. Surveys* (2023).
[33] Daniel C Hallin and Paolo Mancini. 2004. *Comparing Media Systems: Three Models of Media and Politics*. Cambridge University Press.
[34] Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated Identification of Media Bias in News Articles: An Interdisciplinary Literature Review. *International Journal on Digital Libraries* (2019).
[35] Austin Hegland, Annie Li Zhang, Brianna Zichettella, and Josh Pasek. 2022. A Partisan Pandemic: How COVID-19 Was Primed for Polarization. *The ANNALS of the American Academy of Political and Social Science* (2022).
[36] Sounman Hong and Sun Hyoung Kim. 2016. Political Polarization on Twitter: Implications for the Use of Social Media in Digital Governments. *Government Information Quarterly* (2016).
[37] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. Algorithmic Amplification of Politics on Twitter. *Proceedings of the National Academy of Sciences* (2022).
[38] Julie Jiang, Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020. Political Polarization Drives Online Conversations about Covid-19 in the United States. *Human Behavior and Emerging Technologies* (2020).

[39] Eunji Kim, Yphtach Lelkes, and Joshua McCrain. 2022. Measuring Dynamic Media Bias. *Proceedings of the National Academy of Sciences* (2022).

[40] Haewoon Kwak, Jisun An, Joni O. Salminen, Soon-Gyo Jung, and Bernard Jim Jansen. 2018. What We Read, What We Search: Media Attention and Public Attention Among 193 Countries. *Proceedings of the 2018 World Wide Web Conference* (2018).

[41] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media?. In *Proceedings of the International Conference on World Wide Web*.

[42] Rabindra Lamsal, Aaron Harwood, and Maria Rodriguez Read. 2022. Twitter Conversations Predict the Daily Confirmed COVID-19 Cases. *Applied Soft Computing* (2022).

[43] JooYoung Lee, Siqi Wu, Ali Mert Ertugrul, Yu-Ru Lin, and Lexing Xie. 2022. Whose Advantage? Measuring Attention Dynamics across YouTube and Twitter on Controversial Topics. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[44] Yphtach Lelkes. 2016. Mass Polarization: Manifestations and Measurements. *Public Opinion Quarterly* (2016).

[45] Haiko Lietz, Claudia Wagner, Arnim Bleier, and Markus Strohmaier. 2014. When Politicians Talk: Assessing Online Conversational Practices of Political Parties on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[46] Sonia Livingstone. 2003. On the Challenges of Cross-National Comparative Media Research. *European Journal of Communication* (2003).

[47] James Lo, Sven-Oliver Proksch, and Thomas Gschwend. 2014. A Common Left-Right Scale for Voters and Parties in Europe. *Political Analysis* (2014).

[48] MBFC. 2023. Media Bias/Fact Check. https://mediabiasfactcheck.com/mbfc-fact-checks. (2023). [Online; accessed Apr-15-2023].

[49] Scott D McClurg. 2003. Social Networks and Political Participation: The Role of Social Interaction in Explaining Political Participation. *Political Research Quarterly* (2003).

[50] Stuart E Middleton, Giorgos Kordopatis-Zilos, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. *ACM Transactions on Information Systems* (2018).

[51] Nic Newman, Richard Fletcher, C Robertson, Kirsten Eddy, and R Nielsen. 2022. Reuters Institute Digital News Report 2022. (2022).

[52] Juergen Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, Dennis Assenmacher, Siqi Wu, Diyi Yang, Cornelia Brantner, et al. 2023. Just Another Day on Twitter: A Complete 24 Hours of Twitter Data. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[53] John Platt. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers* (1999).

[54] Politico. 2022. 338Canada: Why Canada's conservatives are keeping quiet on abortion. https://www.politico.com/news/2022/05/05/canada-conservatives-abortion-00029908. [Online; accessed Apr-15-2023].

[55] Matthew Powers and Rodney Benson. 2014. Is the Internet Homogenizing or Diversifying the News? External Pluralism in the US, Danish, and French Press. *The International Journal of Press/Politics* (2014).

[56] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2018).

[57] Ana Lucía Schmidt, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2017. Anatomy of News Consumption on Facebook. *Proceedings of the National Academy of Sciences* (2017).

[58] M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. 2009. Extracting the Multiscale Backbone of Complex Weighted Networks. *Proceedings of the National Academy of Sciences* (2009).

[59] Leo Graiden Stewart, Ahmer Arif, A Conrad Nied, Emma S Spiro, and Kate Starbird. 2017. Drawing the Lines of Contention: Networked Frame Contests within #Blacklivesmatter Discourse. *Proceedings of the ACM on Human-Computer Interaction* CSCW (2017).

[60] Natalie Jomini Stroud. 2010. Polarization and Partisan Selective Exposure. *Journal of Communication* (2010).

[61] Hiroki Takikawa and Kikuko Nagayoshi. 2017. Political Polarization in Social Media: Analysis of the "Twitter Political Field" in Japan. In *2017 IEEE International Conference on Big Data (Big Data)*.

[62] Robert P Vallone, Lee Ross, and Mark R Lepper. 1985. The Hostile Media Phenomenon: Biased Perception and Perceptions of Media Bias in Coverage of the Beirut Massacre. *Journal of Personality and Social Psychology* (1985).

[63] Isaac Waller and Ashton Anderson. 2021. Quantifying Social Organization and Political Polarization in Online Platforms. *Nature* (2021).

[64] Siqi Wu and Paul Resnick. 2021. Cross-Partisan Discussions on YouTube: Conservatives Talk to Liberals but Liberals Don't Talk to Conservatives. In *Proceedings of the International AAAI Conference on Web and Social Media*.

[65] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. 2020. Variation across Scales: Measurement Fidelity under Twitter Data Sampling. In *Proceedings of the International AAAI Conference on Web and Social Media.*

[66] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with Local and Global Consistency. *Advances in Neural Information Processing Systems* (2003).

A COVID-19 KEYWORDS AND CRAWLER SETTINGS

We used 12 Twitter sub-streams to collect the COVID-19 tweets. The settings and assigned keywords are displayed below in JSON format. The `language` attribute indicates the corresponding collected language of tweets, with empty list indicating all languages.

All keyword list. [“coronavirus”, “covid19”, “covid”, “covid−19”, “COVID—19”, “pandemic”, “covd”, “ncov”, “corona”, “corona virus”, “sars-cov-2”, “sarscov2”, “koronavirus”, “wuhancoronavirus”, “wuhanvirus”, “wuhan virus”, “chinese virus”, “chinesevirus”, “china”, “wuhanlockdown”, “wuhan”, “kungflu”, “sinophobia”, “n95”, “world health organization”, “cdc”, “outbreak”, “epidemic”, “lockdown”, “panic buying”, “panicbuying”, “socialdistance”, “social distance”, “socialdistancing”, “social distancing”]

sub-stream 1

- keywords: [“wuhanlockdown”, “wuhan”, “kungflu”, “sinophobia”, “n95”, “world health organization”, “cdc”, “outbreak”, “epidemic”]
- languages: []

sub-stream 2

- keywords: [“lockdown”, “panic buying”, “panicbuying”, “socialdistance”, “social distance”, “socialdistancing”, “social distancing”]
- languages: []

sub-stream 3

- keywords: [“pandemic”, “covd”, “ncov”],
- languages: [“en”, “es”]

sub-stream 4

- keywords: [“coronavirus”]
- languages: [“en”]

sub-stream 5

- keywords: [“covid”]
- languages: [“en”]

sub-stream 6

- keywords: [“covid−19”, “COVID—19”, “covid19”]
- languages: [“en”]

sub-stream 7

- keywords: [“corona”, “corona virus”, “sars-cov-2”, “sarscov2”, “koronavirus”, “wuhancoronavirus”, “wuhanvirus”, “wuhan virus”, “chinese virus”, “chinesevirus”, “china”]
- languages: [“en”]

sub-stream 8

- keywords: [“coronavirus”, “corona”, “corona virus”]
- languages: [“es”]

sub-stream 9

- keywords: [“covid19”, “covid”, “covid−19”, “COVID—19”, “sars-cov-2”, “sarscov2”, “koronavirus”, “wuhancoronavirus”, “wuhanvirus”, “wuhan virus”, “chinese virus”, “chinesevirus”, “china”],
- languages: [“es”]

sub-stream 10

- keywords: [“coronavirus”, “corona”, “corona virus”, “covid19”, “covid”, “covid−19”, “COVID—19”, “sars-cov-2”, “sarscov2”, “koronavirus”, “wuhancoronavirus”, “wuhanvirus”, “wuhan virus”, “chinese virus”, “chinesevirus”, “china”]
- languages: [“th”, “it”, “fr”, “tr”]

sub-stream 11

- keywords: [“coronavirus”, “corona”, “corona virus”, “covid19”, “covid”, “covid−19”, “COVID—19”, “sars-cov-2”, “sarscov2”, “koronavirus”, “wuhancoronavirus”, “wuhanvirus”, “wuhan virus”, “chinese virus”, “chinesevirus”, “china”]
- languages: [“in”, “ko”, “ja”, “und”]

sub-stream 12

- keywords: [“coronavirus”, “corona”, “corona virus”, “covid19”, “covid”, “covid−19”, “COVID—19”, “sars-cov-2”, “sarscov2”, “koronavirus”, “wuhancoronavirus”, “wuhanvirus”, “wuhan virus”, “chinese virus”, “chinesevirus”, “china”]
- languages: [“pt”, “zh”, “ar”, “de”, “tl”, “cs”, “vi”, “pl”, “ru”, “sr”, “el”, “nl”, “hi”, “da”, “ro”, “is”, “no”, “hu”, “fi”, “lv”, “et”, “bg”, “ht”, “uk”, “lt”, “cy”, “ka”, “ur”, “sv”, “ta”, “sl”, “iw”, “ne”, “fa”, “am”, “te”, “km”, “ckb”, “hy”, “eu”, “bn”, “si”, “my”, “pa”, “ml”, “gu”, “kn”, “ps”, “mr”, “sd”, “lo”, “or”, “bo”, “ug”, “dv”, “ca”]

B POLITICAL DIVIDING HASHTAGS FOR THE 2020 US PRESIDENTIAL ELECTION

Received January 2023; revised April 2023; accepted July 2023

cult45sucks, ditchmitch, lockuptrump, trumppandemic, thanksobama, impeach, joebiden2020, stevencrowderdrinksdogcum, byedon2020, impeachthemf, trumpisdone, votebiden, leftwing, bunkerbabybonespurs, removetrump, dumptrumphesachump, 45notmypresident, 25thamendment, trumpisatrator, wrongtrump, wearredvoteblue, fuckdonaldtrump, rumpresident, bernietulsi2020, covidiots, uniteblue, medicare4all, noantiblackracism, turnthesenateblue, yanggang2020, onlybernie, americaneedsyang, cult45, yangorbust, shit_trump_would_say, joebiden2020forpresident, nevertrump, faketrumpf17klies, anyonebuttrump, idiottrump, trumpgonnagetschooled, firetheliar2020, pencedemic, trumpsupportersneedtobekilled, warren2020, taketheorangeclownout2020, settleforbiden, stillvotingyang, moretrumplies, prochoiceoneverthingoneverthing, traitors4trump, thetrumpvirus, trumptrash, votebernie, bidenharris2020touniteandrebuildamerica, cowdercrowder, prochoice, republicansforjoebiden, putinspuppet, outtrump, womenforbernie, fuckthegop, trumpsucks, bluetsunami, bunkerbaby, bluewave, liberal, trumpliespeopledie, impeachtrump, trumpvirus2020, trumpforprison2020, leftisbest, voteagainsttrump, typhoidtrump, biden2020, whiteprivilegeisreal, trumpshutdowncps, imwithher, trumptarded, trumpisthehoax, bernieorvest, trumpliesaboutcoronavirus, berniecomeback, trumpexposed, crimesagainsthumanity, obamaenvy, yanggangforever, berniebeatstrump, bidenforpresident, trumpcult, theresistance, yangmediablackout, resistance, voteprogressives, votebiden2020, tre45on, lockhimup, gojoe, trump4prison2020, trumpisnotwell, trumpforprison, trumpplague, traitortrump, neverdonaldtrump, trumpimpeachment, magaisformorons, voteblue2020, byedon, shoottrumpnow, fucktrump, soundslikeyang, donaldwho, bernieismypresident, votebluetoendthisnightmare, trumpslump, trumptard, resisttrump, bernieorbust, orangemoron, trumpresign, dumptrump2020, berniesanders2020, medicareforall, itrustbernie, trumplies, couldahadyang, feelthebern2020, joe2020, impeachdrumf, dumpthetrump, cult45goppathetic, trumpmafia, humanityfirst, obamagreat, liberatethewhitehouse, writeinbernie, gopprodeath, donthecon, trumpflu, impeachbarr, blackwomenforyang, usnotme, trump4prisondent2020, ilikebernie, presidentialfailure, impotus, fakepotus, russiagate, byetrump, bernie2020, cult45gop, yang2020, bluenomatterwho, notmypresident, impotus45, ripgop, firetrump, resist, nevadafor presidenttrump, teampelosi, nevertrumper, jailtrump, failedtrump, bluedowntheballot2020, lettulsispeak, dictatordrumpf, lockthemallup, trumpcrash, foreverbernie, drumpf, swampydon, voteblue, republicanpandemic, bunkerboy, votingallblue2020, bidenharris, votebluenomatterwho2020, bernthednc, resigntrump, stopdonaldtrump, republicansagainsttrump, bluewave2020, endtrumpism, andrewyang2020, trumpthecorrupt, votebluenomatterwho, crookeddonald, andrewyang4pres, votethemallout, euthanizetrumpsupporters, throwouttrump, joebiden4president, bluemaga, moscowmitch, treasonoustrump, sanders2016, presidentsanders, cuomo2024, yanggang2024, yangbeatstrump, magavirus, lockthemup, trumpgate, yangganglove, presidentpussyassbitch, votethemout, votetrumpout2020, gothsforbernie, youtubeandrewyang, imwithhilary, votetrumpout, fakepresident, coldfeetcrowder, votebluein2020, sanders2020, onevoicel, trumpdepression, letyangspeak, trumptardforprison2020, makeamericahonestagain, justicedemocrats, diaperdonald, mybodymychoice, dementiadonnie, 1yang2024, time4trumptoburn, yangmentum, downwithnazitrump, bernie2020ournations21stcenturyfdr, traitor45forprison2020, magats, christiansagainsttrump, voteblueinnovember, cuomoforpresident, trumpfuckedup, countryaboveparty, takehimout, yang2024, shameontrump, trumpthemurderer, magatards, presidentberniesanders, votebluetosaveamerica, notmeus, ridinwithbiden, fakepotusrealcriminal, flipthesenate, berniebros, resignnow, republictards, regredvoteblue, trumpandemic, tulsi2020, presidentcuomo, anyonebuttrump2020, locktrumpup, feelthebern, bunkerbitch, stillsanders, americaortrump, standwithtulsi, donaldsucks, impeachdt, votehimout, onlybernie2020, nevertrumpers, stoptrump, bernieorbust2020, trumpsfault, trumppenceoutnow, dumptrump, trumpfraud, donaldjtrumpfebruaryfailure, comradetrump, berniebro, trumpthelyinginsanefascistbigot, yangwillwin, drainthewhitehouse, voteblueacrosstheballot, trumpthetraitor, connecttheleft, resistance2020, liarinchief, bidenridenbluewave, trumpfearsyang, liberal, democrat, democrats, democratic, bidenwins2020, reopentard, despicabledonaldtrump, trumpsupportersareinferior, voteforbernie, fucktrump2020, bernie2016, trumpprison2020, bidenharris2020, yangwasright, cuomo2020, defundtrump, bunkerboytrump, thanksbiden, bluetsunami2020, impeached, trumpcrimefamily, farmersagainsttrump, trumpvirus, trumpviruscoverup, bloomberg2020, republicansareterrorists, bernienina2020, berniecares, m4a, worstpresidentever

Table 3. Left-leaning hashtags.

walkedaway, blacks4trump2020, supertrump, wwg1wgaww, patriotsforafreeamerica, wakeupamerica, defundthedemocrats, nevervotedemokratagain, magahats, nra, darktolight, greatawakening, voterfraud, patriotsunite, rednationrising, latinos4trump, thinblueline, fourmoreyears, impeachmentmeansnothing, shutdowntheleft, thecollectiveq, donjr2024, trumpnation, qanon, bluethinline, trumpforever, demokkkrat, trumpcheerleaders, teamtrump, kingtrump4ever, trump2020landslidevictory, buildthewall, prolife, pence2020, wwg1wga, kaga2020, draintheswamp, godwins, qanonproof, bluelifematters, australiafortrump, trump2020wwg1wga, imwithq, blacksfortrump2020, sayno2vaccines, puertoricans4trump, stupidliberal, hispanicwomen4trump, keepamericagreat, veteransfortrump, demokkkrats, pro2a, magababymaga, democratsaredummies, trump2028, drainthatswamp, womenfortrump, trump2024, kaga, potustrumpmaga, supporttheblue, redpill, lovepresidenttrump, liberalssuck, godwinsalways, qarmy, libtards, defundthedemoncrats, qdigitalarmy, votredtsunami2020, trumppence2020, qanon8chan, democratssuck, trumpusamerica, voteredtosaveamerica, wethepeople, voteredtosaveamerica2020, america1st, ccot, filmyourhospital, gaysfortrump, itsnotyourbody, obamagate, 4mreyrs, trumpisyourpresident, latiosfortrump, blacks4trump, trumpderangementsyndrome, trumptrain2020, lovepotus3545, qanons, kag, lationsfortrump, wedonotconsent, bestpresidentever, americansfirst, maga2020, stillyourpresident, qanonmerch, trump2020, draintheswamp, kag2020, trump4eva, 4moreyears, noblackvoteforbiden, trumpocrat, proudpatriot, maga, votedemsout, uniteright, magarapper, istandwithtrump, qpatriot, bestpresidentever45, lovemypresident, trump4life, trump2q2q, godblessourpresident, bidenishidingbecausehismindissliding, trumpforallallfortrump, redpilldiaries, magabeanie, impeachdem, patriots, 2astrong, lawandorder, witnessingthegreatawakening, mastertrump, americafirst, lockherup, votedemout, godlives, chicanoconservative, trumpourimmortalbeaconoflight, blackvoicesfortrump, walkway, democratsfortrump, votebluetored, thegreatawakening, makeamericagreatagain, wwg1wga_ww, djt2020, tucker2024, demexitedin2016, vivatrump, conservatism, votetrump, buildthewallandcrimewillfall, glockherup, dethronethedemoncrats, magaland, uncensoredpatriot, gays4trump, redwaverising, trustq, ivanka2024, walkawaycampaign, trump2016, blacksfortrump, impeachpelosi, armyfortrump, jesusmatters, athiestdonaldjtrumpsupportertrumpandus, mrpresidenttrump, screwliberals, backtheblue, shadowgate, standforourflag, filmyourhospitals, democratsaredestroyingamerica, trump2048, presidenttrump, mexicans4trump, 2a, redwave, conservativenews, wwg1wgall, 45isthebest, trillionairesfortrump, kag2q2q, votetrump2020, candaceowens2024, wwg1wgaworldwide, demonrats, shapiro2024, pray4djt, redpillhardcore, votered, abolishdemocrats, trumptrain, clintonbodycount, conservative, latinos4trump2020, pro2ndamendment, trumpismypresident, qalert, patheticdemocrats, redpilltheyouth, screwthedemocraticparty, gotrump, realq, patriot, bluelivesmatter, redwave2020, liberalismiscancer, conservatism, conservative, republican, republicans, qed, trumplandslide2020, awakening, altright, redpilltaken, votedemsoutofoffice, trumpgirlamericafirst, trumpwwg1wga, keepamericagreatwithfourmoreyears, backthepolice, staywoke, whenyoutakeredpills, wherewegoonewegoall, wakeupisrael, liberalismisamentaldisorder, latinosfortrump, wwg1wwa, mypresident, democratshateamerica, alexjonesisright, tcot, trumpsupporter, wwg1wga_worldwide, redpillhardcoreradioshow, democratscreatedthekkk, abortionismurder, redpilled, avidtrumpsupporter, wwg1wgallllll, trumpjr2024, latinosfortrump2020, walkaway, buildthatwall, godisreal, wgw1wga, clintonsforprison, trump2020landslide, 2ndamendment, trump2020nowmorethanever

Table 4. Right-leaning hashtags.