**Who is Curating My Feed? Characterizing Political Exposure of Registered U.S. Voters on Twitter**

Assaf Shamir1, Jennifer Oser2, and Nir Grinberg1

1 Department of Software and Information Systems Engineering, Ben-Gurion University, Israel

2 Department of Politics and Government, Ben-Gurion University, Israel

**Abstract**

Social media platforms offer people a variety of options to engage with politics, from directly following elected officials to discussing politics with social peers. Despite major advances in recent research on online political exposure through the lens of selective exposure, filter bubbles, and ideological echo chambers, little is known about the fundamental questions of what types of political content people are exposed to on social media, and what kind of people are exposed to distinctive content types. We address this gap in the literature by analyzing a unique panel of more than 600,000 registered U.S. voters during the 2020 U.S. Presidential campaign. The findings identify distinct types of political consumers on Twitter that vary in the amount of political content in their feeds, the political alignment of the content available to them, and the composition of sources who provide the content either directly or indirectly (e.g., media organizations, journalists, politicians, opinion leaders, and social peers). Our identification of prototypical exposure types and how they vary across key socio-demographic characteristics advances our understanding of the way citizens learn about politics, and paves the way for next-step research to identify the causal effect of political content exposure on individuals’ political attitudes and political behavior.

Keywords: political content exposure, online media diets, prototypical exposure types, social media, Twitter

Funding acknowledgment: This work is supported by the Ben-Gurion University Data Science Research Center’s Young Faculty Interdisciplinary Grant 2021-2022 [to Grinberg and Oser as co-PI’s]; by Ben-Gurion University’s Interdisciplinary Fellowship for Outstanding Graduate Students [to Shamir]; and by the Israel Science Foundation [to Oser, grant number 1246/20].

# Introduction

The tectonic shifts in the media environment and the rise of social media platforms over the past two decades significantly changed the ways in which people are exposed to news and political information worldwide [(Fletcher and Nielsen, 2018; Shearer, 2018)](https://www.zotero.org/google-docs/?HMtDVg). This trend has been particularly swift in the United States, as Americans are now exposed to news more often on social media than in print, and for younger generations social media is the dominant channel for news [(Shearer, 2018)](https://www.zotero.org/google-docs/?CluboY).

In this increasingly networked media environment, the information that populates one's feed is an amalgamation of curation decisions taken by others, including social peers, journalists, politicians, advertisers, and proprietary ranking algorithms [(Thorson and Wells, 2016)](https://www.zotero.org/google-docs/?2KCWwf). For example, Bakshy et al. [(2015)](https://www.zotero.org/google-docs/?Us6Fpv) showed how selective exposure on Facebook is partially determined by Facebook's Newsfeed ranking algorithm and more dominantly determined by the individual's choice of whom to follow. Of course, the effects of social media and digital media use writ large extend beyond the online world, with a growing body of research showing mobilization effects, where digital media use of various kinds can lead to more traditional forms of political participation offline such as voting [(Oser and Boulianne, 2020; Vaccari et al., 2015)](https://www.zotero.org/google-docs/?J6TdrH). Therefore, it is no surprise that issues of power and control [(Barzilai-Nahon, 2008)](https://www.zotero.org/google-docs/?xv53IS), limits of free speech [(Morrow et al., 2021)](https://www.zotero.org/google-docs/?qk4n1h), and individual choice [(Bakshy et al., 2015)](https://www.zotero.org/google-docs/?9GSzQV) in political exposure on social media are some of the most contested topics of our time. Better understanding of the extent to which different actors can wield that power is critical for advancing our understanding of political communication in the 21st century and for developing guiding principles for designing the next generation of these systems.

Yet, we know relatively little about political exposure on social media, the prevalence of different actors in people’s social feeds, and how political exposure varies across socio-demographic and political groups. Currently, no social media platform provides precise individual-level or comprehensive aggregate-level information about exposure. The Social Science One initiative [(King and Persily, 2020)](https://www.zotero.org/google-docs/?GcD1GK) does provide aggregate information about viewership, but this information is currently limited to Facebook data, includes only URLs and not all political content, does not distinguish eligible from non-eligible voters, and does not provide information about the person who posted the content. In lieu of more precise measurement, researchers have relied on self-reported measures of political consumption and general-purpose web tracking trace data [(Guess, 2021; Weeks et al., 2017; Wojcieszak et al., 2022b)](https://www.zotero.org/google-docs/?8jAs9A).

In this study, we leverage a large panel of 1.8 million U.S. registered voters and their activity on Twitter to identify the prototypical types of exposure to political information and examine their distribution across the U.S. registered voter population. The combination of these two data sources (i.e., Twitter data and registered voter data) creates the opportunity to simultaneously analyze patterns of individuals’ political content exposure and their socio-demographic characteristics. We conduct this analysis by building on the curated flows theoretical framework [(Thorson and Wells, 2016)](https://www.zotero.org/google-docs/?Ky9nUg) to identify the political content available to individual registered voters on Twitter that originates in distinctive streams of political content curated by different actors, including media organizations, journalists, politicians, opinion leaders, and social peers. By using unsupervised clustering methods, we identify the prototypical modes of political exposure by registered voters. Finally, we identify socio-demographic covariates that significantly associate with different modes of exposure, and examine the prevalence of these modes among distinctive socio-demographic groups of registered voters. This research design allows us to address a key problem identified by Prior (2013: 102) in his comprehensive review of the connection between media exposure and political attitudes almost a decade ago, namely that we need to better understand “how many and what kind of people are exposed to which messages.” Further, we identify the actors that are responsible for this content exposure.

Therefore, our contributions are twofold. First, we provide new empirical evidence about the prototypical modes of political exposure – both in terms of quantity of political content, and composition of different actors who curate this content – by a large and representative sample of registered U.S. voters on Twitter. Second, we offer findings on the varying levels and compositions of political exposure by different socio-demographic groups of registered U.S. voters on Twitter. Taken together, our contributions begin to address some of the most basic, yet unanswered, questions at the heart of the curated flow framework based on analysis of uniquely comprehensive individual-level data, namely: who are the most significant curators and for whom.

# 

# The Importance of Political Exposure Online and on Social Media

Numerous studies show that online political exposure and information consumption on social media are related to political attitudes and behaviors, both online and offline. A recent meta-analysis found that incidental exposure, an unintended form of exposure that is common on social media, is positively associated with a variety of pro-democratic attitudes and behaviors including news use, political knowledge, political participation, expressive engagement, and political discussion [(Nanz and Matthes, 2022)](https://www.zotero.org/google-docs/?Xw5WQL).

The impact of political messaging also depends on the identity of the messenger who is delivering it, as the same political message received from different types of sources may have a differential impact on societal attitudes and behavior. For example, recent research indicates that statements by political figures and online celebrities seem to influence the public's real-world beliefs compared to similar statements by non-celebrities [(Alatas et al., 2019; Suuronen et al., 2021)](https://www.zotero.org/google-docs/?YXhlxR). In the realm of media sources, research has shown that high levels of exposure to media outlets with high levels of political content shape political knowledge and behavior, including the propensity to vote [(de Vreese and Boomgaarden, 2006)](https://www.zotero.org/google-docs/?CHYeJh). Turning to the domain of peer networks, research by Graham et al. [(2015)](https://www.zotero.org/google-docs/?oMoBRy) showed that over half of the political discussions in online forums in the U.K. led to at least one political action. The importance of identity cues on opinions and online behaviors as a causal mechanism is evident in Taylor et al’s [(2022)](https://www.zotero.org/google-docs/?eOPQuL) large-scale longitudinal field experiment, which provides robust evidence of meaningful effects of the identity of content providers on how viewers vote and reply to shared content. Taken together, we observe that this emerging research indicates that understanding the identity of the messenger may be as important as the message itself for assessing the impact of political content exposure on individuals’ subsequent attitudes and behaviors.

# 

# Who is Heard and by Whom in Political Communication

A central element in democratic theory is individual expression of different ideas in ways that allow some form of public information-sharing and deliberation [(Habermas, 1984)](https://www.zotero.org/google-docs/?xC5jjq). While research on the ways citizens construct their information diets certainly precedes the digital era [(Katz and Blumler, 1974; Sears and Freedman, 1967)](https://www.zotero.org/google-docs/?GN5bql), the shift to online media – accompanied by the weakening of traditional gatekeepers, and the context collapse that is common on social media [(Davis and Jurgenson, 2014)](https://www.zotero.org/google-docs/?2AenPu) – calls for renewed attention to the fundamental question of who is being heard in modern political communication. Addressing this question is important for advancing our understanding of the extent to which social media, and information systems more broadly, fulfill their egalitarian potential [(Allen, 2015)](https://www.zotero.org/google-docs/?XcTn0A) or alternatively reinforce old political structure and power as the weapon of the strong [(Hindman, 2009)](https://www.zotero.org/google-docs/?6uuyS1).

The theoretical and empirical importance of examining who is being heard is highlighted by Thorson and Wells' [(2016)](https://www.zotero.org/google-docs/?xOjqAM) discussion of the role of individual-level "curation" for understanding media exposure and its effects. In particular, the curated flows framework lists a number of key actors including social peers, journalists, politicians, advertisers, and proprietary ranking algorithms. However, there is little empirical work that shows the relative prevalence of different actors in the public's political exposure. A notable exception is the recent work by Wojcieszak et al. [(2022b)](https://www.zotero.org/google-docs/?uGXImo), which used browsing histories of nearly 700 survey participants to yield new insights on the channels (search engine, social media, aggregators, etc.) that lead people to news. The research design of Wojcieszak et al.’s [(2022b)](https://www.zotero.org/google-docs/?60Kni7) study shed new light on sources of news exposure through active engagement (i.e., consider exposure that results in a "click" or a page visit).

Wojcieszak et al.’s [(2022b)](https://www.zotero.org/google-docs/?XcWs2w) contribution raises the important next-step question of what are the prototypical types of exposure to political content in general, in addition to exposure to news. Answering this question requires attention to both the amount of political content people are exposed to, as well as the different kinds of actors who convey this political content. This type of individual-level attention to the quantity of political content exposure and the source of this curated content is necessary to pave the way for next-step causal examination of how distinct types of political exposure may influence individuals' political attitudes and behaviors. Therefore, our first research question is the following:

**(RQ1) What are the prototypical types of political exposure on social media both in terms of overall quantity and the composition from different types of actors?**

A key element in the composition of political exposure is political ideology and the range of ideas being represented. Some recent work indicates that exposure to political content through online social networks may serve to increase political polarization [(Bail et al., 2018; Garrett et al., 2014; Levy, 2020; Shmargad and Klar, 2020)](https://www.zotero.org/google-docs/?RPQmr1). Yet, other studies indicate that social media exposure through weak ties and the visibility of social endorsements reduce polarization by offering diversity of exposure [(Barberá, 2015; Messing and Westwood, 2014)](https://www.zotero.org/google-docs/?hsuFLC). As noted by Zhuravskaya et al. [(2020)](https://www.zotero.org/google-docs/?9oZRLI), the literature on these topics to date has not yet reached consensus on the strength and the direction of the connection between social media and polarization. Importantly, recent work shows that the ideology of media and politicians accounts that people follow on Twitter varies considerably for different ideological groups [(Eady et al., 2019; Wojcieszak et al., 2022a)](https://www.zotero.org/google-docs/?YgbcnN), but there is still substantial overlap in the ideology of media content between liberals and conservatives [(Guess, 2021)](https://www.zotero.org/google-docs/?XFPHxq). Clearly, political exposure varies by ideology, and thus it is important that we model left-right ideology jointly with the composition of exposure from different sources.

ocio-demographic characteristics of political consumers also matter for the attention people pay to politics. Firstly, there is a well-documented age gradient observed in the level of interest in politics and the self-efficacy of individuals [(Verba et al., 1995)](https://www.zotero.org/google-docs/?KY9X8d). As younger generations increasingly get their news on social media [(Shearer, 2018)](https://www.zotero.org/google-docs/?MIQ5v6), it is important for researchers to pay attention to the types of political content to which they are exposed. In general, research has tended to show that those who have more traditionally advantaged socio-demographic backgrounds (e.g., more male, older) are more actively politically engaged in general, including in their efforts to seek out political news and content [(Schlozman et al., 2018, 2012)](https://www.zotero.org/google-docs/?mGwZf1). Yet research suggests that social media and online participation may have differential mobilization effects for different socio-demographic groups. Specifically, research focused on the connection between social media and political participation suggests that online participation, including social media engagement, may recruit younger groups and women more actively into politics than traditional offline channels do [(Oser et al., 2013; Oser and Boulianne, 2020; Xenos et al., 2014)](https://www.zotero.org/google-docs/?g61F5N). Prior work also observed how different publics and counterpublics pay attention to different topics on social media and offline [(Jackson et al., 2018; Jackson and Welles, 2015; Shugars et al., 2021)](https://www.zotero.org/google-docs/?jDCaoj). Therefore, our second research question is the following:

**(RQ2) How do the prototypical types of political exposure on social media vary for distinctive socio-demographic groups?**

Informed by this review of the features of political content exposure that may have theory-based implications for contemporary political attitudes and behaviors, we now turn to assess the methodological challenges and opportunities for creating a robust empirical digital-political footprint for each citizen in our panel.

# Measuring Political Exposure in the Digital Age

Observational survey data have long been a leading source of vital information for examining political news consumption habits [(Iyengar and Hahn, 2009; Lazarsfeld et al., 1944; Mutz, 2001)](https://www.zotero.org/google-docs/?vjBa3m). However, observational, self-reported surveys pose multiple methodological challenges that social scientists have recently aimed to address more robustly in efforts to both improve existing survey methods [(Berinsky, 2017; Guess, 2015)](https://www.zotero.org/google-docs/?qGau64), and to identify alternate and complementary data and methodologies for gaining new knowledge about social and political phenomena with a focus on causal identification [(Samii, 2016)](https://www.zotero.org/google-docs/?69KSIW). For the study of online communication and media diets more broadly, prior research shows that self-reported survey data suffer from limited reliability due to self-reported measures, which tend to be biased due to socio-political attributes [(Prior, 2013; Scharkow, 2016)](https://www.zotero.org/google-docs/?rDXm3p). There are also large discrepancies between actual and reported frequency of posting about politics for the more active users and more recent activity [(Guess et al., 2019; Henderson et al., 2021)](https://www.zotero.org/google-docs/?L7fgtT).

Digital trace data provide new and complementary ways to measure individuals’ behavior directly. Recent research on digital phenomena increasingly applies methodological approaches for gathering digital trace data, often collected through dedicated software installed by participants [(Flaxman et al., 2016; Garrett, 2009)](https://www.zotero.org/google-docs/?rfgEwk). For example, Guess (2021) uses web browsing data to characterize Americans’ media consumption habits and examine whether internet use indeed facilitates selective exposure to like-minded views. As modeled in Guess’s (2021) study, this type of digital trace data approach is ideally combined with survey responses. While this approach provides the most comprehensive picture of both objective and subjective measures of political engagement, it is often limited to a few thousand participants who are willing to volunteer their data. In addition to the selection issues, the sample quickly becomes statistically underpowered for obtaining accurate descriptions of subgroups and their heterogeneity of activity [(Hughes et al., 2021)](https://www.zotero.org/google-docs/?yoMKGY).

A recently developed alternate approach for directly gathering data on individuals’ behavior is to use publicly available social media data. Due to the active engagement of media outlets and political figures on the platform, Twitter has been a uniquely important context in which to investigate behavioral exposure of a large user base to political content in recent years [(Bail et al., 2018; Barberá, 2015; Eady et al., 2019; Guess, 2021)](https://www.zotero.org/google-docs/?YJ1Xln). Around one-in-five Americans use Twitter [(Odabaş, 2022)](https://www.zotero.org/google-docs/?gINu4T), and almost seven in ten of them say they get their news regularly through the platform [(Mitchell et al., 2021)](https://www.zotero.org/google-docs/?T8mVMP). While Twitter users in the U.S. were found to be younger and more likely to be Democrats in comparison to the general public [(Wojcik and Hughes, 2019)](https://www.zotero.org/google-docs/?us2UEu), prior work has shown that differences between Twitter users and non-users are due mostly to the demographic composition of social media users, which can be addressed by controlling for demographic variables [(Mellon and Prosser, 2017)](https://www.zotero.org/google-docs/?Tu2FFc). Importantly, rigorous empirical work on the representativeness of Twitter users shows some modest demographic differences between Twitter users and the general population that can be accounted for analytically [(Hughes et al., 2021)](https://www.zotero.org/google-docs/?slG7Vx).

Nevertheless, currently no social media platform offers public access to data about individuals' exposure to distinctive types of political content, and as a result hardly any research has directly measured it. An increasingly prominent approach for approximating exposure involves the collection of content posted by accounts followed by the focal user on social media [(Eady et al., 2019; Grinberg et al., 2019)](https://www.zotero.org/google-docs/?j2KhVU). This approach does not guarantee exposure, i.e. that an individual actually saw a particular post, but it does directly speak to the content available to people in their social feeds from their ego-network.

As detailed in the following section, in the current study we apply novel methodological approaches to measuring and analyzing publicly available behavioral data on social media to obtain robust measurements of prototypical types of political content exposure and how these types vary across key socio-demographic groups.

# 

# Data and Methods

## *Twitter Panel and Representativeness*

The primary dataset used in this work is a panel of nearly 1.8 million Twitter users that were successfully matched to public U.S. voter registration records. Following the same approach described in prior work, a Twitter account was matched to a voter record if and only if their full name exactly matched and they were the only person with that name in either the city- or state-level geographic area specified in both datasets (see Grinberg et al., 2019 and Shugars et al., 2021 for more details). Importantly, this matched dataset provides comprehensive data on individuals’ social media behavior through Twitter, as well as the basic socio-demographic information available in public voter registration records.

We further restrict our analysis to the set of 606,112 panel members who were minimally active on Twitter during the 2020 presidential election (August to November, inclusive). By focusing on a period of a presidential election, we examine political exposure at its peak [(Grinberg et al., 2019; Peterson et al., 2021)](https://www.zotero.org/google-docs/?1WHuIt). Our criterion for inclusion in the analysis is that panelists have been minimally active on the platform during the four months of the analysis by posting or liking at least one tweet during the entire period.

As our target population is restricted to registered U.S. voters on Twitter, we do not make claims about the important, yet omitted, populations of eligible non-registered voters and of people inactive on Twitter. Moreover, the dependence on full names and disclosed locations raises concerns about potential selection bias. However, a rigorous comparison of this panel with a gold-standard survey conducted by Pew Research Center shows that only small demographic and ideological differences exist between the two samples of registered U.S. voters, which can be corrected by standard post-stratification methods [(Hughes et al., 2021)](https://www.zotero.org/google-docs/?li1xsm).

To model exposure, we follow the approach used in prior work that approximates individuals' exposure using the content available from the accounts they follow [(Eady et al., 2019; Grinberg et al., 2019)](https://www.zotero.org/google-docs/?BpRW7Y). Our panel follows a total of 51 million unique Twitter users, which make it impossible to collect all of their tweets through the Twitter API. Instead, we work around this limitation by analyzing the content in a 10% random sample of Twitter ("decahose") posted by accounts followed by the panel, similar to the approach in Grinberg et al. (2019). An important limitation to this technique is that it only provides a sample of the content from the individual's network and not all of it. Specifically, this technique does not include ads and does not consider algorithmic ranking. Yet, in lieu of more precise information from social media platforms about exposure, this approach reflects the most accurate and reproducible estimate currently available for the composition of people's social feeds.

## 

## *Identifying Political Tweets*

We train a Machine Learning classifier to identify political tweets and validate its accuracy against human coders, similar to the approach used in prior work [(Bakshy et al., 2015; Eady et al., 2019; Grinberg et al., 2019)](https://www.zotero.org/google-docs/?zjVOdV). In particular, we define a set of whitelist terms including political keywords, candidate names and usernames, and hashtags relevant to the election that identify political tweets with high probability. In order to capture the varying nature of politics during an election cycle, we trained our logistic regression classifier daily to distinguish tweets containing whitelist terms from a random sample, which enables the classifier to learn the additional terms that associate with politics each day. We evaluated the performance of our classifier by crowdsourcing labels for a random sample of about 2000 tweets that were stratified over time. The classifier resulted in precision of 88.8% and recall of 80.0% for political tweets, and a recall of 96.4% when further restricted to subcategory of election-related tweets. More details about the classifier and its validation are in Appendix A.

## 

## *Identifying Different Actors in Political Exposure*

Following the curated flows framework, we examine different types of actors that curate political content for individuals in our panel. We focus on four types of actors directly mentioned in Thorson and Wells' (2016) framework – social peers, politicians, journalists and media organizations – and include a fifth category of "opinion leaders" who have been identified as important in recent research. Specifically, opinion leaders’ accounts have a large followership on social media and a demonstrated ability to influence public opinion [(Alatas et al., 2019)](https://www.zotero.org/google-docs/?bQzh45). We identify influential accounts by following Bail et al.'s [(2018)](https://www.zotero.org/google-docs/?0eGmSD) approach of considering as an opinion leader any account followed by 15 or more active members of congress (MoC). It is most likely that by using such a definition, leads to inclusion of politicians’, media outlets’, and reporters’ accounts together with other opinion leaders (i.e., public figures and organizations). To categorize and separate these accounts…

Remove accounts from existing lists (magda, yphtach, ours)

To cope would also be included in this pool of accounts. combine a curated of public figures, organizations and other influential entertainment accounts, curated by CITE YPHTACH; Politician accounts are comprised of two sources; the first is an original list we have compiled by linking an official list of the 116th MoC names to a list of MoC accounts on Twitter [(Wrubel and Kerchner, 2020)](https://www.zotero.org/google-docs/?43zIsF). The list includes both politicians' accounts and their election campaigning accounts, which is important for capturing all messages that originate from politicians during an election cycle. We manually validated the politicians' list for completeness and accuracy. The second source for politician accounts is a list by CITE MAGDALENA. For media organizations, we used a list that started with a seed list of known media organizations' accounts and used snowball sampling to iteratively add organizations that appear on Twitter lists (curated by Twitter users) with many of the identified accounts[[1]](#footnote-1). For journalists’ accounts, we combine a list provided by CITE MAGDALENA with an additional list of 95 journalists’ accounts that we identified (and manually validated) while compiling the media organizations’ list, as these journalists included their media outlet’s name as part of their screen name, and thus, enabled an easy identification.

Importantly, people can play multiple roles in sharing political content, and the same content can be attributed to multiple people due to retweets, quotes, replies and mentions. For example, a politician can also be an influencer and a social peer. To maximize interpretability, we assign accounts only to a single category by first assigning politicians and media accounts, then influencers, and finally social peers. The influencers list excludes active MoC and media organization accounts by using inexact string matching and manual validation (see Appendix D for details). To support different attributions of content and interpretations of the results, we distinguish between *direct* and *indirect* exposure. Direct exposure comes from directly following the accounts of politicians, journalists, opinion leaders, and media organizations. Indirect exposure is mediated through social peers who retweet, quote, mention or reply to a tweet by these actors.

## 

## *Measuring Political Alignment of News, Politicians and Opinion Leaders*

Modeling the ideological leaning of news content and politicians is fundamental to assessing people’s online media diet, and different methods have been proposed for such characterization. Our analysis focuses on three aspects of citizens’ political exposure: (i) measuring exposure to right- and left-leaning MoC, (ii) measuring exposure to right- and left-leaning influencer account, and (iii) quantifying an alignment score that encapsulates the political affinity of news websites to be shared by Republicans or Democrats (i.e., URLs shared on Twitter linking news outlets outside the platform).

For politicians, we consider their party affiliation to be representative of their political leaning (excluding four Independents and one Libertarian). For influencers, their political leaning of influencers was derived based on the Members of Congress who follow them: if more Republican (Democrat) MoC followed an account then it is considered a Republican (Democrat) influencer.

For news websites, we learn a representation for each domain based on the co-sharing of domains. Then, we train a model that learns the association between domain representation and its audience alignment scores, i.e. the fraction of Republicans or Democrats who shared it. A comparison of alignment scores to the list obtained by Bakshy et al. [(2015)](https://www.zotero.org/google-docs/?M65e2m) shows a 0.82 Pearson correlation, which indicates a high correlation with an influential study in the field. Finally, to assess the ideological slant of each panel member, we averaged the ideological alignment scores of the domains they were exposed to during the studied period (See Appendix G for further detail). To address exposure to hyper-partisan media outlets, we identified political domains that were shared exclusively by one side of the political map (i.e., more than 90 percent of users who have shared these media outlets were either Democratic or Republican). Unlike measuring exposure to politicians and political influencers, in this case, we measured exposure to hyper-partisan media URLs rather than measuring exposure to hyper-partisan media accounts.

## 

## *Inferring Prototypical Types of Political Exposure*

We use state-of-the-art clustering methods to identify prototypical types of political exposure on Twitter. Our goal is to group people with similar political exposure characteristics and to derive the main modes of political consumption. The granular nature of our analysis, which takes into account direct and indirect exposure to multiple channels of information, leads to increased dimensionality of the panel members’ descriptive features (see Appendix G for a full list of the features used in the model).

Distance between nearest neighbors (i.e., users), which is essential for effective clustering, is not meaningful as the dimension of the problem becomes high, a phenomenon known as the Curse of Dimensionality in machine learning [(Aggarwal et al., 2002)](https://www.zotero.org/google-docs/?vJQsEn). Therefore, we use the machine learning approach of first reducing the dimensionality of the data and only then applying the clustering algorithm similar to other prior work [(Allaoui et al., 2020; Grootendorst, 2022)](https://www.zotero.org/google-docs/?BCJz7O). We use Uniform Manifold Approximation and Projection (UMAP) to reduce dimensionality, which outperforms other common dimension reduction techniques (e.g., PCA and t-SNE) in maintaining both global and local representations in high-dimensional data [(McInnes et al., 2020)](https://www.zotero.org/google-docs/?3FIPoL). Moreover, UMAP’s superior run-time performance and lack of computational restrictions on initial dimensionality, make it an ideal technique for analyzing our data, which consists of over 600,000 users, each described by a set of fifteen features.

For clustering, we use an algorithm that is well-suited to identifying islands of higher density of points (i.e., users with similar exposure characteristics) amid a sea of sparser noise. Specifically, we use the high density-based spatial clustering of applications with noise (HDBSCAN) [(McInnes et al., 2017)](https://www.zotero.org/google-docs/?OGCwz0). A unique feature of this clustering approach is that it is more robust to outliers, which is an important advantage when modeling a large and diverse sample of online activity such as our Twitter panel.

# 

# Results

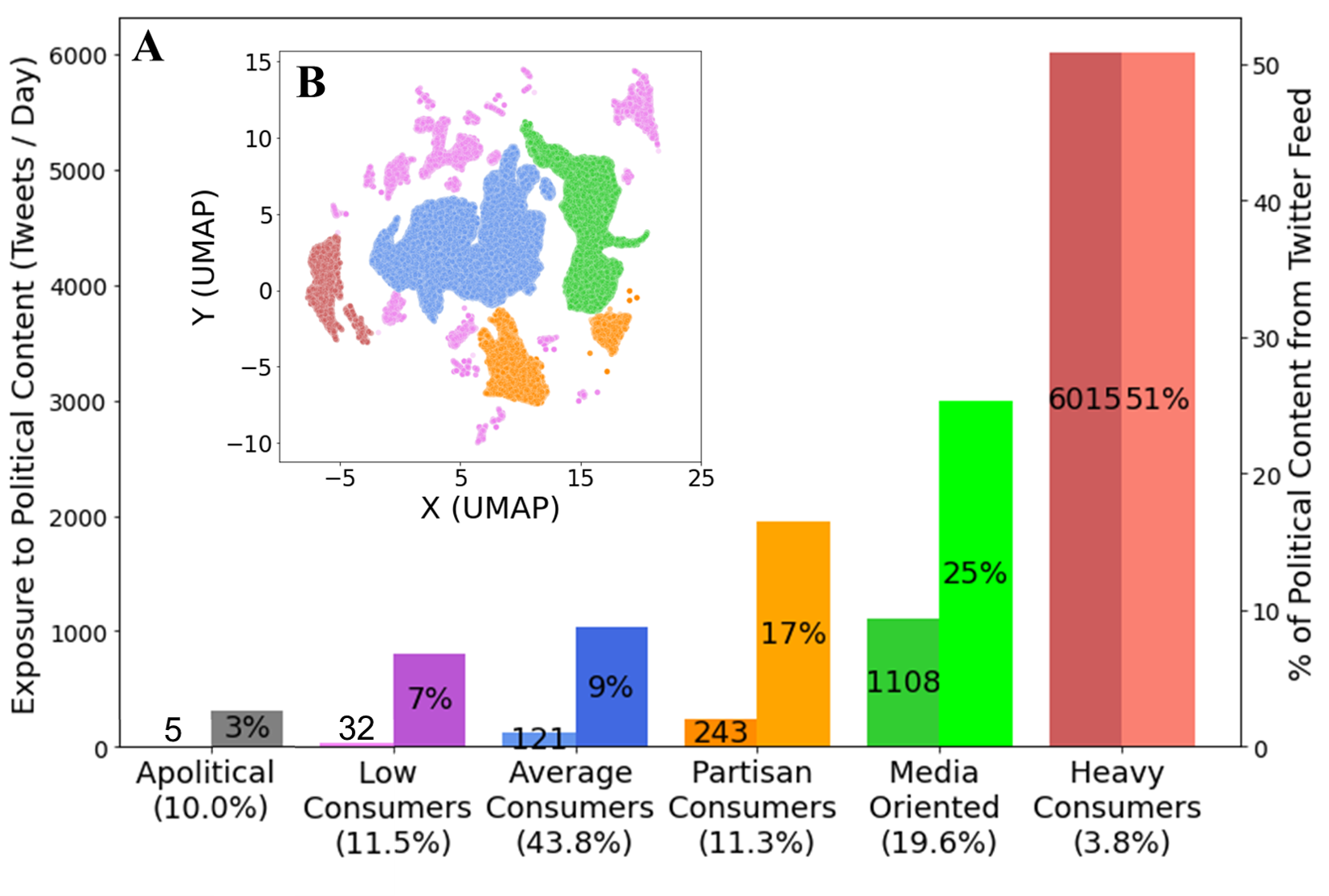
In this section, we report results regarding our two research questions: What are the prototypical types of political exposure on Twitter, and how does the distribution of these exposure types vary for distinctive socio-demographic groups?

In order to identify robust patterns of political exposure, we need to model users with a sufficient amount of political exposure. Therefore, all users that did not meet a minimum threshold were assigned to a separate cluster of ‘apolitical’ users. Consistent with prior research, we set this threshold to be one observed political tweet a day in the decahose (Grinberg et al. 2019), or a total of 122 observations throughout the entire election period. Based on this criteria, 10.0% of the population was directly assigned to the apolitical cluster, and we refrain from subdividing their consumption.

Applying clustering on the political exposure of all users with sufficient overall exposure to politics resulted in 28 clusters and 4.9% percent of outlier accounts. Twenty of the smallest clusters varied in some aspects of exposure, but had the common characteristic of having a relatively low level of political exposure. Therefore, we merged these small clusters into a single cluster with 11.5% of the population. The remaining clusters were further collapsed if they had similar levels of overall exposure to politics with a similar breakdown of actor types. Outliers were excluded from further analysis.

Figure 1 presents the prototypical types identified by clustering the political exposure of panel members after collapsing smaller clusters. Each point in Fig. 1B represents an individual and their political consumption at the reduced two-dimensional space computed by the UMAP algorithm with its color designating its cluster. Points that are closer together represent individuals that are similar to one another in some aspects of political exposure. Fig. 1A shows the median amount of political exposure in people's feeds[[2]](#footnote-2) and its share out of all content available to people on Twitter for each cluster separately. For example, the cluster referred to as 'heavy consumers' consists of 3.8% of the population. The median user in this cluster tweets every three days, and has over 6,000 political tweets available to them each day, which amounts to a little more than half of the overall tweets available to them. While we cannot definitively determine how many of those tweets are actually seen, no prior work has shown that nearly 4% of the population has social feeds that are *dominated* by politics.

***Figure 1****: Clustering of political exposure of individuals. Each point in panel (A) represents the political exposure of single panel member, reduced to two dimensions using the UMAP algorithm, and colored by the cluster assignment obtained from HDBSCAN. Panel (B) shows the median number of political tweets available to individuals per day (left, dark-colored bars), and their percent in all tweets available on Twitter (right, light-colored bars). The share of each cluster in the population is specified along with the cluster label on the x-axis.*



Moreover, one can see in Fig. 1B four distinct types of political consumers and a fifth cluster that is scattered in the periphery of the graph (colored in pink). The largest cluster, located in the center of the figure (in blue), consists of 43.8% of the population and we refer to it as 'average consumers' due to its size. Fig. 1A shows that the median person in this “average” cluster has about 120 political tweets available to them each day, which form about 9% of their Twitter Timeline. Three other clusters surround the main cluster and have a larger share of politics in their feeds: the heavy consumers cluster (shown in red), and two other clusters we refer to as 'media-oriented' (in green) and 'partisan' (in orange) that we describe in greater detail below. Finally, the cluster scattered throughout the map (in pink) consists of 11.5% of people and is characterized by having a relatively low level of political exposure that borders the apolitical group. Thus we refer to this cluster as 'low consumers' of politics.

In addition to the overall level of exposure to politics, the clusters we identified vary in the composition of political exposure from different "curators". Figure 2 shows the breakdown of political exposure by different actor types including influencers (in green), mainstream media (in yellow), active members from the 116th U.S. Congress (in red) and social peers (in blue). Lighter-colored bars indicate content from a peer that indirectly referred to MoC, influencer, or a media account by either retweeting, quoting, replying, or mentioning them. For example, the group of average consumers receives only 2% of their political diet directly from media organizations and 4.9% of indirect exposure to media organizations, which is almost 2.5 times the amount of direct exposure. In stark contrast, the heavy consumers group receives nearly 90% of their political exposure directly from news organizations. It is likely that heavy consumers simply follow many of the most active political news organizations on the platform.

***Figure 2****: Political diet composition across the five political clusters: the content is partitioned by (i) direct and indirect content and by (i) sources - political influencers (green shades), mainstream media (yellow shades), MoCs (red shades) and personal peer network (blue).*

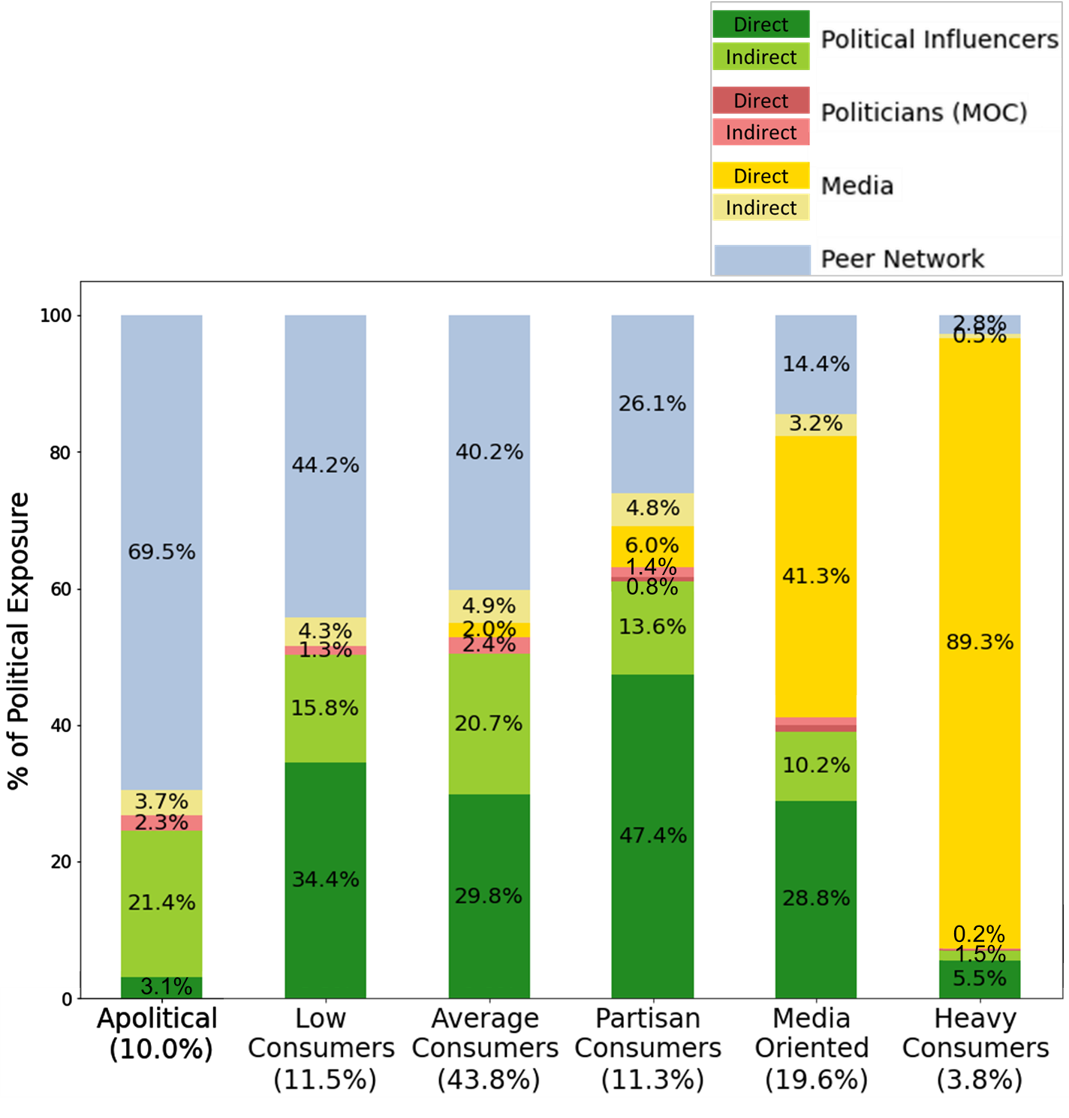


Figure 2 provides several key insights. First, low and average consumers, which together amount to about 55% of the population, receive hardly any political content directly from media organizations or national politicians, and most of their political exposure is provided by social peers and influencer accounts. This may reflect an active avoidance of news organizations and politicians, or that network formation is dominated by factors other than political interests. Either way, this large portion of the population is exposed to a non-negligible amount of political content in their feeds (7-9% according to Fig. 1A) and this exposure is almost exclusively curated by non-traditional gate-keepers and influencer accounts. Another interesting finding is that the group labeled as partisan consumers receives more than 60% of their political exposure from influencers (47.4% directly and 13.6% through peers). This suggests that influencers have the potential to play a much bigger role than previously thought in polarizing some people. Likewise, this finding indicates that influencers can be potentially more effective in delivering interventions that reduce polarization in comparison to curators like politicians or media outlets, due to their larger audience of followees. Finally, the media-oriented cluster, which consists of nearly 20% of the population, receives over 40% of its political exposure directly from mainstream media accounts.

Our modeling of political exposure will not be complete without investigating the political orientation of content and curators that populate one's political feed. To that end, we leveraged the average alignment scores we computed for panelists based on domains, influencers, and MoC in their feeds. We also visually examined the spatial distribution of alignment scores in the reduced two-dimensional space produced by UMAP. While some clusters had small areas of higher density of individuals with opposite alignment scores relative to their cluster, none of them was nearly as close to the sharp dichotomy observed in the two subclusters that make up our partisan cluster (hence its name).

Figure 3 focuses on the two subclusters that form the partisan cluster. The figure shows two measures of political leaning for each subset of the population: the average composition of political exposure from liberal and conservative influencers (left, solid bars) and from liberal or conservative politicians (right, light-colored bars). To the right of each set of bars, the density of news alignment scores is presented. Based on the alignment scores we labeled one subcluster of the partisan cluster as left-leaning and one subcluster as right-leaning. For example, one can see that for right-leaning partisans 82% of their influencers' content comes from conservative influencers, and that stands in stark contrast to the 4% observed for left-leaning partisans. To put these percentages in context, the figure also includes statistics for the sample as a whole, and averages for voters who registered with the Democratic or Republican parties. While left-leaning partisans show only a slightly more polarized consumption than the average registered Democrat, the subgroup of right-leaning partisans exhibits substantially more polarized consumption than the average registered Republican from both influences and Members of Congress. Similar asymmetries appear when examining subcluster exposure to hyper-partisan media by these subclusters.

***Figure 3****: Comparison of hyper-partisan clusters with similar political diet characteristics (in terms of consumption magnitude and sources’ composition); bars (left y-axis) represent percent of liberal and conservative MoCs and political influencers; density plots (right y-axis) represent alignment scores distribution; sample medians are displayed for reference.*

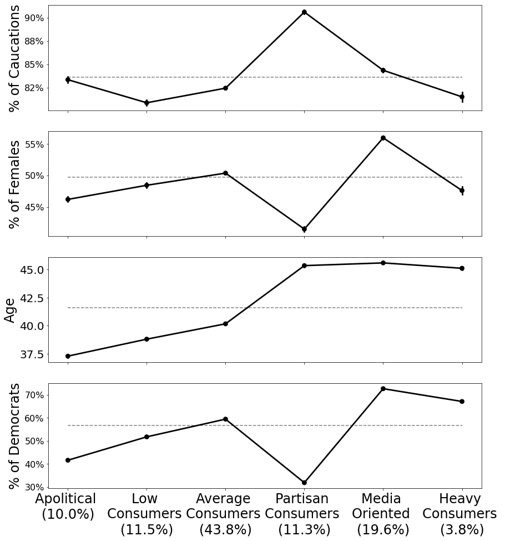


The findings in Figure 3 are consistent with Benkler et al.'s [(2018)](https://www.zotero.org/google-docs/?mbCtte) view of asymmetric polarization in American politics, but to the best of our knowledge, prior work mostly discussed the asymmetry in media consumption habits and in politicians being followed. Together with the observations from Figure 2, these findings on potential exposure of the partisan cluster identified in our results highlight the importance of further examining the causal effects of influencers and the political exposures they generate as a potential mechanism for affecting political attitudes and preferences such as left-right polarization, since this kind of political content exposure is much more prevalent for partisan consumers.

We now turn to our second research question, which focuses on how different socio-demographic groups engage with different types of political consumption. Figure 4 shows how ethnicity, gender, age, and party affiliation are distributed across the different exposure types. In particular, it shows the percentage of Caucasians, women, and Democratic panel members in each of the clusters. Age is presented using averages. The dashed horizontal line in each panel designates the sample average as a baseline for comparison.

Several key observations are relevant for assessing our second research question on socio-demographic distinctions in political content exposure patterns. First, there is a clear positive association between political content exposure and age, as the literature would predict (e.g., Verba, Schlozman and Brady 1995). While apolitical users and lower consumers of politics are below the average age, the clusters we label as partisan, media oriented and heavy consumers of politics are above the average age. A more detailed inspection shows that the youngest age cohort in our data (18-29) are indeed significantly less exposed to political content on Twitter, as they are overrepresented by 19.2% in the apolitical group in comparison to their representation in the sample.

***Figure 4:*** *Socio-Demographic characteristics among different political exposure modes. Sample averages are marked in a grey dashed line. Error bars represent 95% confidence intervals.*



Interestingly, breakdown by gender reveals a sizable group of women (56.0% in comparison to 49.8% in our sample) classified as media-oriented consumers, which are only second to heavy consumers in terms of both magnitude and interest in politics on Twitter, and constitute one in five registered voters in our sample.

From the panel members’ political affiliation data, excluding low and average political consumers, the remaining users (posing almost half of our sample; 44.7%) show a dichotomic characteristic based on the type of political consumer; that is, media-oriented and heavy political consumers show a larger representation of liberals, while the groups labeled as apolitical and partisan consumers of political content consist of a larger share of conservatives. Finally, results comparing how modes of political exposure relate to ethnicity, show over-representation of Caucasians as partisan consumers. However, this can be attributed to the overrepresentation of conservatives within partisan consumers (57.9% in comparison to a sample average of 31.2%), as conservatives within our sample have a higher percentage of Caucasians (94.4%) in comparison to the average user (83.6%).

# 

# Discussion

Much of the discussion about societal factors that may be contributing to democratic backsliding in advanced democracies – including rising populism, decreasing trust in media and political establishment, increased polarization, and misinformation – has been linked to the increased prevalence of digital media, and in particular, to social media. Social platforms are, indeed, widely adopted as a source of political information, and a primary source for many young adults. These trends in political content exposure call for better theoretical understanding of political exposure on these platforms including next-step causal examination of the impact of different types of political exposure on subsequent political attitudes and political behaviors. Robust analysis of these phenomena requires new computational methods for making valid inferences based on digital trace data that complement traditional methods, and allow us to go beyond previous research.

Grounded in the curated flows theoretical framework, this work contributes to the conceptualization of actors responsible for this curation, along with empirical findings that describe the types of actors that are responsible for political content distribution to registered U.S. voters on Twitter and the demographic characteristics of distinctive types of political consumers. We found that 55% of the population, including some of the lowest consumers of politics on the platform, still have considerable amounts of political content in their feeds during the election, which is primarily curated by social peers and influencer accounts. We also observe the dominance of influencer accounts in the political exposure of right-leaning partisan consumers. To the best of our knowledge, no prior work showed this increased prevalence of influencers in partisan consumption of politics.

Future work can leverage the observational findings of the current study to investigate the causal impact that influencers may have on people’s attitudes and behaviors, such as left-right ideological polarization and affective polarization.

Another important finding is that for almost a fifth of the population in our sample, media organizations are the largest source of political information and that they reach those voters directly without any mediation by peers. These findings contribute to the debate about the erosion of traditional gatekeepers, as the media organizations on our lists have, fundamentally, the same editorial processes that Kurt Lewin wrote about when he first introduced Gatekeeping theory [(Lewin, 1943)](https://www.zotero.org/google-docs/?DO3jGE). Our results show that a substantial proportion of modern consumers of political content on Twitter *choose* to replicate traditional gatekeeping in new media. Women are surprisingly over-represented in this group, and despite the documented left-leaning bias of the media on Twitter, about a third of the people in this group are registered Republicans. Future research could investigate the curation roles and impacts that these media-oriented individuals have on their local network and examine the role of media organizations in influencing women’s subsequent political attitudes and behaviors.

Along with these contributions, this research has a number of important limitations. First, it is unclear to what extent our findings will generalize to other social media platforms and to other populations. If additional data become available from Facebook, either through the Social Science One initiative or other means, future research could empirically tackle this question directly. The focus on registered voters may suggest the behaviors we observed on Twitter may generalize to other social platforms, but at the same time there are good reasons to believe that platform affordances may have a meaningful impact on our findings. Another key limitation is the focus of our analysis on content that is available to people and not necessarily the content that is actually seen by them. The difference between these two sets may be systematically biased by factors such as the time when individuals visit their Timeline, the duration of their visits, and the algorithmic content ranking conducted by social platforms. Moreover, since we relied on manually curated lists and verification for identifying distinctive curation actors (e.g., media organizations, opinion leaders), we cannot guarantee the comprehensiveness of lists. For example, the list of politicians does not include state and local politicians, which may have different levels of exposure and audiences.

There are also several avenues for future work to expand this research. In terms of theory, the curated flows framework puts much of its emphasis on the actor who is doing the curation. We believe that there is room to expand the theory to consider the producer of the content in addition to the person who curates it as it propagates through the network. Content attribution is also a major challenge that calls for methodological contributions. Furthermore, future research can examine how the different types of political content exposure are related to pro-democratic attitudinal measures known to be crucial for robust democratic functioning, such as political knowledge, and political efficacy. In addition, it is increasingly important to understand the relationship between the online and offline worlds, and examine how different types of political consumers engage in and mobilize to political action both online and offline.

# 

# References

[Aggarwal, C., Hinneburg, A., Keim, D., 2002. On the Surprising Behavior of Distance Metric in High-Dimensional Space. First Publ Database Theory ICDT 200 8th Int. Conf. Lond. UK January 4 - 6 2001 Jan Van Den Bussche Eds Berl. Springer 2001 Pp 420-434 Lect. Notes Comput. Sci. 1973.](https://www.zotero.org/google-docs/?FHjHwm)

[Alatas, V., Chandrasekhar, A.G., Mobius, M., Olken, B.A., Paladines, C., 2019. When celebrities speak: A nationwide twitter experiment promoting vaccination in Indonesia. National Bureau of Economic Research.](https://www.zotero.org/google-docs/?FHjHwm)

[Allaoui, M., Kherfi, M.L., Cheriet, A., 2020. Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. pp. 317–325.](https://www.zotero.org/google-docs/?FHjHwm)

[Allen, D., 2015. Reconceiving Public Spheres, in: From Voice to Influence: Understanding Citizenship in a Digital Age. University of Chicago Press, p. 179.](https://www.zotero.org/google-docs/?FHjHwm)

[Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.B.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A., 2018. Exposure to opposing views on social media can increase political polarization. Proc. Natl. Acad. Sci. 115, 9216–9221. https://doi.org/10.1073/pnas.1804840115](https://www.zotero.org/google-docs/?FHjHwm)

[Bakshy, E., Messing, S., Adamic, L.A., 2015. Exposure to ideologically diverse news and opinion on Facebook. Science 348, 1130–1132. https://doi.org/10.1126/science.aaa1160](https://www.zotero.org/google-docs/?FHjHwm)

[Barberá, P., 2015. How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the U.S. 46.](https://www.zotero.org/google-docs/?FHjHwm)

[Barzilai-Nahon, K., 2008. Toward a theory of network gatekeeping: A framework for exploring information control. J. Am. Soc. Inf. Sci. Technol. 59, 1493–1512. https://doi.org/10.1002/asi.20857](https://www.zotero.org/google-docs/?FHjHwm)

[Benkler, Y., Faris, R., Roberts, H., 2018. Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford Univ. Press.](https://www.zotero.org/google-docs/?FHjHwm)

[Berinsky, A.J., 2017. Measuring Public Opinion with Surveys. Annu. Rev. Polit. Sci. 20, 309–329. https://doi.org/10.1146/annurev-polisci-101513-113724](https://www.zotero.org/google-docs/?FHjHwm)

[Davis, J.L., Jurgenson, N., 2014. Context collapse: theorizing context collusions and collisions. Inf. Commun. Soc. 17, 476–485. https://doi.org/10.1080/1369118X.2014.888458](https://www.zotero.org/google-docs/?FHjHwm)

[de Vreese, C.H., Boomgaarden, H., 2006. News, Political Knowledge and Participation: The Differential Effects of News Media Exposure on Political Knowledge and Participation. Acta Polit. 41, 317–341. https://doi.org/10.1057/palgrave.ap.5500164](https://www.zotero.org/google-docs/?FHjHwm)

[Eady, G., Nagler, J., Guess, A., Zilinsky, J., Tucker, J.A., 2019. How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data. SAGE Open 9, 215824401983270. https://doi.org/10.1177/2158244019832705](https://www.zotero.org/google-docs/?FHjHwm)

[Flaxman, S., Goel, S., Rao, J.M., 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. Public Opin. Q. 80, 298–320.](https://www.zotero.org/google-docs/?FHjHwm)

[Fletcher, R., Nielsen, R.K., 2018. Are people incidentally exposed to news on social media? A comparative analysis. New Media Soc. 20, 2450–2468.](https://www.zotero.org/google-docs/?FHjHwm)

[Garrett, R.K., 2009. Echo chambers online?: Politically motivated selective exposure among Internet news users. J. Comput.-Mediat. Commun. 14, 265–285. https://doi.org/10.1111/j.1083-6101.2009.01440.x](https://www.zotero.org/google-docs/?FHjHwm)

[Garrett, R.K., Gvirsman, S.D., Johnson, B.K., Tsfati, Y., Neo, R., Dal, A., 2014. Implications of Pro- and Counterattitudinal Information Exposure for Affective Polarization: Partisan Media Exposure and Affective Polarization. Hum. Commun. Res. 40, 309–332. https://doi.org/10.1111/hcre.12028](https://www.zotero.org/google-docs/?FHjHwm)

[Graham, T., Jackson, D., Wright, S., 2015. From everyday conversation to political action: Talking austerity in online ‘third spaces.’ Eur. J. Commun. 30, 648–665. https://doi.org/10.1177/0267323115595529](https://www.zotero.org/google-docs/?FHjHwm)

[Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., Lazer, D., 2019. Fake news on Twitter during the 2016 U.S. presidential election. Science 363, 374. https://doi.org/10.1126/science.aau2706](https://www.zotero.org/google-docs/?FHjHwm)

[Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure.](https://www.zotero.org/google-docs/?FHjHwm)

[Guess, A., Munger, K., Nagler, J., Tucker, J., 2019. How accurate are survey responses on social media and politics? Polit. Commun. 36, 241–258.](https://www.zotero.org/google-docs/?FHjHwm)

[Guess, A.M., 2021. (Almost) Everything in Moderation: New Evidence on Americans’ Online Media Diets. Am. J. Polit. Sci. ajps.12589. https://doi.org/10.1111/ajps.12589](https://www.zotero.org/google-docs/?FHjHwm)

[Guess, A.M., 2015. Measure for Measure: An Experimental Test of Online Political Media Exposure. Polit. Anal. 23, 59–75. https://doi.org/10.1093/pan/mpu010](https://www.zotero.org/google-docs/?FHjHwm)

[Habermas, J., 1984. The theory of communicative action. Beacon press.](https://www.zotero.org/google-docs/?FHjHwm)

[Henderson, M., Jiang, K., Johnson, M., Porter, L., 2021. Measuring Twitter use: validating survey-based measures. Soc. Sci. Comput. Rev. 39, 1121–1141.](https://www.zotero.org/google-docs/?FHjHwm)

[Hindman, M., 2009. The myth of digital democracy. Princeton university press, Princeton (N.J.).](https://www.zotero.org/google-docs/?FHjHwm)

[Hughes, A.G., McCabe, S.D., Hobbs, W.R., Remy, E., Shah, S., Lazer, D.M.J., 2021. Using Administrative Records and Survey Data to Construct Samples of Tweeters and Tweets. Public Opin. Q. 85, 323–346. https://doi.org/10.1093/poq/nfab020](https://www.zotero.org/google-docs/?FHjHwm)

[Iyengar, S., Hahn, K.S., 2009. Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. J. Commun. 59, 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x](https://www.zotero.org/google-docs/?FHjHwm)

[Jackson, S.J., Bailey, M., Foucault Welles, B., 2018. # GirlsLikeUs: Trans advocacy and community building online. New Media Soc. 20, 1868–1888.](https://www.zotero.org/google-docs/?FHjHwm)

[Jackson, S.J., Welles, B.F., 2015. Hijacking #myNYPD: Social Media Dissent and Networked Counterpublics. J. Commun. 65, 932–952. https://doi.org/10.1111/jcom.12185](https://www.zotero.org/google-docs/?FHjHwm)

[Katz, E., Blumler, J.G., 1974. The uses of mass communications: Current perspectives on gratifications research.](https://www.zotero.org/google-docs/?FHjHwm)

[King, G., Persily, N., 2020. A new model for industry–academic partnerships. PS Polit. Sci. Polit. 53, 703–709.](https://www.zotero.org/google-docs/?FHjHwm)

[Lazarsfeld, P.F., Berelson, B., Hazel, G., 1944. The People’s Choice: How the Voter Makes Up His Mind in a Presidential Campaign. New York, NY: Columbia University Press.](https://www.zotero.org/google-docs/?FHjHwm)

[Levy, R., 2020. Social media, news consumption, and polarization: evidence from a field experiment. News Consum. Polariz. Evid. Field Exp. July 16 2020.](https://www.zotero.org/google-docs/?FHjHwm)

[Lewin, K., 1943. Forces behind food habits and methods of change. Bull. Natl. Res. Counc. 108, 35–65.](https://www.zotero.org/google-docs/?FHjHwm)

[McInnes, L., Healy, J., Astels, S., 2017. hdbscan: Hierarchical density based clustering. J. Open Source Softw. 2. https://doi.org/10.21105/joss.00205](https://www.zotero.org/google-docs/?FHjHwm)

[McInnes, L., Healy, J., Melville, J., 2020. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.](https://www.zotero.org/google-docs/?FHjHwm)

[Mellon, J., Prosser, C., 2017. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. Res. Polit. 4. https://doi.org/10.1177/2053168017720008](https://www.zotero.org/google-docs/?FHjHwm)

[Messing, S., Westwood, S.J., 2014. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. Commun. Res. 41, 1042–1063.](https://www.zotero.org/google-docs/?FHjHwm)

[Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. ArXiv Prepr. ArXiv13013781.](https://www.zotero.org/google-docs/?FHjHwm)

[Mitchell, A., Shearer, E., Stocking, G., 2021. News on Twitter: Consumed by Most Users and Trusted by Many. Pew Research Center.](https://www.zotero.org/google-docs/?FHjHwm)

[Morrow, G., Swire-Thompson, B., Polny, J.M., Kopec, M., Wihbey, J.P., 2021. The emerging science of content labeling: Contextualizing social media content moderation. J. Assoc. Inf. Sci. Technol.](https://www.zotero.org/google-docs/?FHjHwm)

[Mutz, D.C., 2001. Facilitating Communication across Lines of Political Difference: The Role of Mass Media. Am. Polit. Sci. Rev. 95, 97–114. https://doi.org/10.1017/S0003055401000223](https://www.zotero.org/google-docs/?FHjHwm)

[Nanz, A., Matthes, J., 2022. Democratic consequences of incidental exposure to political information: A meta-analysis. J. Commun. 72, 345–373.](https://www.zotero.org/google-docs/?FHjHwm)

[Odabaş, M., 2022. 10 facts about Americans and Twitter. Pew Res. Cent.](https://www.zotero.org/google-docs/?FHjHwm)

[Oser, J., Boulianne, S., 2020. Reinforcement Effects between Digital Media Use and Political Participation: A Meta-Analysis of Repeated-Wave Panel Data. Public Opin. Q.](https://www.zotero.org/google-docs/?FHjHwm)

[Oser, J., Hooghe, M., Marien, S., 2013. Is Online Participation Distinct from Offline Participation? A Latent Class Analysis of Participation Types and Their Stratification. Polit. Res. Q. 66, 91–101. https://doi.org/10.1177/1065912912436695](https://www.zotero.org/google-docs/?FHjHwm)

[Peterson, E., Goel, S., Iyengar, S., 2021. Partisan selective exposure in online news consumption: evidence from the 2016 presidential campaign. Polit. Sci. Res. Methods 9, 242–258. https://doi.org/10.1017/psrm.2019.55](https://www.zotero.org/google-docs/?FHjHwm)

[Prior, M., 2013. Media and Political Polarization. Annu. Rev. Polit. Sci. 16, 101–127. https://doi.org/10.1146/annurev-polisci-100711-135242](https://www.zotero.org/google-docs/?FHjHwm)

[Samii, C., 2016. Causal empiricism in quantitative research. J. Polit. 78, 941–955.](https://www.zotero.org/google-docs/?FHjHwm)

[Scharkow, M., 2016. The accuracy of self-reported internet use—A validation study using client log data. Commun. Methods Meas. 10, 13–27.](https://www.zotero.org/google-docs/?FHjHwm)

[Schlozman, K.L., Brady, H.E., Verba, S., 2018. Unequal and Unrepresented, in: Unequal and Unrepresented. Princeton University Press.](https://www.zotero.org/google-docs/?FHjHwm)

[Schlozman, K.L., Verba, S., Brady, H.E., 2012. The unheavenly chorus, in: The Unheavenly Chorus. Princeton University Press.](https://www.zotero.org/google-docs/?FHjHwm)

[Sears, D.O., Freedman, J.L., 1967. Selective exposure to information: A critical review. Public Opin. Q. 31, 194–213.](https://www.zotero.org/google-docs/?FHjHwm)

[Shearer, E., 2018. Social media outpaces print newspapers in the US as a news source. Pew Res. Cent. 10.](https://www.zotero.org/google-docs/?FHjHwm)

[Shmargad, Y., Klar, S., 2020. Sorting the News: How Ranking by Popularity Polarizes Our Politics. Polit. Commun. 37, 423–446. https://doi.org/10.1080/10584609.2020.1713267](https://www.zotero.org/google-docs/?FHjHwm)

[Shugars, S., Gitomer, A., McCabe, S., Gallagher, R.J., Joseph, K., Grinberg, N., Doroshenko, L., Foucault Welles, B., Lazer, D., 2021. Pandemics, Protests, and Publics: Demographic Activity and Engagement on Twitter in 2020. J. Quant. Descr. Digit. Media 1. https://doi.org/10.51685/jqd.2021.002](https://www.zotero.org/google-docs/?FHjHwm)

[Suuronen, A., Reinikainen, H., Borchers, N.S., Strandberg, K., 2021. When Social Media Influencers go Political: An Exploratory Analysis on the Emergence of Political Topics Among Finnish Influencers. Javn. - Public 1–17. https://doi.org/10.1080/13183222.2021.1983367](https://www.zotero.org/google-docs/?FHjHwm)

[Taylor, S.J., Muchnik, L., Kumar, M., Aral, S., 2022. Identity effects in social media. Nat. Hum. Behav. https://doi.org/10.1038/s41562-022-01459-8](https://www.zotero.org/google-docs/?FHjHwm)

[Thorson, K., Wells, C., 2016. Curated flows: A framework for mapping media exposure in the digital age. Commun. Theory 26, 309–328.](https://www.zotero.org/google-docs/?FHjHwm)

[Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J.T., Nagler, J., Tucker, J.A., 2015. Political expression and action on social media: Exploring the relationship between lower-and higher-threshold political activities among Twitter users in Italy. J. Comput.-Mediat. Commun. 20, 221–239.](https://www.zotero.org/google-docs/?FHjHwm)

[Verba, S., Schlozman, K.L., Brady, H.E., 1995. Voice and equality: Civic voluntarism in American politics. Harvard University Press.](https://www.zotero.org/google-docs/?FHjHwm)

[Weeks, B.E., Lane, D.S., Kim, D.H., Lee, S.S., Kwak, N., 2017. Incidental exposure, selective exposure, and political information sharing: Integrating online exposure patterns and expression on social media. J. Comput.-Mediat. Commun. 22, 363–379.](https://www.zotero.org/google-docs/?FHjHwm)

[Wojcieszak, M., Casas, A., Yu, X., Nagler, J., Tucker, J.A., 2022a. Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. Sci. Adv. 8, eabn9418. https://doi.org/10.1126/sciadv.abn9418](https://www.zotero.org/google-docs/?FHjHwm)

[Wojcieszak, M., Menchen-Trevino, E., Goncalves, J.F., Weeks, B., 2022b. Avenues to news and diverse news exposure online: Comparing direct navigation, social media, news aggregators, search queries, and article hyperlinks. Int. J. Press. 27, 860–886.](https://www.zotero.org/google-docs/?FHjHwm)

[Wojcik, S., Hughes, A., 2019. Sizing Up Twitter Users. Pew Res. Cent.](https://www.zotero.org/google-docs/?FHjHwm)

[Wrubel, L., Kerchner, D., 2020. 116th U.S. Congress Tweet Ids. https://doi.org/10.7910/DVN/MBOJNS](https://www.zotero.org/google-docs/?FHjHwm)

[Xenos, M., Vromen, A., Loader, B.D., 2014. The great equalizer? Patterns of social media use and youth political engagement in three advanced democracies. Inf. Commun. Soc. 17, 151–167.](https://www.zotero.org/google-docs/?FHjHwm)

[Zhuravskaya, E., Petrova, M., Enikolopov, R., 2020. Political Effects of the Internet and Social Media. Annu. Rev. Econ. 12, 415–438. https://doi.org/10.1146/annurev-economics-081919-050239](https://www.zotero.org/google-docs/?FHjHwm)

Online Appendix for

**How Registered U.S. Voters' Engage with Politics on Twitter**

**Content**

**Appendix A: Detection of Political Content**

**Appendix x: Mapping Political Elites Accounts on Twitter**

**Appendix C: 116th Members of Congress Twitter Accounts**

**Appendix D: Media Twitter Accounts**

**Appendix E: Opinion Leaders’ Twitter Accounts**

**Appendix F: Inferring Categories of Political Elites’ Accounts**

**Appendix G: Modeling Political Alignment**

**Appendix H: List of Political Exposure Feature**

**Appendix A: Detection of Political Content**

A critical component for our analysis includes identifying and separating political content (i.e., tweets) from other non-politically-related content, for which we have used the political classifier. A first step in creating the classifier was to attain labels for tweets that will be used to train the classifier. Labels were created based on a keyword list containing (i) general politically-related keywords, (ii) candidates' names and usernames on Twitter, and (iii) various hashtags which were specifically relevant for the 2020 U.S. Presidential election. With intention to minimize false positives in the keyword list, each tweet that contained one or more keywords from the list was labeled as political content (positive).

For training the classifier, all non-English tweets and tweets not containing URLs were filtered out. Then, labeling of the tweets was done based on the keywords list. From “negative” (non-political) tweets, we randomly sampled an equal number of tweets as “positives” (political) tweets, to establish a balanced dataset. The classifier was trained on daily data, to identify changes in political discussion over time. Predictions of the classifier were labeled as follows: tweets labeled as political by the classifier and tweets that contained one or more of the keywords (in the political keywords list) were labeled as political content, and non-political otherwise.

|  |
| --- |
| **General terms** |
| election, presdebate, vpdebate, democratic, gop, dnc, rnc, politics, political, voter, senate, senator, 2020election, election2020, electionday, votebymail, votersuppression, ballot, mailin, mail-in, mail in, russiahoax, qanon, obamagate, mailinballots, nakedballots, presidential, vote-by-mail, votingsquad, votethemout, wewillvote, blackvotesmatter. |
| **Elected officials and Candidates** |
| mike pence, michael pence, mikepence, michaelpence, pence, kamala harris, kamala, harris, spike cohen, angela walker, kamalaharris, senkamalaharris, joe biden, joebiden, biden, votebiden, bluewave2020, votebidenharris2020, ridinwithbiden, nomalarkey**,** biden2020, bidenharris2020, bidenforpresident, bidenkamala2020, joebiden2020, bidenharris, votehimout**,** Dumptrump, nevertrump, bluewave, fucktrump, bidenwarroom, voteblue, demconvention, votebluetosaveamerica, wakeupamerica, trumpisanationaldisgrace, trumpvirus, trumpisalaughingstock, traitortrump,jo jorgensen, jojorgensen, joanne marie jorgensen, jorgensen2020, beboldvotegold, donald trump, don trump, realdonaldtrump, donaldtrump, donaldjtrump, donald j trump,, trump, trumpwarroom, teamtrump, the donald, trump2020, maga, draintheswamp, keepamericagreat, neverbiden, trumppence2020, makeamericagreatagain, kag, presidenttrump, notmypresident, americafirst, redwave, votered, sleepy joe, sleepyjoe, hidenbiden, creepyjoebiden, bidenukrainescandal, rnc2020, kag2020, maga2020, trump2020landslide, tulsatrumprally, voteredtosaveamerica, trumpforpresident, backtheblue, howiehawkins, howie hawkins, howiehawkins2020, hawkins2020 |

*Table A1: Whitelist terms used for identifying political tweets with high probability*

**Political Classifier Validation**

For evaluating the classifier’s performance, we collected manual labels for 2065 tweets in our sample, stratified over the days which are included among the examined time period. All tweets that were selected for hand labeling were annotated by at least two workers on Amazon's Mechanical Turk. Where disagreements occurred, the authors agreed on a final label.

The annotators were asked to label the tweets according to the following categories:

(1) U.S. Presidential Election

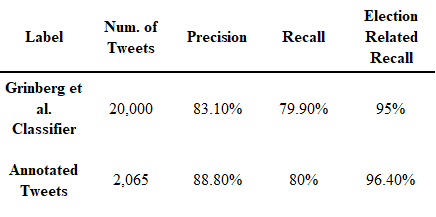
(2) U.S. Politics

(3) Non-U.S. politics

(4) Other

(5) I don't know

In the next step, we compared the classifier’s results to the manual annotations. “U.S. Presidential Election” and “U.S. Politics” were merged into a single category of “Politics”. Using the manual annotations as ground truth, we validated the classifier’s performance and reached precision of 88.8% and recall of 80.0%. Specifically for the “U.S. Presidential Election” category, the classifier demonstrated recall performance of 96.4%. These results are comparable to the classifier’s performance used by Grinberg et al. (2019) with a higher precision for the “Politics” category and a higher recall for the “U.S. Presidential Election” category.



*Table A2: Political classifier performance results*

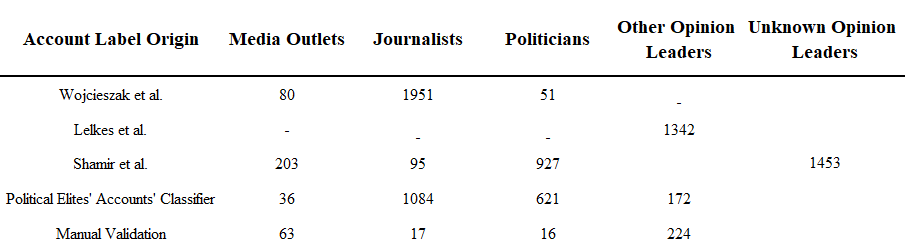
**Appendix B: Mapping Political Elites Accounts on Twitter**

Identifying the political elites’ Twitter accounts citizens are exposed to, provides the heart of the evidence in this analysis. The four main categories of political elites are defined as (i) news outlets, (ii) journalists, (iii) politicians and (iv) other opinion leaders (any opinion leader, as defined in Appendix E, whom do not belong to one of the three initial categories).

During the curation of the political elites accounts’ list we’ve used a combination of few sources and methods; we have initially started with identifying accounts of members of the 116th U.S. congress (see Appendix C). We then identified XXX general opinion leaders’ accounts (as detailed in Appendix x, by flagging accounts with a followership of more then 15 MoCs). To obtain some of these accounts underlying elites’ categories, and moreover to extend our list of political elites, we make use of political accounts lists produced by CITE MAGDA and CITE YPHTACH. These curated lists contribution is twofold; first (i) it directly extends our list of political elites’ accounts, and secondly (ii) it improve our accounts’ classifier (see Appendix F) performance (by adding training data), thereby indirectly enables a further extension of our list of political elites’ accounts.

After removal of XXX overlapping accounts between our initial collection of opinion leaders’ accounts list and the aforementioned valuable sources, XXX accounts remained without a label. These accounts were labeled to one of the four elite classes using our political elites’ accounts’ classifier, as detailed in Appendix F.

In table B1, we present a summary of the accounts used in this work and their originators.

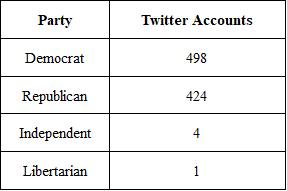
****

*Table B1: Number of political elites’ accounts used in this work by source; “Other Opinion Leaders” category include public figures, organizations and other accounts not political by nature (entertainement, sports and brands accounts); the remaining “Unknown Opinion Leaders” accounts were labeled as plainly “Other Opinion Leaders” accounts along the analysis*

**Appendix C: 116th Members of Congress Twitter Accounts**

To estimate content reaching citizens from active members of the 116th U.S. Congress (MoC), we have compiled a list of 942 Twitter personal and campaign accounts matching 533 representatives and 5 non-voting members from the 116th U.S. Congress, which convened on January 3, 2019 and ended on January 3, 2021, and thus presided through the examined period.

To compile this list, we first extracted a list of representatives from the official CONGRESS.GOV website and then merged it with a list of politicians’ Twitter account usernames published by [(Wrubel and Kerchner, 2020)](https://www.zotero.org/google-docs/?Sg1cfR). The accounts’ usernames were validated manually through the Twitter platform to match either a personal or a campaign account by the official list provided by CONGRESS.GOV. Finally, Twitter account IDs were extracted by using the Twitter API. The list contains 942 Twitter accounts IDs and screen names, matching 533 representatives from the 116th U.S. Congress and 5 non-voting members (one missing politician closed their Twitter account; the second missing politician was not found to be active during the examined period).



*Table B1: Number of Twitter accounts obtained 116th Members of U.S. Congress*

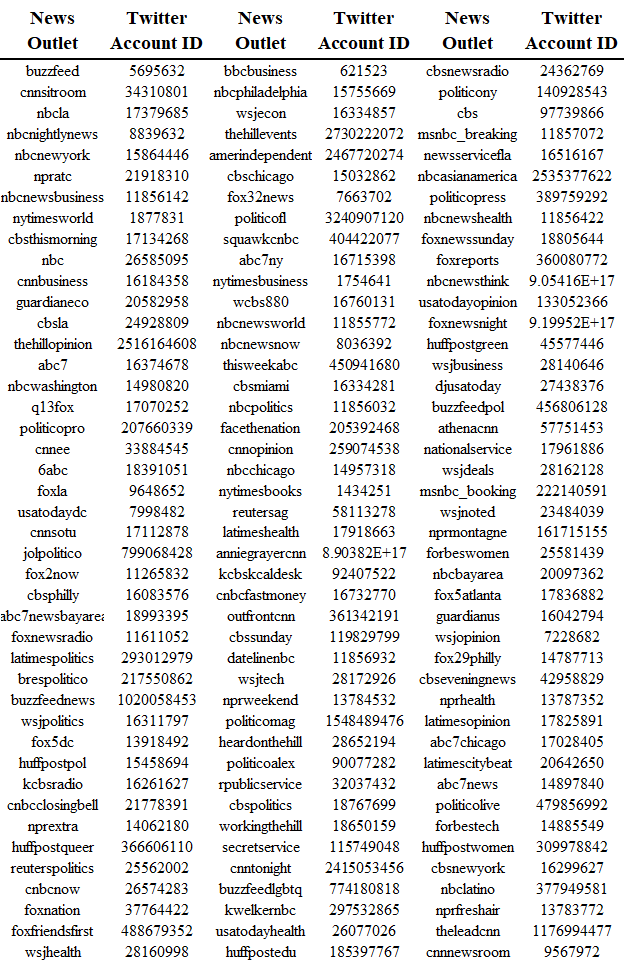
**Appendix D: Media Twitter Accounts**

The media list that was used to identify exposure to political news from media outlets included 76 Twitter accounts available in the following data repository: <https://github.com/sdmccabe/new-tweetscores>.



*Table D1: List of accounts used to detect media exposure; list in Table D2 extends it*

An additional 127 media accounts from the same news outlets were identified, from the opinion leaders original list (see Appendix E) which included any user with more than fifteen politician followers).



*Table D2: List of media Twitter accounts (multiple accounts for media outlets in Table D1)*

**Appendix E: Opinion Leaders’ Twitter Accounts**

To broaden our lens of political elites, we also consider the definition of *opinion leaders* as was used by Bail et al. [(2018)](https://www.zotero.org/google-docs/?9i4jRN), which includes accounts which are followed by at least 15 MoCs. To infer their political affiliation, we measured the average political affiliation of the MoCs who follow them, relying on it to signal the ideology of the political influencer. The list of “opinion leaders” also contained some of the MoCs in our list, which assisted in verifying the reliability of our political ideology inference method. As noted in the article, these were categorized strictly as politicians, and not as influencers.

To collect the list of followees of all MoCs (from which we derive the opinion leaders), we used the Twitter API. One limitation of the collection process is that we collect accounts after the examined time frame in this study, and therefore, (i) some accounts were deleted, suspended and changed to protected status leading to missing accounts, and moreover (ii) we cannot be certain that the connections collected for our study existed along the examined time frame.

After collecting a list of accounts followed by more than 15 MoCs, we needed to detect and remove media news accounts, as most likely these attract a large followership by active politicians. To do so, first we filtered out all media accounts used in our analysis (see Appendix D). Second, we created a list of keywords to detect media account names; the list contained “news” and “media” keywords, and media outlet names, which were extracted from the list of media accounts (i.e., *BBC* was extracted from *BBCWorld* account name), assuming news outlets maintained multiple active Twitter accounts. We identified 203 media accounts, and among them all 76 accounts from the media list. The remaining 127 accounts were multiple accounts of the same news outlets, and were manually validated. After removal of these accounts, we have finalized a list of 5119 political influencers.

**Appendix F: Inferring Categories of Political Elites’ Accounts**

xxx…

**Appendix G: Modeling Political Alignment**

Modeling the political news ideological alignment that people consume online is fundamental to our work. We used a three-step process to obtain such alignment scores: first, we used an embedding model based on domains’ sharing characteristics, and created a 100-dimension embedding vector that encapsulates the intricate characteristics of each media outlet. As a second step, we derived ideology alignment scores for the 17,901 popular domains (domains shared by over 30 panel members), by calculating the average of political ideology of users who have shared them. At the third step, we have used a Deep-Neural-Network to predict alignment scores for the remaining 20,622 websites, using the learned embedding vectors.

*Creating Media Outlets Embedding Vectors*

To create embedding vectors that will encapsulate the characteristics of news outlets, we model each user’s news outlets’ sharing history as a sentence, and each of the news outlets shared as a word. In this manner, we assume that there is a meaning to sharing sequences of news domains by an individual. Our approach is based on the Word2Vec model, a method from the world of natural-language-processing (NLP), which was introduced by [(Mikolov et al., 2013)](https://www.zotero.org/google-docs/?nOwhas) to learn word representations. After modeling the users’ sharing histories as sentences, we have used the Word2Vec API to insert them into a neural network (NN), and at each iteration the NN is introduced with a shared domain, and tries to predict its surrounding domains (i.e., the domains that were shared before and after the current domain). The end result is vectors of 100 dimensions for each of the domains we have modeled. To validate that our model training is not biased by either hyper-sharers or highly infrequent sharers (which can add significant noise to the model), we have removed (i) any user who has shared more than 100 domains per day during the examined period, (ii) the top 1% of sharers, and (iii) any users with under 10 shares during the examined period.

In the next step, we calculated ideological alignment scores, in a similar manner to [(Bakshy et al., 2015)](https://www.zotero.org/google-docs/?FM5ruF) by calculating the average political ideology of users who have shared them and normalize the scores between minus to plus one (i.e., if 75 out of 100 users who have shared a certain domain are Republican, this particular domain’s alignment score would be 0.5). These alignment scores arelater inserted into our model for establishing the main modes of political exposure on Twitter. The reasoning for calculating the domain embeddings is that alignment scores were calculated only for domains which were shared by more than 30 different users in our panel (to reduce noise). This criteria matched 17,901 domains, leaving 20,622 shared domains (by under 30 users) without a matching alignment score. To produce an alignment score for these domains, we trained a neural network to predict an alignment score out of an embedding vector (using data from the 17,901 domains which had both alignment scores and domain embeddings), and predicted alignment scores using the trained network for the remaining domains. The end result is a list of 38,523 web domains and their matching ideology alignment scores. A comparison of alignment scores to the list obtained by Bakshy et al. (2015) shows 0.82 Pearson correlation.

**Appendix H: List of Political Exposure Features**

A list of the full set of features used in the model for inferring the prototypical modes of political exposure is documented in Table F1. Each of the features were measured and averaged per panel member along the examined time period (August to November, inclusive, 2020), and relate to an individual’s Twitter feed.

|  |  |
| --- | --- |
| 1. | Number of political tweets to per day (Log2) |
| 2. | Fraction of political tweets from Twitter feed |
| 3. | Fraction of political tweets from political influencers (direct) |
| 4. | Fraction of political tweets from political influencers (indirect) |
| 5. | Fraction of political tweets from conservative political influencers (direct) |
| 6. | Fraction of political tweets from liberal political influencers (direct) |
| 7. | Fraction of political tweets from 116th Members of U.S. Congress (direct) |
| 8. | Fraction of political tweets from 116th Members of U.S. Congress (indirect) |
| 9. | Fraction of political tweets from Republican 116th Members of U.S. Congress (direct) |
| 10. | Fraction of political tweets from Democrat 116th Members of U.S. Congress (indirect) |
| 11. | Fraction of political tweets from media (direct) |
| 12. | Fraction of political tweets from media (indirect) |
| 13. | Fraction of political tweets from Left-leaning Hyper-Partisan websites |
| 14. | Fraction of political tweets from Right-leaning Hyper-Partisan websites |
| 15. | Average alignment score |

*Table F1: List of political exposure features used in our model*

1. List is available <https://github.com/sdmccabe/new-tweetscores> and in Appendix D. [↑](#footnote-ref-1)
2. The amounts are estimated by multiplying the amounts observed in the 10% random sample of Twitter content ("decahose") by ten. [↑](#footnote-ref-2)