Mapping the political landscape from data traces: multidimensional opinions of users, politicians and media outlets on X

Antoine Vendeville, Jimena Royo-Letelier, Duncan Cassells, Jean-Philippe Cointet, Maxime Crépel, Tim Faverjon, Théophile Lenoir, Béatrice Mazoyer, Benjamin Ooghe-Tabanou, Armin Pournaki, et al.

Mapping the political landscape from data traces: multidimensional opinions of users, politicians and media outlets on X

Antoine Vendeville[a,b,c,†], Jimena Royo-Letelier[a,†], Duncan Cassells[d,a,c], Jean-Philippe Cointet[a], Maxime Crépel[a], Tim Faverjon[a], Théophile Lenoir[a], Béatrice Mazoyer[a], Benjamin Ooghe-Tabanou[a], Armin Pournaki[e,f,a], Hiroki Yamashita[b,g], and Pedro Ramaciotti[b,a,c,*,†]

[a]Sciences Po médialab, Paris, France
[b]Complex Systems Institute of Paris Ile-de-France CNRS, Paris, France
[c]Learning Planet Institute, Learning Transitions Unit, CY Cergy Paris University, Paris, France
[d]Sorbonne Université, CNRS, LIP6, Paris, France
[e]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
[f]Laboratoire Lattice, École Normale Supérieure - PSL - CNRS - Univ. Sorbonne Nouvelle, Montrouge, France
[g]École des Hautes Études en Sciences Sociales, Paris, France

Abstract

Studying political activity on social media often requires defining and measuring political stances of users or content. Relevant examples include the study of opinion polarization, or the study of political diversity in online content diets. While many research designs rely on operationalizations best suited for the US setting, few allow addressing more general political systems, in which users and media outlets might exhibit stances on multiple ideology and issue dimensions, going beyond traditional Liberal-Conservative or Left-Right scales. To advance the study of more general online ecosystems, we present a dataset pertaining to a population of X/Twitter users, parliamentarians, and media outlets embedded in a political space spanned by dimensions measuring attitudes towards immigration, the EU, liberal values, elites and institutions, nationalism and the environment, in addition to left-right and liberal-conservative scales. We include indicators of individual activity and popularity: mean number of posts per day, number of followers, and number of followees. We provide several benchmarks validating the positions of these entities and discuss several applications for this dataset.

Keywords: X/Twitter, ideology scaling, dimensionality, polarization

1 Introduction

The study of phenomena related to public opinion online has become relevant to many fields, from social and political sciences[1–11], to computer science and complex systems[12–18]. However, because many pioneering studies originated in the US, researchers have traditionally produced more studies and data considering opinions as binary choices (e.g., Democrat- vs Republican-leaning), or as positions along single-dimensional political frames (Liberal-Conservative or Left-Right scales). This framework will not be suited to more general cases, for example countries where more political dimensions are required to analyze national politics[1,19–22].

Most landmark works have relied on large datasets of behavioral traces collected from online platforms or panels of users matched to additional data or surveys. X, in particular, has been extensively studied[23,24], due to the combination of a large active user base (including most of the prominent political figures in the Western world), a profusion of political content, and an accessible and well-documented API. However, research efforts are now impeded as the API was officially locked behind an expensive paywall in February 2023. Data is not easily accessible either on other major platforms such as Reddit or Facebook, while smaller platforms are often restricted to fringe user bases[25–27] or do not propose well-documented APIs[28,29]. Research efforts can be supported through collaboration with the companies themselves[13,30–33], but this approach raises concerns related to independence and reproducibility[34,35]. Finally, despite the large quantity of empirical social media research, few researchers release publicly the datasets they collect.

We address these limitations with the release of a large-scale, high-granularity and anonymized dataset containing multi-dimensional opinions of $N = 978,933$ users in the French political sphere on X, including all of the 883 French Members of Parliament who had accounts by February 2023. The political opinions of users in our dataset are provided in the form of real numbers that indi-

cate individual positions on $d = 16$ ideological scales and issue dimensions, including the Left-Right axis, anti-elite sentiment, attitudes towards nationalism, immigration, E.U. integration, and environmental policies. Using links pointing towards press articles shared by the users, we also provide positions of 400 popular French websites, including most relevant media outlets, blogs, and forums. Finally, we include individual indicators of activity and popularity for the users, as well as indicators of popularity for the media outlets.

To derive the opinions of users, we applied the methodology proposed by Ramaciotti et al. (2022)[36]. This methodology first performs a multidimensional ideology scaling[37] on a bipartite follower network linking MPs with their followers[38]. This network was built via a comprehensive collection of follow links in February 2023. Following the method by Ramaciotti et al. (2022)[36], we then selected a political survey dataset to calibrate and identify the multidimensional latent positions resulting from the ideology scaling procedure. For this, we selected the 2019 and 2023 Chapel Hill Expert Surveys (CHES)[39,40] to map the output of the multidimensional ideology scaling onto the dimensions or scales of the CHES datasets. This second step lets us overcome the traditional identification problem of ideology scaling, which embeds users in latent spaces that do not necessarily correspond to actual ideological or political axes. The first dimension of such spaces is often a good proxy of the most salient line of division in national politics, but it can fail to produce readily usable political positions in national settings that display several salient and independent lines of division.

The methodology we use is solely based on follow relationships, to the exclusion of textual data. To validate the contents of our dataset we look at textual self-descriptions provided by the users on their X profile bios, which are independent from data used for the inference of positions. We annotated these profiles according to the political stances that they express, both with human annotators and with generative AI annotation protocols. These annotations used in validation are also provided in the dataset. The positions of the web domains that we include in the dataset (computed on the bases of the positions of users that shared posts with URL from these domains) are in good agreement with a categorical classification of the most important French media on the Left-Right dimension by previous research[41,42]. The use of political positions of media outlets is well-known to the scientific literature in many fields, and the dataset here presented enables researchers to conduct many of these investigations considering now several relevant ideology and issue dimensions, beyond Left-Right or Liberal-Conservative scales.

Opinion inference based on social media data counts several and diverse methods. Other relevant publicly available datasets include: a French X dataset collected during the 2017 Presidential Elections, comprising retweet and mention counts between 20K users manually labeled with their preferred political party[43,44]; an American dataset of popular web domains and political leanings of Facebook users sharing them in 2014[13,45]; a dataset comprising survey responses and web-browsing data collected from voluntary participants in several countries in 2021-2022[46,47]; a dataset comprising over 200K messages posted in Indian Whatsapp groups during the 2019 General Elections, with 3.8K of these messages manually labeled to indicate support or disapproval of political parties[48,49]; a dataset of American political blogs and hyperlinks between one another, collected during the 2004 Presidential Elections[50,51].

In comparison with most previously available datasets for studying phenomena related to political opinions in online settings, our main contribution is the release of the first public dataset with multidimensional and continuous opinion estimates, alongside indicators of activity and popularity, for almost a million X users, parliamentarians and hundreds of media domains.

Next, we present the methodology for the construction of the dataset, a detailed description of the data records and files included in the dataset, validation benchmarks for the ideological and issue positions provided by the dataset, and, finally, a brief discussion of limitations and the data and code availability statements.

2 Methodology

Abundant evidence suggests that the choice of which political actors to follow is an informative signal of ideology[36,38,52–56]. The methodology we use for opinion inference relies on the follow relationships that exist between Members of Parliament (MPs) and the general public on X, relying on the existence of political homophily on the part of users choosing to follow MPs. We start by describing the data collection process. Then, we detail the opinion inference methodology, which unfolds in two steps. Next, we explain how we derived individual indicators of activity and popularity. Finally, we explain how we derived positions for a selection of French media domains on ideology and issue dimensions.

2.1 Collection of the MP-followers network

The data collection process reproduces that done by Ramaciotti et al., (2022), in which they used data collected in 2019. In February 2023, we identified the X profiles of all 886 French MPs present on the platform (out of 925 in the French parliament), and collected all their followers using the software `minet 1.00.0-a15`[57]. Both chambers of the parliament are concerned, the lower (*Assemblée*

[1]MPs who resigned include mostly those who accepted a position in the government after the elections. Resigned MPs and their substitutes amount to $N = 46$ of the MPs present in the data.

nationale) and the upper (*Sénat*) houses. We include MPs who resigned from their mandate before December 2022, as well as their substitutes[1]. Their followers were then filtered based on the criteria from Barberá[38], keeping only those who followed at least 3 MPs. This ensures that the users left in our dataset have a sufficient knowledge of national politics. Individuals that are knowledgeable about national politics are more accurately represented in their political opinion by spatial models, a property related to *political sophistication*[58].

Barberá[38] also suggests removing users with less than 25 followers, to filter out bots and inactive accounts. To allow for fine-grained studies of the relationships between popularity and opinions we keep these accounts in the data, and propose an alternative dataset with those users excluded (see data availability). We show in the Appendix that keeping these accounts did not alter significantly the computed political positions.

We remove MPs who are left without followers after this filtering step. We end up with $M = 883$ MPs and $N = 978,206$ followers. From there, we build a directed bipartite MP-follower network, yielding a binary adjacency matrix $A \in \{0,1\}^{N\times M}$. Importantly, follower-follower and MP-MP connections are ignored. The network contains 9,601,175 edges, MPs have average in-degree 10,910 and followers have average out-degree 10.

Similar studies performing latent ideological inference often rely on retweet networks[59,60]. While retweet links may also provide a source of homophilic choice data, they could also lead to biased results due to agenda setting and the dynamics of issue saliency over time. The database of tweets that we rely on for the computation of media positions, covers in particular the period of the invasion of Ukraine by Russia, and the widely contested pension reform in France. Such significant political event have the capability to strongly restructure the political discussion online, resulting in a biased representation of the overall political space. Using follow links on the other hand, ensures that we are capturing a stable political space which we expect to underpin the general dynamics of the public debate in France. Additionally, our collection of follow links is exhaustive, and therefore does not suffer from potential issues due to data sampling. These issues could arise when retrieving individual tweets and retweets from the now defunct Twitter API[61–63].

2.2 Inference of latent ideological positions

As Ramaciotti et al., (2022) describe in their method, we first apply the ideology scaling method originally proposed by Barberá[38] for extending the use of ideology scaling to social media data. It takes as input a bipartite network of connections between users and MPs, and outputs positions in a latent homophily space in which MPs that are close in space are followed by similar sets of users, and, conversely, in which users that are close in space follow a similar set of MPs. The method assumes that the observation (i.e., the choices of whom to follow) obeys a probabilistic homophily law in which followers decide to follow MPs on the bases of unobservable positions on some latent space:

$$\begin{aligned}\text{Prob}(A_{ij} = 1|\alpha_i, \beta_j, \gamma, \phi_i, \phi_j) = \\ \text{logit}^{-1}(\alpha_i + \beta_j - \gamma\|\phi_i - \phi_j\|^2),\end{aligned} \tag{1}$$

where $A_{ij} = 1$ when user i follows user j, α_i and β_j quantify respectively the tendency to follow others and the tendency to be followed of users i and j, and γ is a shape parameter that controls the relative weight between homophilic (ϕ_i and ϕ_j) and idiosyncratic parameters (α_i and β_j). The network of connections is known, and the goal is to infer the unknown latent ideological positions ϕ that best explain this network. These positions will then reflect ideological similarities and differences between the users.

Using a Markov Chain Monte Carlo (MCMC) estimation procedure on the network of U.S. parliamentarians and their X followers, Barberà[38] showed that the latent positions ϕ stand as good indicators of ideological positions. It has been later shown that the inference of the parameters of the equation can be approximated via a Correspondence Analysis (CA)[64]. CA is a dimensionality reduction procedure for categorical data, which has been shown to be a fast and efficient approximation method for the computation of ϕ[38,65,66].

We perform a CA on the adjacency matrix A to obtain individual ideological positions in a latent space. The dimensionality of this latent space can be fixed between 1 and 883. Because we will map positions from this latent space to the space subtended by the dimensions of the political survey we use for calibration—the Chapel Hill Expert Survey data (CHES), see below—and because this map is computed with the positions of political parties on both spaces, the choice of the number of dimensions for the latent space depends on the number of available political parties on both spaces. Our 883 MPs embedded in the latent space can be identified with 11 parties from the CHES data. We compute proxy multidimensional positions of each party in the latent space as the average position of MPs from that party. We remark that positions along dimensions of this latent space may be readily linked to positions on ideologies and issues, but this cannot be assured because of the lack of identification for values ϕ in the probabilistic equation (in particular with respect to isometries such as translations and rotations).

2.3 Mapping latent positions onto ideology and issue dimensions from political surveys

Following the method described by Ramaciotti et al. (2022), we map these latent multidimensional positions onto the ideology and issue dimensions of a political survey instrument used as reference. We use two political survey instruments administered around the time of our

data collection: the 2019[39] and the 2023[40] CHES data. In these surveys, political scientists estimate the positions of the main political parties of multiple European countries on multiple ideology and issue dimensions (e.g. placement on the Left-Right scale, or attitudes towards immigration policy), on a continuous scale from 0 to 10[2]. Crucially, positions on these dimensions are endowed with spatial reference frames. For instance, on the Left-Right dimension, 0 stands for the leftmost position for political parties, 10 for the rightmost, and 5 is the political center. While CHES 2019 is older and less consistent with the date of collection of our data, it contains a much larger set of dimensions (51 versus 11 for CHES 2023), and both waves include 4 dimensions in common (Left-Right, GAL-TAN, EU, and Anti-elite), which we also use as validation (see Fig. 4). Our approach does not seek to infer positions for our dataset on all available CHES dimensions, and we focus instead on those for which we can propose validation tests (see Table 1).

As described in the work detailing the positioning method[36], we compute an affine transformation mapping between our latent space and the space of political dimensions spanned by CHES. This transformation is fitted by using positions of political parties in both the latent space and the survey dimensions. Party positions come readily available in the CHES datasets. Following Ramaciotti et al. (2022), we compute party positions in the latent space as the mean position of MPs from each party. For each CHES dimension of interest d_c, we fit an affine transformation between the multidimensional positions of political parties in the latent space onto the party positions along d_c. We use a Ridge regression[67], with penalty parameter $\alpha = 1.0$. We restrain the number dimensions of the departure space to $P-1$, with P being the number of political parties that exist both in the latent space and in the survey dataset. This value, $P-1$, is the number of dimensions that would fully determine the affine transformation under no regularization[36]. From the parties of the MPs that were manually annotated, 11 were present in the CHES 2023 data, (resp. 9 in CHES 2019), leading to fitting affine transformations using the first 10 and 8 dimensions of the latent space, respectively. See the Appendix and Table 9 for a more detailed description of the parties.

Once the affine transformations are fitted this way, we apply them to the latent positions of each follower and MP. This way, we obtain coordinates for all followers and MPs in the political space spanned by the chosen CHES dimensions. Due to the regularization we apply, the positions of the political parties in our data—computed as the average position of the relevant MPs—may differ slightly from their positions in CHES. This deviation is small in the sense that the average Pearson correlation between party position computed with the affine transformation and given by the CHES data is 0.94. The mean absolute difference in party positions is 0.63 (for reference, CHES dimensions span from 0 to 10).

2.4 Activity and popularity

We include in this dataset metrics of popularity and activity for users positioned in our multidimensional ideology and issue space. These metrics are obtained or calculated from the user metadata obtained during the collection process using the platform's API. They comprise the mean number of posts per day, the number of followers, and the number of friends (also known as followees or followed accounts), at collection date. We compute for each user the mean number of posts per day, dividing their total number of posts by the number of days elapsed between account creation and our date of collection.

2.5 Media sharing collection

To compute positions of media outlets, we use a database of tweets collected during 18 months between 1 January 2022 and 30 June 2023 via the search query `lang:fr filter:links`. The database contains all tweets published in French during that period which contained at least one URL. When a tweet was part of a thread we also collected the entire thread, including the original post. For the collection, we used the Gazouilloire tool version 1.2.0[68] running query calls to X's search API v1.1. We aggregate this way an average of 2.6 million tweets each day and a total of 1.4 billion tweets over the whole period[3].

We filter this database of tweets and keep only original posts produced by the users for which we estimated a political position in Section 2.3, and containing URLs towards a domain belonging to the French media ecosystem. This design choice achieves two objectives. Filtering out quotes, we capture how individuals with a given issue and ideological position propose URL sources, without this source being leveraged or elicited by the context of a discussion unfolding on a thread. Filtering out retweet or shares, we capture behavior in which individuals have personally curated (and thus read with higher probability), instead of effortlessly sharing a post that may have contained a URL source (which they may have not read). We define this ecosystem as a collection of 747 domains, taken from ref. [69] and corresponding to major French media sources—newspapers, regional press, radio and TV channels, blogs and other digital information outlets. The methodology for selection of the domains is detailed in ref. [70]. Filtering the collection of tweets this way, we obtain a collection of 3,429,848 tweets by

[2]Dimensions pertaining to E.U. integration scale from 1 to 7, but we rescale them to $[0-10]$ for the sake of consistency.

[3]Data collected live was lost for the months of January and February 2023 and had therefore to be recollected a few months later, resulting in a little less tweets per day over those two months due to the known attrition of posts either deleted by X or their authors themselves.

Table 1: List of dimensions from the Chapel Hill Expert Survey (CHES) that we consider in the method to calibrate the political positions of X users, with statistics summarizing these positions. The reference points for dimensions are between values 0 and 10 (the extreme positions for political parties in the CHES data), while individuals can be positioned outside these bounds (outliers).

CHES dimension	Description	CHES wave	Mean	Std.	% outliers
lrgen	Left - Right	2019	6.308	2.293	3.078
corrupt_salience	Importance of reducing political corruption	2019	4.708	0.691	0.000
people_vs_elite	Opposes - Favors (direct democracy)	2019	4.802	0.995	0.023
immigrate_policy	Favors - Opposes (immigration)	2019	6.861	1.819	4.008
sociallifestyle	Favors - Opposes (liberal policies)	2019	5.494	2.185	0.592
nationalism	Cosmopolitanism - Nationalism	2019	6.355	2.030	5.518
antielite_salience	Anti-elite sentiment	2023	6.771	1.875	2.081
eu_position	Anti EU - Pro EU	2023	4.832	1.978	0.923
lrecon	Left - Right (economy)	2023	5.445	1.320	0.033
refugees	Opposes - Favors (welcoming Ukrainian refugees)	2023	5.822	1.820	0.407
galtan	Liberal - Conservative	2023	6.004	1.902	0.209
environment	Favors - Opposes (environment over economy)	2019	5.645	1.073	0.190
lrecon	Left - Right (economy)	2019	5.350	1.958	0.049
antielite_salience	Anti-elite sentiment	2019	7.040	1.880	3.320
eu_position	Anti EU - Pro EU	2019	4.705	2.275	0.540
galtan	Liberal - Conservative	2019	5.988	1.584	0.026

71,374 followers. There were 747 domains identified in the original study, of which 692 we find in at least one tweet from our users. To guarantee a good signal quality, we discard media domains whose articles were not tweeted by at least 100 different users. We end up with 400 media domains from which 3,272,574 URLs were shared by 70,604 followers (7.2% of all followers collected before). For each media domain, we provide the number of posts containing an URL from the domain, and the number of users in our collection who emitted these posts. Under this operationalization, the affinity between author and cited source may be of recognition of authority and relevance (i.e., the author deems the source worthy of mention, even if it is to critice the source or the content), ideology and issue affinity, or a combination of both. In the Validation section, we disentangle these possibilities by measuring the issue and ideology regularity of the authors of tweets including URLs from a given source. We also show that the media positions we derive are strongly correlated with the results of an independent study positioning French media outlets on the Left-Right axis.

2.6 Computing positions of media domains on survey dimensions

For each media domain, and for each one of the selected issue and ideology dimensions from the CHES data, we compute the position of a media in a dimension as the mean position of the followers that cited URLs from the domain. We do not take into account the number of tweets per user but solely a binary variable that takes value 1 if the user ever tweeted a link to a webpage from the domain, and 0 otherwise. This design choice is motivated by the fact that polarized users tend to be more active on social media. We seek to capture who finds a source worthy of mention, and not how many times. Operationalized this way, the computed positions must be interpreted as those of the public that deems the corresponding media worthy of citing. These positions do not necessarily represent the positions of contributors, journalists and editors of the media, which can display considerable heterogeneity. These data on media positions are useful as a measure of attention and authority granted by individuals. For studies that would want to leverage these positions for studies related to media bias, we also compute and provide in the data standard deviations and perform Hartigan's dip test[71] to evaluate the spread and the potential bimodality of the opinion distributions across users who share each domain. Using these additional metrics, researchers can select medias that are highly cited by users with coherent positions in our selected issue and ideology dimensions.

Finally, to give an example illustration of the data we release, Fig. 1 shows the positions of MPs, parties and three major media domains along four selected dimensions.

3 Data records

The released files contain seven tables in CSV format, one Jupyter notebook and several images. The CSV files are the following:

- `mps_positions.csv`
- `followers_positions.csv`
- `followers_human_annotations.csv`
- `followers_llm_annotations.csv`
- `mps_activity.csv`
- `followers_activity.csv`
- `domains_positions.csv`

We detail below the size and content of these tables. Most of them exhibit sets of columns pertaining to similar features declined across the multiple political dimensions and annotation labels present in our data ; we do not list

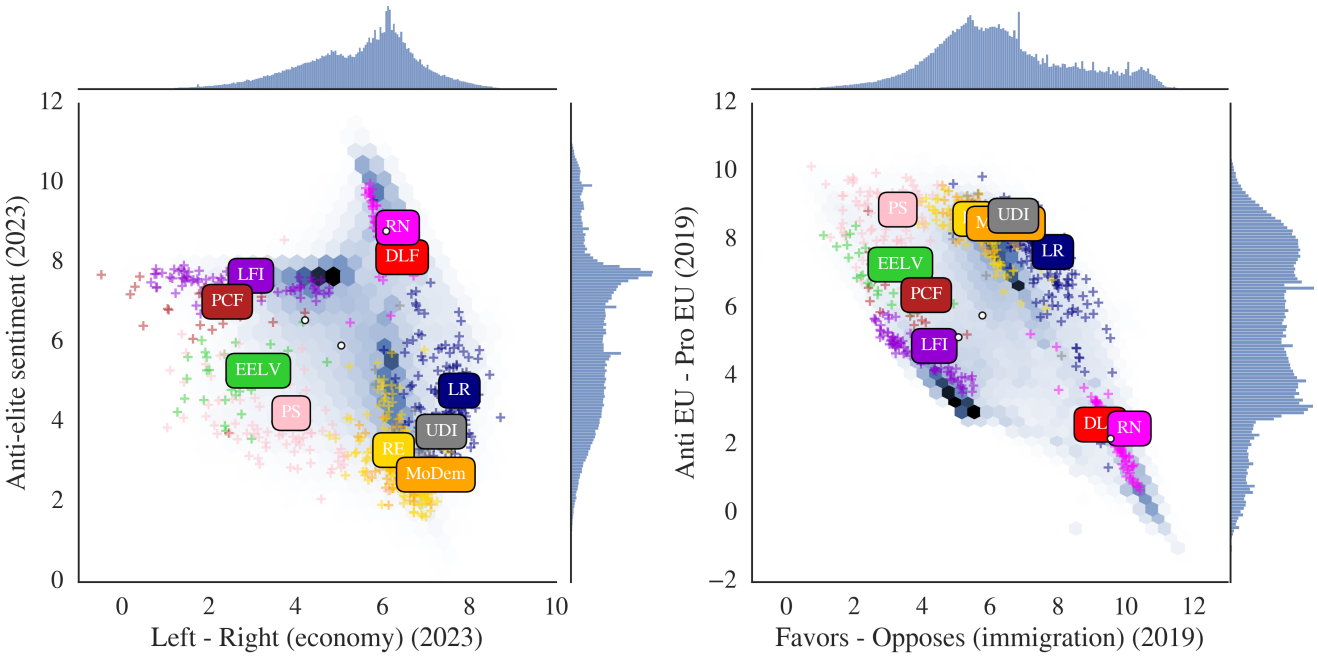

Figure 1: Positioning of users, MPs, parties and three selected media domains on four political axes: economic Left-Right and anti-elite sentiment (left), European integration and ecology (right). Colored crosses represent MPs. Colored rectangles indicate the average position of each party's parliamentarians. Marginal densities and hexagons indicate the distribution of followers positions. White dots indicate the positions of three selected media domains.

all such columns in the tables below but give a representative example and the total number of such columns each time. All floats are given with three decimals, except for mean_tweets_per_day (five decimals). Names of MPs are included, but other users have been anonymized and all information allowing for potential identification has been removed. For the sake of practicality, we endowed each user and MP with a unique random *pseudo id* that allows for matching data across the different tables.

3.1 Positions of MPs

The file `mps_positions.csv` contains the names, party affiliations and political positions of MPs. It has 883 rows and 19 columns. Each row corresponds to one MP that had a platform account at the time of our manual annotation procedure. The columns are detailed in Table 2.

3.2 Positions of followers of MPs

The file `followers_positions.csv` contains the political positions of the followers. It has 978,050 rows and 17 columns. The columns are the same as for `mps_positions.csv` minus the party and name columns. Each row corresponds to one user.

3.3 Human annotations of followers bios

The file `followers_human_annotations.csv` contains human annotations of the bios of followers, used for the validation. It has 5,864 rows and 9 columns—users with all NaN annotations were discarded. Each row corresponds to one follower. These annotations correspond to labels produced by a set of human annotators, prompted to read profile bios of users in our dataset, and annotate whether they display information allowing us to classify them into groups relevant for our validation metrics. The columns are detailed in Table 3.

3.4 LLM annotations of profile bios of followers

The file `followers_llm_annotations.csv` contains annotations of the followers bios produced by LLMs and used for the validation. It has 317,684 rows and 16 columns—users with all NaN annotations were discarded. Each row corresponds to one follower. The columns are detailed in Table 4.

3.5 Activity and popularity of MPs

The file `mps_activity.csv` contains the names, party affiliations, activity-related and popularity-related metrics for the MPs. It has 883 rows and 6 columns. Each row corresponds to one MP. All MPs are included. The columns are detailed in Table 5.

3.6 Activity and popularity of followers

The file `followers_activity.csv` contains activity and popularity metrics for the followers. It has 978,050 rows and 4 columns. Each row corresponds to one follower.

Table 2: Detail of columns for `mps_positions.csv`.

Column name	Description	Format
pseudo_id	Unique (artificial) identifier of the user.	string
name	First name and surname of the MP.	string
party	Party affiliation of the MP.	string
xxx_yy (e.g. lrecon_23)	Position on dimension xxx in the space of CHES 20yy. 16 such columns.	float

Table 3: Detail of columns for `followers_human_annotations.csv`.

Column name	Description	Format
pseudo_id	Unique (artificial) identifier of the user.	string
xxx (e.g. eurosceptic)	Human annotations of the bios for class xxx. 8 such columns.	1.0 (identified as xxx), nan (unidentified).

Table 4: Detail of columns for `followers_llm_annotations.csv`.

Column name	Description	Format
pseudo_id	Unique (artificial) identifier of the user.	string
xxx	LLM annotations of the bios for class xxx. 15 such columns.	0.0 (identified as not xxx), 1.0 (identified as xxx), nan (unidentified).

Table 5: Detail of columns for `mps_activity.csv`.

Column name	Description	Format
pseudo_id	Unique (artificial) identifier of the user.	string
name	First name and surname of the MP.	string
party	Party affiliation of the MP.	string
mean_tweets_per_day	Average number of posts per day.	float
followers	Number of followers.	float
followees	Number of followees.	float

Table 6: Detail of columns for `followers_activity.csv`.

Column name	Description	Format
pseudo_id	Unique (artificial) identifier of the user.	string
mean_tweets_per_day	Average number of posts per day.	float
followers	Number of followers.	float
followees	Number of followees.	float

Some followers did not appear in our database of tweets and are excluded from the table. The columns are detailed in Table 6.

3.7 Positions of media domains

The file `domains_positions.csv` contains the positions of the media domains. It has 400 rows and 84 columns. Each row corresponds to one media domain. The rows are sorted in order of decreasing user_count (see below). The columns are detailed in Table 7.

4 Validations

In this section we present several computations showing the quality and consistency of the positions computed for users and for medias. Our method for position estimation is purely structural, in that it relies solely on the *follow* relationships that exist between MPs and other users on X, and on relations between medias and users whenever they include them in their tweets. Notably, our methods so far exclude textual data. To demonstrate the robustness of our results, we compare the positions we computed with political stances stated by the users

Table 7: Detail of columns for `domains.csv`.

Column Name	Description	Format
domain	Web address of the domain.	string
media_category	Media category from ref. [42].	string
user_count	Number of users who tweeted URLs pointing towards this domain.	integer
tweet_count	Number of tweets containing URLs pointing towards this domain.	integer
xxx_yy_mean (e.g. lrecon_23_mean)	Average position of users who shared URLs from this domain on dimension xxx in the space of CHES 20yy. 16 such columns.	float
xxx_yy_std (e.g. lrecon_23_std)	Standard deviation of the positions of users who shared URLs from this domain. 16 such columns.	float
xxx_yy_quantile (e.g. lrecon_23_quantile)	Quantile (20%) calculated from the distribution of the values of columns xxx_yy_mean across all domains. 16 such columns.	integer
xxx_yy_dip (e.g. lrecon_23_dip)	Statistic of Hartigan's dip test [71] performed on the distribution of the positions of users who shared URLs pointing toward this domain, in the corresponding dimension. 16 such columns.	float
xxx_yy_pval (e.g. lrecon_23_pval)	p-value of Hartigan's dip test. 16 such columns.	float

through text in their profile bios (text self-descriptions curated by users to appear at the top of their personal profiles in the platform). In these bios, users may often convey, as we show below, information accusing their ideological stances (e.g., presenting themselves explicitly as Left- or Right-leaning individuals) and their stances on issues (e.g., explicitly taking position on immigration or environmental protection).

We now present the protocols through which we identified users expressing political stances in their profile bios, and how we leveraged them in computing metrics showing the quality of the ideology and issue positions in our dataset. All of the results and figures presented in this section can be reproduced using the code provided with the data. The exact text of bios are not provided in the data to protect the anonymity of users.

4.1 Annotation of profile bios

For each of the selected ideology and issue dimensions, we establish a pair of labels associated with a distinguishable leaning or stances of users, for which the dimension should provide an order. For instance, using the text profile bios, we label each user as being Left-leaning ("left") or Right-leaning ("right") whenever possible (admittedly, not all users hint at their leaning on their profiles) to quantify the degree to which the continuous Left-Right dimension we inferred orders these users. For the E.U. dimension, we annotate users as"eurosceptic" or "pro-european" whenever possible. Dimensions and the corresponding annotation labels are listed in Table 8. We aim to identify users that have distinguishable self-reported leanings and stances, and not all possible ways in which users convey their stances. In other words, our labels are suited for our validation protocol, but do not exhaust the many possible ways in which users hint at their political positions through text in their bios. This strategy borrows from that of previous ideology scaling using follower networks by Barbera [38], in which he identified users self-identifying as liberal or conservative (among other labels) to validate their estimated ideological positions. Our strategy expands that of Barbera, adapting a greater number of dimensions, similar to what Ramaciotti et al. [72] proposed for validating two dimensions: Left-Right and anti-elite positions. We compute these labels in two independent ways: using LLM automatic annotation, and, for fewer dimensions (because of the involved cost), using human annotations. See in Table 8 the labels used and the corresponding dimensions in which we used them for validation.

The criteria and instructions used for annotation (both for automatic and human) rely on the description of dimensions contained in the political survey data we use for calibration, i.e., the CHES data. The CHES data codebooks contain detailed descriptions of the substance of each dimension. Some dimensions do not span between dychotomous extremes (like Left-Right), but measure the degree to which a given attitude is held. The most notable example is the CHES anti-elite dimension, measuring the degree of adoption of anti-elite and anti-establishment populist rhetoric. In this case, to produce dichotomous labels, we label users as being from "elite" groups or displaying "populist" rhetoric to validate the Anti-elite dimension. Individuals displaying populist rhetoric (e.g., subscribing to a vision of society divided into the "people" and "elites") may be *also* part of different elites. The validation we propose relies on the fact that *most* users engaging in acute anti-elite rhetoric will not be part of the elite groups, with high probability.

LLM annotation. To submit the profile bios to an LLM for annotation, we first translated each follower profile bio text to English using `M2M100 1.2B`[73], to minimize the probability of the quality of our annotations depending on the language. We then submitted each English-translated bio to the Large Language Model `zephyr-7B-`β[74], producing binary annotations for the collection of political labels considered. The accuracy and utility of LLM annotations for the classification of political content has been examined in several recent studies[75,76]. The prompts used for annotation were chosen after an iterative series of tries, manually evaluating the

quality of the annotations for a small subset. Here are two example prompts that we use; an exhaustive list is provided in Appendix.

- LLM label "Left": *You are an expert in French politics. Please classify the following X profile bio as "Left- leaning" or "Not-Left" according to whether the author of the text (who is from France) is politically Left-leaning or not. The response should be in the form of a single term with the name of the category: "Left-leaning" or "Not-Left": [TEXT OF THE BIO].*
- LLM label "Eurosceptic": *You are an expert in European politics. Please classify the following X profile bio as "Eurosceptic" or "Not-Eurosceptic" according to whether the author of the text (who is from France) holds negative views of the European Union or not. The response should be in the form of a single term with the name of the category: "Eurosceptic" or "Not-Eurosceptic": [TEXT OF THE BIO].*

Our prompts produce a very large proportion of outputs in the intended requested form. We discarded outputs that do not correspond to one of the two requested allowed categories specified in the prompt. Then we examine LLM annotations produced for pairs of dichotomous labels ("Left" and "Right", "Pro-European" and "Eurosceptic") to further discard annotations that are contradictory (e.g. users annotated as being both "Left"- and "Right"-leaning by the LLM). Upon manual examination, we noted that this happened mainly because of ambiguous political messages, e.g. "I am a left-winger because I believe there should not be private property but a right-winger because I believe we should expel immigrants" (fictional example). Table 10 reports the number of bios for which we were able to obtain a label. We compare the annotations for each pair of opposite categories (e.g. "left" and "right", "populist" and "elite") with the corresponding CHES dimensions (e.g. lrecon_23, antielite_salience_23). The full list of correspondence between pairs of labels and CHES dimensions is summarized in Table 8.

We leveraged annotated profiles in two forms of validation: examination of the monotonicity of the concentration of users with labels along dimensions, and measurement of the order and separation between labeled users along dimensions.

Human annotation. Rather than evaluating LLM annotations with human annotators, we seek to produce an independent measure of validity of estimated positions of users using human annotations. Because of the cost involved in producing human annotations, we employed a different strategy and protocol. First, we only considered a few dimensions, namely, Left-Right dimensions (ideological and economic), attitudes towards the EU, immigration, and anti-elite and anti-establishment sentiments. Second, because cost of human annotations were prohibitive, we adopted neither a purely *descriptive* nor purely *prescriptive* paradigm[77], but a mixed approach. Instead of providing annotators with the same instructions given to LLMs, we allowed for annotators to propose criteria to define the labels in a two-phase protocol. The two steps of this descriptive-prescriptive annotation strategy are: 1) first, annotators examine the bios to specify a set of criteria, to then 2) apply these criteria to produce the labels. While the annotators know the prompts given to LLMs, and thus the number of labels that need to be produced and the main concept underlying them, they are free to propose the rest of the annotation protocol to follow for classifying texts into labels. Concretely, their proposed protocols allow for them to filter and select the bios that they will then annotate (e.g., based on the presence of keywords they think are important). This strategy can thus assure only a low rate of false positives and a high rate of true positives, but says little about true and false negatives. More details about the human annotation protocol are provided in the Appendix. Sections 4.2 & 4.3 demonstrate the convergence validity between LLM and human annotations, and show that both strategies lead to a validation of the estimated positions of users along all targeted CHES dimensions, as quantified by classification metrics.

4.2 Concentration of labeled users along dimensions

First, we assess whether labeled users are coherently positioned along the corresponding CHES dimensions. We divide the range of values of each CHES dimension into one-sized bins (0-1, 1-2, 2-3, 3-4, 4-5, 5-6, 6-7, 7-8, 8-9, and 9-10) and group users accordingly. We then proceed to examine the fraction of labeled users among the bins, for the labels corresponding to the selected dimension (Table 8). We also compute Clopper-Pearson confidence intervals[78] for the fraction of labeled users on each bin. Our first validation consists in evaluating that the concentration of labeled users is monotonic, and that it increases or decreases coherently across the corresponding dimension. The figures for 15 labels (each associated to a CHES dimension), are included with the reproducibility material. The result for a couple of labels is shown here for illustration.

Fig. 2 shows the distribution of users labeled as having "left" leaning (both by human and LLM annotators), showing that their concentration increases monotonically towards lowers values of the CHES Left-Right dimension. Fig. 2 also shows additional examples, including the distribution of users labeled (by humans) as "eurosceptic", monotonically concentrated towards lower values of the CHES EU dimension, and users labeled as displaying "liberal immigration" stances (by the LLM) along the CHES dimension measuring stances towards immigra-

tion. These selected examples illustrate how human and LLM annotation of text profile bios produce labels for users that are concentrated coherently with the ideology or stances on issues that the CHES dimensions capture. We find convincing results for all dimensions. Plots for all combinations of dimension and labels are provided with the data, and can also be reproduced using the reproducibility code available with this article.

4.3 Order and separation of labeled users along CHES dimensions

We now seek to evaluate how well users with opposite labels are separated along the corresponding CHES dimensions. To this end, we fit a logistic regression model between labels and positions in our opinion space. For the example of "left" and "right" labeled users along CHES Left-Right dimensions, we consider one label as *failures* and the other as *successes*, we use positions along Left-Right to fit the model, and validate the induced separation and order of labeled users with goodness of fit metrics. Because logistic regression is sensible to class imbalance, and because some of our labels are imbalanced, we systematically weight the samples of labeled users by the size of their label class[4]. Once we have fitted a logistic regression model, we assess its goodness of fit by measuring ROC AUC and F1 scores[79], presented in Table 8. Because F1-score is asymmetric (i.e., achieves different results depending on which label is taken as success), we compute it both ways (i.e., considering one class as success and then as failure) and average the result. Asymmetric F1-scores for each label ordering are provided in Appendix (Table 10), while Precision and Recall are computed in the reproducibility code accompanying the data. In other words, we use the training accuracy of a logistic classifier for pairs of labels on the corresponding dimensions as a goodness of fit metric.

Upon evaluation, we observe high ROC AUC scores (greater than 0.7) for all the selected dimensions that we test. For most dimensions we obtained remarkably high scores, greater than 0.9. Goodness of fit metrics are coherent when comparing human and LLM annotation, although human annotations always achieve better goodness of fit scores. All of our validations except three achieve F1-score values above 0.6. Fig. 3 illustrates the goodness of fit metrics on selected examples.

4.4 Consistency between dimensions present in the CHES 2019 and 2023 waves

Four dimensions exist both in the 2019 and 2023 waves of the CHES data. The 2019 and 2023 are both relevant to our estimates because we used follower links collected in 2023, but produced by users before that. These dimensions are economic Left-Right, attitudes towards the EU, Liberal-Conservative, and Anti-elite sentiments. To assess both the reliability of the position inference method we use, and the sensibility of the method to data from these waves, we compare positions of parties, MPs, and their followers. To compare between positions on a 2019 and a 2023 dimension, we compute the value of Pearson correlation. As shown in Fig. 4, we find high correlations, indicating consistency across the two surveys (all equal or greater than 0.869).

4.5 Positions of media outlets

We now validate the position of the web domains that correspond to French media outlets in our political CHES dimensions. To this end, we leverage a previously curated categorical classification of the most relevant French media outlets by Cointet et al.[41,42]. This dataset comprises a list of 478 French media domains that users on Twitter (now X) cite most frequently in posts. In this data by Cointet et al., news outlets and politically engaged medias are categorized into the following categories: "Centre", "Hyper-centre", "Left Wing", "Right Wing", "Revolutionary Right" and "Identitarian". These categories correspond to clusters of the hyperlink citation network of articles from these outlets, computed using the crawled hyperlink network and Stochastic Block Model inference (see the related publication[41] for details). The categories "Centre" and "Hyper-centre" are not named after the political stance of the corresponding medias, but rather because of the central role they occupy in citation networks, gathering the most important medias that reach a national audience and may thus be read by a large and diverse crowd. These central positions are, as shown in Fig. 5 and other related plots available with the data, tightly related to centrist ideological stances.

Among our 400 media domains, 217 are categorized in the work of Cointet et al. We use the categorization of those domains to validate our political positioning of the domains. For each dimension of the political space, we compare the distribution of the domains positions in the political space across the different categories. We illustrate the result by plotting the distribution of positions along the considered dimension for each category of media. Two examples are shown in Fig. 5; additional plots are provided with the data and can also be reproduced with the reproducibility code.

5 Limitations and perspectives

Our dataset offers the potential for several applications. Researchers interested in using this data should however keep in mind that it concerns a specific population, that of X users who follow French Members of Parliament. As online users are in general not representative of the overall population of a country[80,81], conclusions drawn from analyses of this database should not be generalized to

[4]We use for this the implementation provided with the `scikit-learn` library, providing the parameter `class_weight='balanced'`.

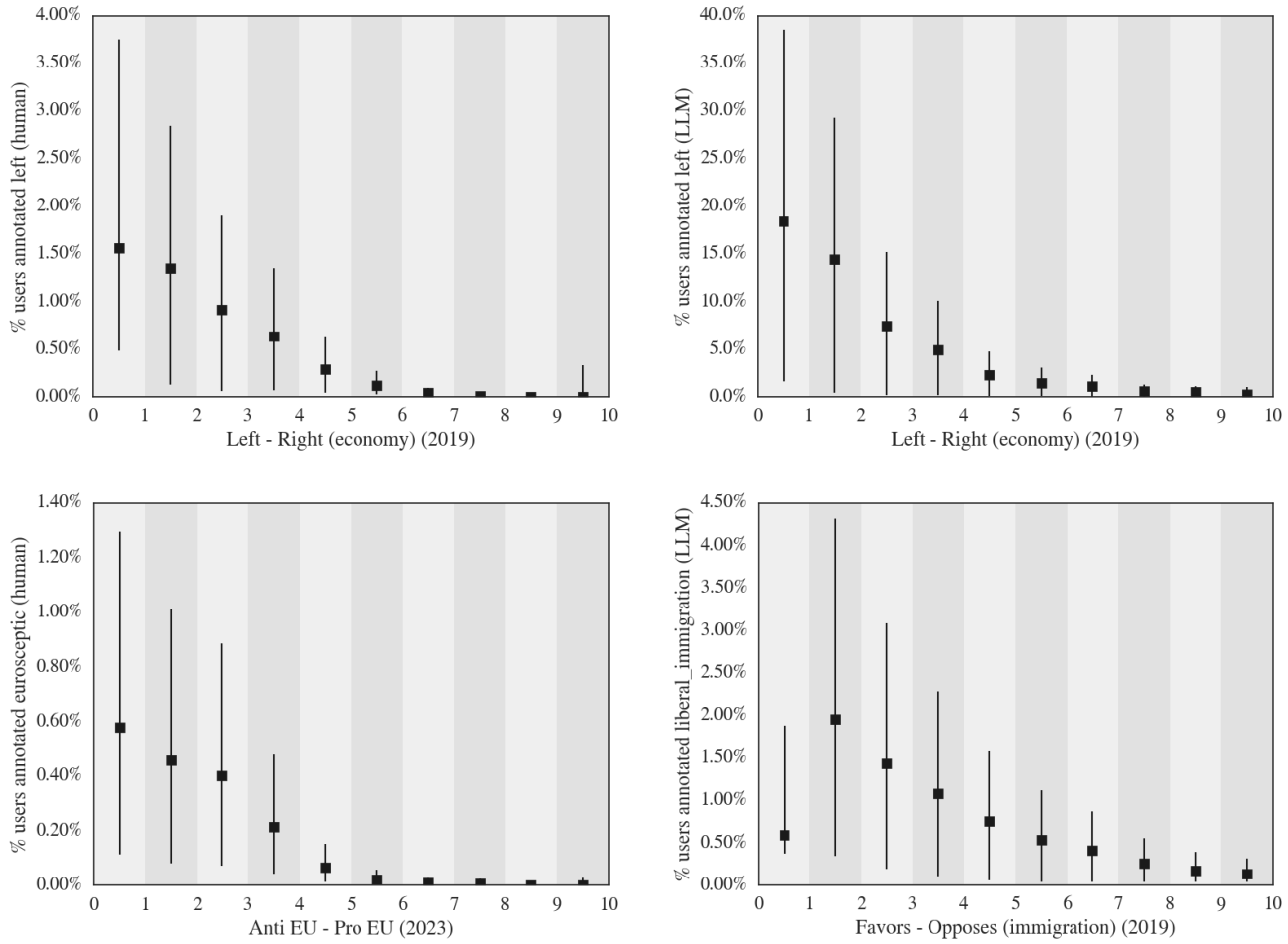

Figure 2: Proportion of users having a given label indicating ideological or issue stances, as provided by both human and LLM annotations, along the corresponding CHES dimensions they intend to validate. Vertical bars delimit Clopper-Pearson confidence intervals[78] at the $\alpha = 0.05$.

the whole ensemble of French citizens without a thorough evaluation of the validity of such procedure. In fact, this database can also serve as the basis of a proper analysis of the ideological similarities and differences between the online and the general populations, by performing comparisons with respondent surveys such as the *European Social Survey*[82].

Because we rely on the Chapel Hill Expert Surveys to position users along the political dimensions, these positions are inherently constrained by the structure of the opinion space subtended by the French political parties.

In the future, this database could be further enriched by subsequent collections to study the evolution of the political space. Further research shall also strive to develop databases of individual political positions for other countries and other online platforms relevant for political communication, such as YouTube, TikTok or Bluesky.

6 Data availability

The dataset and the code needed to reproduce analyses in this article are available at https://doi.org/10.17605/OSF.IO/AT5Q2.

A filtered dataset where users with less than 25 followers are discarded (see Methods) is available at https://doi.org/10.17605/OSF.IO/V28KH.

Funding

This work has been partially funded by the "European Polarisation Observatory" (EPO) of CIVICA Research co-funded by EU's Horizon 2020 programme (Grant Agreement No. 101017201), by the project SoMe4Dem funded by EU's Horizon Europe programme (Grant No. 101094752), by the Project Liberty Institute's project "AI-Political Machines", and by the *Very Large Research Infrastructure* (TGIR) Huma-Num of CNRS, Aix-Marseille Université and Campus Condorcet.

Table 8: Summary of results for goodness of fit for the logistic regression models, sorted by decreasing ROC AUC, serving as main validation of the position of users along selected Chapel Hill Expert Survey (CHES) data dimensions from the 2019 and 2023 waves. Each CHES dimension is validated using users labeled with pairs of dichotomous labels that should be well ordered and separated by the dimension. For the sake of space, "imm." stands for "immigration".

Dimension	Year	Annotator	Label A	Label B	ROC AUC	Avg. F1
lrgen	2019	human	left	right	0.992	0.961
lrecon	2019	human	left	right	0.992	0.952
lrecon	2023	human	left	right	0.979	0.946
immigrate_policy	2019	human	liberal_immigration	restrictive_immigration	0.972	0.924
eu_position	2023	human	eurosceptic	pro_european	0.969	0.924
eu_position	2019	human	eurosceptic	pro_european	0.967	0.921
refugees	2023	human	liberal_immigration	restrictive_immigration	0.962	0.924
galtan	2023	LLM	conservative	liberal	0.943	0.882
galtan	2019	LLM	conservative	liberal	0.941	0.883
sociallifestyle	2019	LLM	conservative	liberal	0.941	0.877
lrgen	2019	LLM	left	right	0.931	0.863
lrecon	2019	LLM	left	right	0.929	0.859
eu_position	2023	LLM	eurosceptic	pro_european	0.927	0.823
immigrate_policy	2019	LLM	liberal_immigration	restrictive_immigration	0.922	0.864
eu_position	2019	LLM	eurosceptic	pro_european	0.917	0.803
nationalism	2019	LLM	cosmopolitan	nationalist	0.915	0.864
antielite_salience	2023	LLM	elite	populist	0.915	0.697
refugees	2023	LLM	liberal_immigration	restrictive_immigration	0.911	0.843
antielite_salience	2023	human	elite	populist	0.903	0.824
lrecon	2023	LLM	left	right	0.894	0.854
antielite_salience	2019	LLM	elite	populist	0.886	0.684
antielite_salience	2019	human	elite	populist	0.866	0.813
corrupt_salience	2019	LLM	elite	populist	0.848	0.609
people_vs_elite	2019	LLM	elite	populist	0.772	0.563
corrupt_salience	2019	human	elite	populist	0.762	0.640
people_vs_elite	2019	human	elite	populist	0.723	0.647
environment	2019	LLM	climate_denialist	pro_environment	0.653	0.362
environment	2019	LLM	economic_focus	pro_environment	0.639	0.546

CRediT Author Statement

Antoine Vendeville: Software, Validation, Formal analysis, Investigation, Data curation, Writing, Visualization. **Jimena Royo-Letelier:** Software, Formal analysis, Resources, Data curation. **Duncan Cassells:** Data curation. **Jean-Philippe Cointet:** Data curation. **Maxime Crépel:** Data curation. **Tim Faverjon:** Data curation. **Théophile Lenoir:** Data curation. **Béatrice Mazoyer:** Resources, Data curation. **Benjamin Ooghe-Tabanou:** Resources, Data curation. **Armin Pournaki:** Resources, Data curation. **Hiroki Yamashita:** Data curation. **Pedro Ramaciotti:** Conceptualization, Methodology, Investigation, Resources, Data curation, Writing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of competing interests

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Jennifer McCoy, Tahmina Rahman, and Murat Somer. Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, 62(1):16–42, January 2018. ISSN 0002-7642. doi: 10.1177/0002764218759576.

[2] Delia Baldassarri and Andrew Gelman. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):404–446, 2008.

[3] Andres Reiljan. 'fear and loathing across party lines' (also) in europe: Affective polarisation in european party systems. *European Journal of Political Research*, 59(2):376–396, 2020. ISSN 1475-6765. doi: 10.1111/1475-6765.12351.

[4] Levi Boxell, Matthew Gentzkow, and Jesse M. Shapiro. Cross-country trends in affective polarization. *The Review of Economics and Statistics*, 106(2):557–565, March 2024. ISSN 0034-6535. doi: 10.1162/rest_a_01160.

[5] Yunus Emre Orhan. The relationship between affective polarization and democratic backsliding: Comparative evidence. *Democratization*, 29(4):714–735, May 2022. ISSN 1351-0347. doi: 10.1080/13510347.2021.2008912.

[6] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22(1):129–146, May 2019. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev-polisci-051117-073034.

[7] Morris P Fiorina and Samuel J Abrams. Political polarization in the american public. *Annu. Rev. Polit. Sci.*, 11:563–588, 2008.

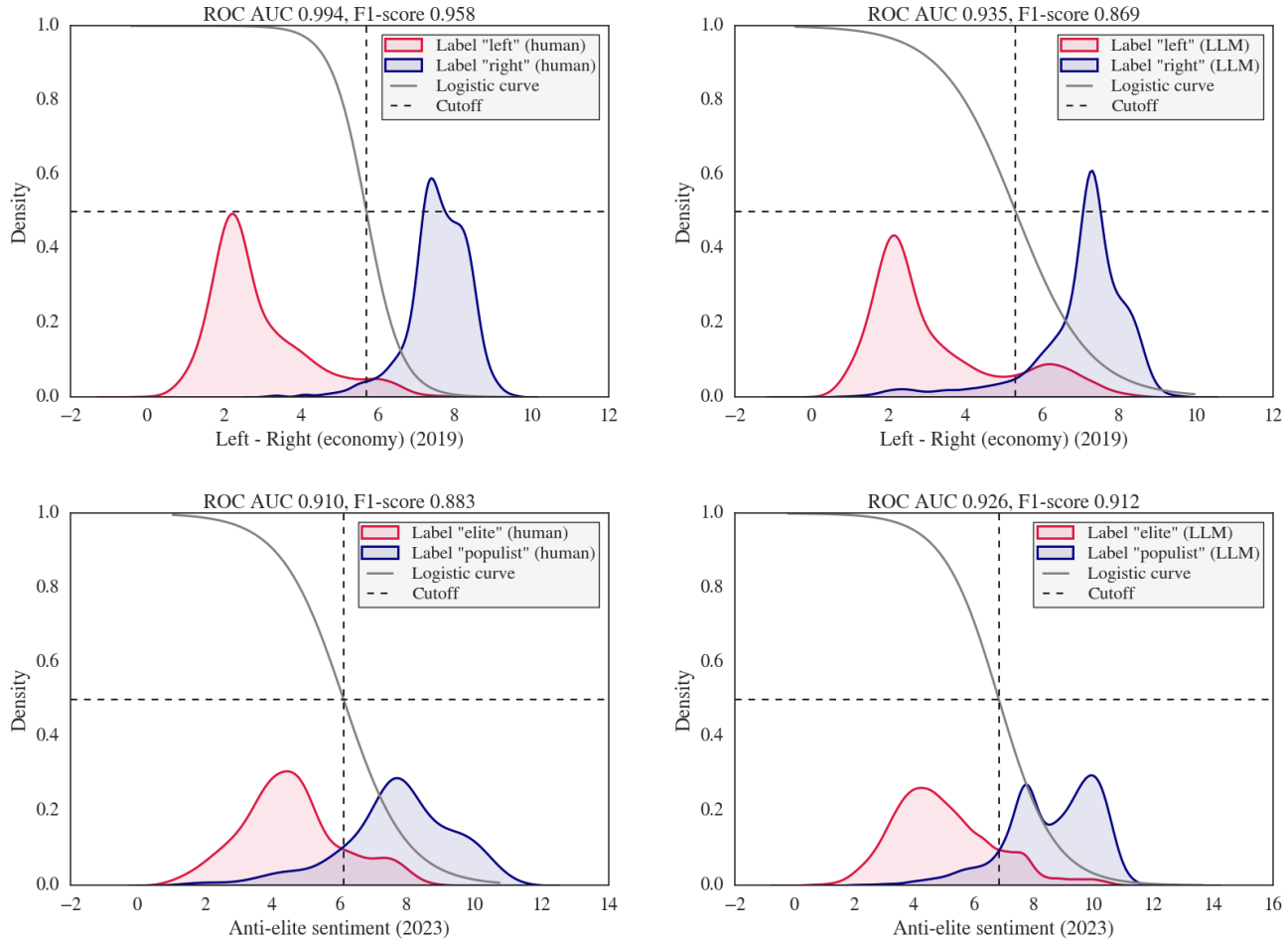

Figure 3: Illustration of the logistic regression model for the economical left-right dimension (CHES 2019), and the anti-elite sentiment dimension (CHES 2023). We show results obtained with human as well as with LLM annotations. Blue and red areas indicate the distributions of the political position of the annotated users. The cutoff determines where the model changes its prediction. We also indicate ROC AUC and F1 scores.

[8] 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, Linda J. Skitka, Joshua A. Tucker, Jay J. Van Bavel, Cynthia S. Wang, and James N. Druckman. Political sectarianism in america. *Science*, 370(6516):533–536, October 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abe1715.

[9] James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. Affective polarization, local contexts and public opinion in america. *Nature human behaviour*, 5(1):28–38, 2021.

[10] Austin C. Kozlowski and James P. Murphy. Issue alignment and partisanship in the american public: Revisiting the 'partisans without constraint' thesis. *Social Science Research*, 94:102498, February 2021. ISSN 0049089X. doi: 10.1016/j.ssresearch.2020.102498.

[11] Yphtach Lelkes. Mass polarization: Manifestations and measurements. *Public Opin Q*, 80(S1): 392–410, January 2016. ISSN 0033-362X. doi: 10.1093/poq/nfw005.

[12] Antonio F. Peralta, Pedro Ramaciotti, János Kertész, and Gerardo Iñiguez. Multidimensional political polarization in online social networks. *Phys. Rev. Res.*, 6:013170, Feb 2024. doi: 10.1103/PhysRevResearch.6.013170.

[13] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348 (6239):1130–1132, 2015. doi: 10.1126/science.aaa1160. URL https://www.science.org/doi/abs/10.1126/science.aaa1160.

[14] Venkata Rama Kiran Garimella and Ingmar Weber. A long-term analysis of polarization on twitter. *Proceedings of the International AAAI Conference*

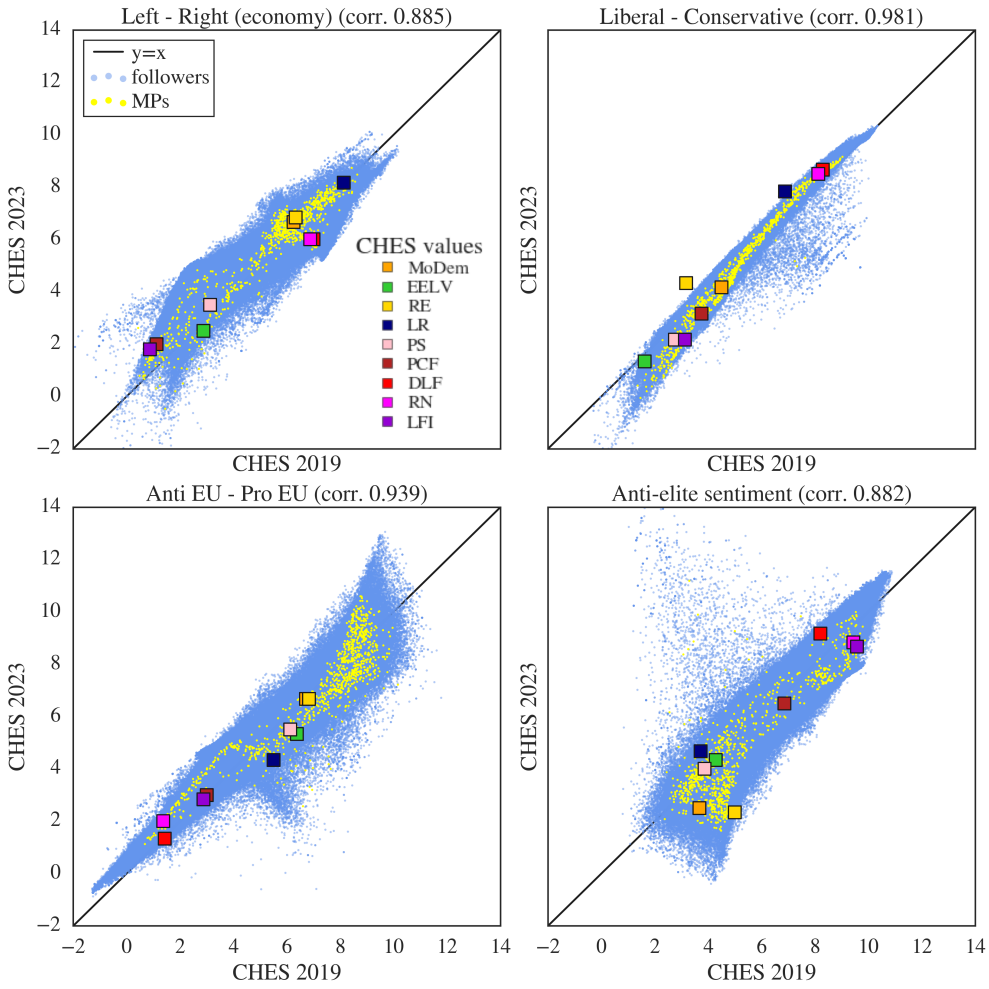

Figure 4: Positions of followers, MPs and party centroids along the four dimensions present in the 2019 and 2023 waves of CHES surveys. Blue dots are followers, yellow dots are MPs, colored squares represent party positions in the CHES surveys. We show the $y = x$ line in black. We also indicate Pearson correlations between the positions of users on the two axes.

on Web and Social Media, 11(1):528–531, May 2017. ISSN 2334-0770. doi: 10.1609/icwsm.v11i1.14918.

[15] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96, 2011. doi: 10.1609/icwsm.v5i1.14126.

[16] James Flamino, Alessandro Galeazzi, Stuart Feldman, Michael W Macy, Brendan Cross, Zhenkun Zhou, Matteo Serafino, Alexandre Bovet, Hernán A Makse, and Boleslaw K Szymanski. Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections. *Nature Human Behaviour*, 7(6):904–916, 2023.

[17] Tuan M. Pham, Andrew C. Alexander, Jan Korbel, Rudolf Hanel, and Stefan Thurner. Balance and fragmentation in societies with homophily and social balance. *Scientific Reports*, 11(1):17188, December 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-96065-5.

[18] Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini. Emergence of polarized ideological opinions in multidimensional topic spaces. *Phys. Rev. X*, 11:011012, Jan 2021. doi: 10.1103/PhysRevX.11.011012. URL https://link.aps.org/doi/10.1103/PhysRevX.11.011012.

[19] Ryan Bakker, Seth Jolly, and Jonathan Polk. Complexity in the European party space: Exploring dimensionality with experts. *European Union Politics*, 13(2):219–245, 2012. doi: 10.1177/1465116512436995.

[20] Sylvia Kritzinger Martin Dolezal, Nikolaus Eder and Eva Zeglovits. The structure of issue attitudes revisited: A dimensional analysis of Austrian voters and party elites. *Journal of Elections, Pub-*

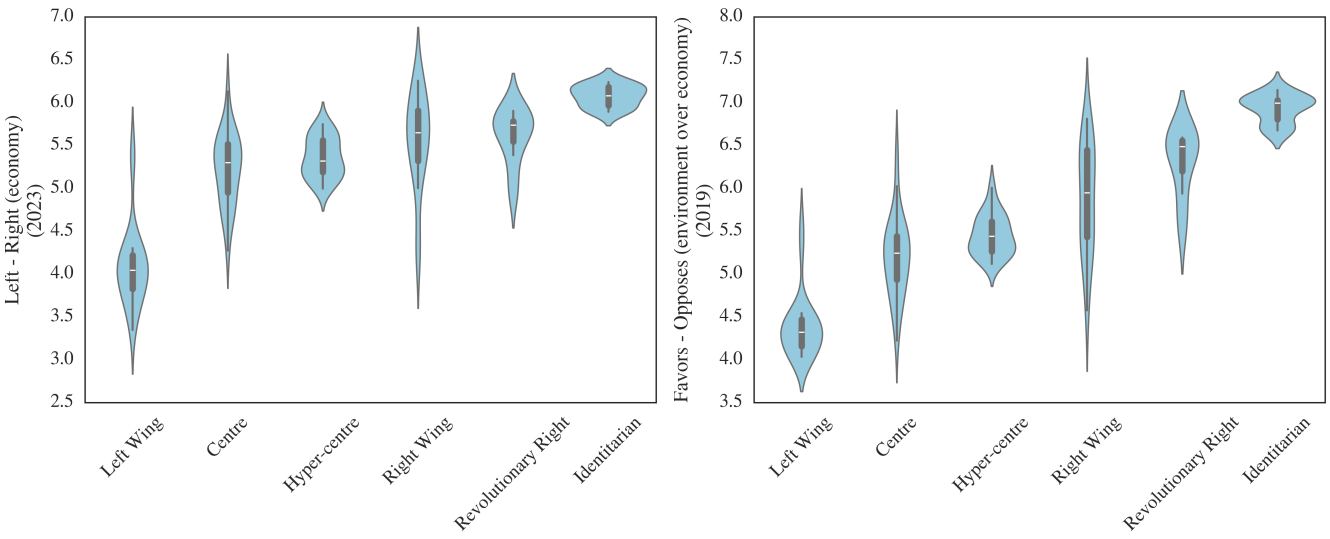

Figure 5: Validation of the media positions along two political axes. We show the distribution of positions among each media category. These figures demonstrate a strong alignment between the categorization of media along the left-right axis and the economical Left-Right dimension as well as the Environment dimension. The latter is consistent with the difference in environmental discourse between the left and the right in France.

lic Opinion and Parties, 23(4):423–443, 2013. doi: 10.1080/17457289.2013.803195.

[21] Heather Stoll. Elite-level conflict salience and dimensionality in western europe: Concepts and empirical findings. *West European Politics*, 33(3):445–473, 2010. doi: 10.1080/01402381003654494.

[22] Simon Bornschier. The new cultural divide and the two-dimensional political space in western europe. *West European Politics*, 33(3):419–444, 2010. doi: 10.1080/01402381003654387.

[23] Suay M. Özkula, Paul J. Reilly, and Jenny Hayes. Easy data, same old platforms? A systematic review of digital activism methodologies. *Information, Communication & Society*, 26(7):1470–1489, May 2023. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2021.2013918.

[24] Emily Kubin and Christian Von Sikorski. The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association*, 45(3):188–206, July 2021. ISSN 2380-8985, 2380-8977. doi: 10.1080/23808985.2021.1976070.

[25] Amin Mekacher, Max Falkenberg, and Andrea Baronchelli. The systemic impact of deplatforming on social media. *PNAS Nexus*, 2(11):pgad346, November 2023. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgad346.

[26] Antonis Papasavva, Jeremy Blackburn, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. "Is it a Qoincidence?": An exploratory study of QAnon on Voat. In *Proceedings of the Web Conference 2021*, WWW '21, page 460–471, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450036.

[27] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. Raiders of the Lost Kek: 3.5 years of augmented 4chan posts from the politically incorrect board. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):885–894, May 2020. doi: https://doi.org/10.1609/icwsm.v14i1.7354. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7354.

[28] Francesco Corso, Francesco Pierri, and Gianmarco De Francisci Morales. What we can learn from TikTok through its research API. In *Companion Publication of the 16th ACM Web Science Conference*, Websci Companion '24, page 110–114, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704536. doi: 10.1145/3630744.3663611.

[29] Amin Mekacher, Max Falkenberg, and Andrea Baronchelli. The Koo dataset: An Indian microblogging platform with global ambitions. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:1991–2002, May 2024. ISSN 2334-0770. doi: 10.1609/icwsm.v18i1.31442.

[30] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y. Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew

Gentzkow, Sandra González-Bailón, Andrew M. Guess, Edward Kennedy, Young Mie Kim, David Lazer, Neil Malhotra, Devra Moehler, Jennifer Pan, Daniel Robert Thomas, Rebekah Tromble, Carlos Velasco Rivera, Arjun Wilkins, Beixian Xiong, Chad Kiewiet De Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, 620(7972):137–144, August 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06297-w.

[31] Andrew M. Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, July 2023. doi: 10.1126/science.abp9364.

[32] Andrew M. Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, 381(6656):404–408, July 2023. doi: 10.1126/science.add8424.

[33] Sandra González-Bailón, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M. Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet De Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, 381(6656):392–398, July 2023. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.ade7138.

[34] Michael W. Wagner. Independence by permission. *Science*, 381(6656):388–391, 2023. doi: 10.1126/science.adi2430.

[35] Jon Roozenbeek and Fabiana Zollo. Democratize social-media research-with access and funding. *Nature*, 612(7940):404, 2022. doi: 10.1038/d41586-022-04407-8.

[36] Pedro Ramaciotti Morales, Jean-Philippe Cointet, Gabriel Muñoz Zolotoochin, Antonio Fernández Peralta, Gerardo Iñiguez, and Armin Pournaki. Inferring attitudinal spaces in social networks. *Social Network Analysis and Mining*, 13(1):14, December 2022. ISSN 1869-5469. doi: 10.1007/s13278-022-01013-4.

[37] Joshua Clinton, Simon Jackman, and Douglas Rivers. The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370, 2004.

[38] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91, 2015. ISSN 1047-1987, 1476-4989. doi: 10.1093/pan/mpu011.

[39] Seth Jolly, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. Chapel hill expert survey trend file, 1999-2019. *Electoral Studies*, 75:102420, 2022. ISSN 0261-3794. doi: 10.1016/j.electstud.2021.102420.

[40] Liesbet Hooghe, Gary Marks, Ryan Bakker, Seth Jolly, Jon Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. The russian threat and the consolidation of the west: How populism and EU-skepticism shape party support for Ukraine. *European Union Politics*, 25(3), 2024. URL https://www.chesdata.eu/ches-europe.

[41] Jean-Philippe Cointet, Dominique Cardon, Andreï Mogoutov, Benjamin Ooghe-Tabanou, Guillaume Plique, and Pedro Morales. Uncovering the structure of the French media ecosystem. 2021. URL https://arxiv.org/abs/2107.12073.

[42] Guillaume Plique, Benjamin Ooghe-Tabanou, Jean-Philippe Cointet, Dominique Cardon, Audrey Baneyx, Paul Girard, Arnaud Pichon, Maxime Crepel, Oubine Perrin, Diego Antolinos-Basso, Benjamin Tainturier, and Tim Fingerhut. Corpus Médias "Polarisation de l'Espace Public Numérique" [dataset]. *data.sciencespo* https://doi.org/10.21410/7E4/HZB8D0, 2021.

[43] Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besançon, and Mohand Boughanem. #Élysée2017fr: The 2017 French Presidential Campaign on Twitter. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*, 2018. doi: 10.1609/icwsm.v12i1.14984.

[44] Ophélie Fraisier, Guillaume Cabanac, Yoann Pitarch, Romaric Besançon, and Mohand Boughanem. #Élysée2017fr: The 2017 French presidential campaign on Twitter [dataset]. *Zenodo* https://doi.org/10.5281/zenodo.5535333, June 2018.

[45] Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on facebook. 2015.

[46] Mariano Torcal, Emily Carty, Josep Maria Comellas, Oriol J. Bosch, Zoe Thomson, and Danilo Serani. The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022). *Data in Brief*, 48:109219, June 2023. ISSN 23523409. doi: 10.1016/j.dib.2023.109219.

[47] Mariano Torcal, Emily Carty, Josep Maria Comellas, Oriol J. Bosch, Zoe Thomson, and Danilo Serani. The dynamics of political and affective polarisation: Datasets for Spain, Portugal, Italy, Argentina, and Chile (2019-2022) [dataset]. *OSF* http://doi.org/10.17605/OSF.IO/3T7JZ, 2023.

[48] Vivek Srivastava and Mayank Singh. PoliWAM: An exploration of a large scale corpus of political discussions on WhatsApp Messenger, 2021. URL https://arxiv.org/abs/2010.13263.

[49] Vivek Srivastava and Mayank Singh. PoliWAM: An exploration of a large scale corpus of political discussions on WhatsApp Messenger [dataset]. *Zenodo* https://zenodo.org/records/4115660#.X5BHSngzZQI, 2021.

[50] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pages 36–43, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595932151. doi: 10.1145/1134271.1134277.

[51] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. election: Divided they blog [dataset]. *Figshare* https://figshare.com/articles/dataset/A_directed_network_of_hyperlinks_between_weblogs_on_US_politics_recorded_in_2005_by_Adamic_and_Glance_/1149954?file=1648333, 2005. Non-official repository.

[52] Pedro Ramaciotti Morales, Jean-Philippe Cointet, and Gabriel Muñoz Zolotoochin. Unfolding the dimensionality structure of social networks in ideological embeddings. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, page 333–338, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391283. doi: 10.1145/3487351.3489441. URL https://doi.org/10.1145/3487351.3489441.

[53] Pedro Ramaciotti Morales, Jean-Philippe Cointet, and Julio Laborde. Your most telling friends: Propagating latent ideological features on twitter using neighborhood coherence. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 217–221, 2020. doi: 10.1109/ASONAM49781.2020.9381468.

[54] Robert Bond and Solomon Messing. Quantifying social media's political space: Estimating ideology from publicly revealed preferences on facebook. *American Political Science Review*, 109(1):62–78, 2015. doi: 10.1017/S0003055414000525.

[55] Pablo Barberá and Gonzalo Rivero. Understanding the political representativeness of Twitter users. *Social Science Computer Review*, 33(6):712–729, 2015.

[56] François Briatte and Ewen Gallic. Recovering the French party space from Twitter data. In *Science Po Quanti*, Paris, France, May 2015. URL https://shs.hal.science/halshs-01511384.

[57] Guillaume Plique, Pauline Breteau, Jules Farjas, Héloïse Théro, Jean Descamps, Amélie Pellé, and Laura Miguel. Minet, a webmining CLI tool & library for Python, April 2023. URL https://doi.org/10.5281/zenodo.7791759.

[58] Robert C Luskin. Explaining political sophistication. *Political behavior*, 12:331–361, 1990.

[59] Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, and Andrea Baronchelli. Growing polarization around climate change on social media. *Nature Climate Change*, 12 (12):1114–1121, December 2022. ISSN 1758-678X, 1758-6798. doi: 10.1038/s41558-022-01527-x.

[60] Max Falkenberg, Fabiana Zollo, Walter Quattrociocchi, Jürgen Pfeffer, and Andrea Baronchelli. Patterns of partisan toxicity and engagement reveal the common structure of online political communication across countries. *Nat Commun*, 15(1):9560, November 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53868-0.

[61] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen Carley. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):400–408, August 2021. ISSN 2334-0770, 2162-3449. doi: 10.1609/icwsm.v7i1.14401.

[62] Rebekah Tromble. Where have all the data gone? a critical reflection on academic digital research in the post-api age. *Social Media + Society*, 7(1):2056305121988929, 2021. doi: 10.1177/2056305121988929.

[63] Gabriele Di Bona, Emma Fraxanet, Björn Komander, Andrea Lo Sasso, Virginia Morini, Antoine Vendeville, Max Falkenberg, and Alessandro Galeazzi. Sampled datasets risk substantial bias in the identification of political polarization on social media, 2024. URL https://arxiv.org/abs/2406.19867.

[64] Michael Greenacre. *Correspondence Analysis in Practice*. CRC Press, Boca Raton, FL, third edition edition, 2017. ISBN 9781315369983. doi: 10.1201/9781315369983.

[65] Will Lowe. Understanding wordscores. *Political Analysis*, 16(4):356–371, 2008. doi: 10.1093/pan/mpn004.

[66] J. Douglas Carroll, Ece Kumbasar, and A. Kimball Romney. An equivalence relation between correspondence analysis and classical metric multidimensional scaling for the recovery of euclidean distances. *British Journal of Mathematical and Statistical Psychology*, 50(1):81–92, 1997. doi: 10.1111/j.2044-8317.1997.tb01104.x.

[67] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

[68] Benjamin Ooghe-Tabanou, Béatrice Mazoyer, Jules Farjas, and Guillaume Plique. Gazouilloire, a command-line tool for long term collections of tweets, July 2023. URL https://doi.org/10.5281/zenodo.8108616.

[69] Maxime Crépel, Benjamin Ooghe-Tabanou, Guillaume Plique, Béatrice Mazoyer, Kelly Christensen, Jean-Philippe Cointet, Sylvain Parasie, Dominique Cardon, Antoine Machut, and Katharina Tittel. French media ecosystem map [dataset]. *SciencesPo data* https://doi.org/10.21410/7E4/VMMY7L, 2024.

[70] Maxime Crépel, Benjamin Ooghe-Tabanou, Kelly Christensen, Béatrice Mazoyer, Guillaume Plique, Sylvain Parasie, Dominique Cardon, Katharina Tittel, Antoine Machut, and Jean-Pilippe Cointet. Digital mapping of the French media ecosystem. Technical Report Sciences Po médialab - DEFACTO - 02/23/24, Médialab ; AFP ; CLEMI ; Wiki, 2024. URL https://sciencespo.hal.science/hal-04868603.

[71] John A Hartigan and Pamela M Hartigan. The dip test of unimodality. *The annals of Statistics*, pages 70–84, 1985.

[72] Pedro Ramaciotti, Jimena Royo-Letelier, Jean-Philippe Cointet, and Armin Pournaki. Attitudinally-positioned european sample dataset. *Report.*, 2024.

[73] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation, 2020.

[74] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023.

[75] Michael Heseltine and Bernhard Clemm von Hohenberg. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239, 2024. doi: 10.1177/20531680241236239.

[76] Gaël Le Mens and Aina Gallego. Positioning political texts with large language models by asking and averaging. *Political Analysis*, page 1–9, 2025. doi: 10.1017/pan.2024.29.

[77] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*, 2021.

[78] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. ISSN 00063444, 14643510. doi: 10.2307/2331986.

[79] Pedro Ramaciotti Morales and Gabriel Muñoz Zolotoochin. Measuring the accuracy of social network ideological embeddings using language models. In *International Conference on Information Technology & Systems*, pages 267–276. Springer, 2022.

[80] Jonathan Mellon and Christopher Prosser. Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, 4 (3):2053168017720008, July 2017. ISSN 2053-1680, 2053-1680. doi: 10.1177/2053168017720008.

[81] Emilio Zagheni and Ingmar Weber. Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25, April 2015. ISSN 0143-7720. doi: 10.1108/IJM-12-2014-0261.

[82] European Social Survey European Research Infrastructure (ESS ERIC). ESS Round 11 - 2023: Social Inequalities in Health, Gender in Contemporary Europe, 2023. URL https://doi.org/10.21338/ess11-2023.

Appendix

A Impact of retaining users with less than 25 followers

In a seminal paper[38], Barbera suggests removing accounts with less than 25 followers before computing political positions, in order to filter out bots and inactive accounts. However, to allow for research projects to investigate the relations between political opinions and activity, we choose to retain these accounts. This changes the structure of the bipartite follow network, and therefore may impact results for all users and MPs. Our full dataset has 980K users and MPs, among which 518K have at least 25 followers. We now demonstrate that our choice does not significantly alter the results. To do so, we create an alternative version of the dataset where political positions are computed solely for users with at least 25 followers. We compare these positions with those obtained for the same users with the full dataset. We find very high Pearson correlations between the two, ranging from 0.909 (dimension: lrecon_23) to 0.997 (eu_position_19), with an average over all dimensions of 0.973. Retaining users with less than 25 followers therefore does not impact the results, justifying our choice to leave them in.

B Parties and political dimensions

We present the different parties that appear in the data and their characteristics in Table 9. Note the following two differences in nomenclature between our data and the CHES surveys: RE in our data corresponds to LREM in CHES, LFI in our data corresponds to FI in CHES. The following additional parties can be found in our data: Horizons (center right), PRV (Parti Radical, center-right), and LC (Les Centristes, center). MPs without party affiliation are noted as Independent.

C Number of annotated bios

Table 10 provides the number of samples per label, i.e., the number of annotated bios, used for the logistic regression models in the validation, as well as the F1-score of the model. For the sake of space we do not include precision and recall, which are computed in the Jupyter notebook accompanying the data.

D Human annotation protocol

In our manual annotation process, we are not looking to capture all the richness of expressions with which an individual may declare identification with an ideological or issue stance. On the contrary, we are looking to identify a small but sufficient group of individuals for which the group identification is certain, so that we can use this certainty in assessing their spatial coherence. We are not concerned by the diverse ways by which individuals might communicate that they identify with ideological leaning or stances on issues that speak to the selected CHES dimensions. For instance, our protocol has high chances of identifying a description such as "I am a proud leftist" as being on the "left", but low chances of identifying a description such as "I will always follow marxist economic policies", or cues that rely on niche knowledge or current events.

The following are the criteria given to annotators for the labels used in validations.

Left and Right. Criterion: only explicit identification with the "left" of the "right". Examples: "Teacher, always on the left" (left); "Neither left nor right" (neither); "I hate leftists" (neither, this is not a right-winger necessarily); "Husband, parent. Conservative" (neither); "I vote for NUPES" (neither, we are interested in measuring party identification and left- or right-wing identification independently); "I've always been a marxist" (this is more complicated: it references belonging to the left via historical markers, and not current political competition. It's up to the annotator to decide.); "we should deport all migrants, legal AND illegal" (neither, this might be expression of stances aligned with left-right, but we are precisely interested at measuring this alignment, rather than relying on it); "I am on the right side of life" (it's up to the annotator, but we've seen that on some countries these are coined phrases of right-wing identification). In general, we favor explicit mentions of identification with groups "left" and "right".

Populists and elites. Criterion for populist: the person refers to "the people", or "the elite" (also simply "elite"), mentions to "politicians" (plural) in a critical way (not specific criticism of a single politician), criticism towards institutions or elites (including references to corruption), or expressed sympathies for strongman rule (e.g. "If Macron was as brave as Putin we wouldn't have terrorism in France"). Criterion for elite: is a member of an "elite group". This includes several groups theorized in the literature: governing elites, economic elites. We exclude political elites (another group theorized in the literature): council members, majors, senators, deputies, etc. Keep in mind that X is an elite space in itself in many ways, and that we're interested in differences within our X populations. We aim at labeling as "elite" indi-

viduals that are part of very selective elites. Examples: high paying jobs, working positions related to finance, tech, international trade, diplomacy. We acknowledge that the definition for the label "populist" is conceptually clearer than that for "elite". Rather than relying on a robust and inclusive delimitation, our strategy relies on an exclusive and conservative delimitation. Users with the label do not capture the totality of users that may display populist views, nor those of elites belonging to elite groups. Instead, those that have the label, hold a small probability of being mislabeled and should thus be coherently spatialized along the corresponding dimension.

Eurosceptics and Pro Europeans. Criterion for pro European: subscribes to positive views on Europe (includes displaying European symbols, or describing oneself as European, or working on European organization). Criterion for eurosceptic: involves criticism towards the E.U. (including negative mentions of Ursula von der Leyen, Brussels or other symbols of the EC and EU institutions).

Favorable and opposed to liberal immigration policies. Following the framing of our reference document, the CHES codebook, the criterion for both labels were left unqualified beyond what the name signifies. Annotators were instructed to label profile bios according to whether they displayed views that could be deemed "favorable" or "opposed" to "liberal immigration policies".

E LLM annotation protocol

The number of contradictory LLM labels (value 1 in both opposite categories) was the following: left and right 1733 (4.35%) ; populist elite 950 (1.64%) ; eurosceptic and pro_european 110 (0.51%) ; liberal_immigration and restrictive_immigration 53 (0.61%) ; cosmopolitan and nationalist 17 (0.09%) ; pro_environment and climate_denialist 6 (0.02%) ; economic_focus pro_environment 765 (2.04%) ; liberal and conservative 49 (0.34%). Those labels were discarded and replaced by Nans in the published version of the data.

We now list all the prompts that were used for the annotation process with the LLM.

Left. *You are an expert in European politics. Please classify the following X profile bio as "Left-leaning" or "Not-Left" according to whether the author of the text (who is from France) is politically Left-leaning or not. The response should be in the form of a single term with the name of the category: "Left-leaning" or "Not-Left": [TEXT OF THE BIO]*

Right. *You are an expert in European politics. Please classify the following X profile bio as "Right-leaning" or "Not-Right" according to whether the author of the text (who is from France) is politically Right-leaning or not. The response should be in the form of a single term with the name of the category: "Right-leaning" or "Not-Right": [TEXT OF THE BIO]*

Populist. *You are an expert in European politics. Please classify the following X profile bio as "Populist" or "Not-Populist" according to whether the author of the text (who is from France) holds populist views or not. Populist views include, among others, believing that society is split between the people and elites, or that political elites are corrupt. The response should be in the form of a single term with the name of the category: "Populist" or "Not-Populist": [TEXT OF THE BIO]*

Elite. *You are an expert in European politics. Please classify the following X profile bio as "Elite" or "Not-Elite" according to whether the author of the text (who is from France) belongs to an elite group, including political or economic elites. The response should be in the form of a single term with the name of the category: "Elite" or "Not-Elite": [TEXT OF THE BIO]*

Pro European. *You are an expert in European politics. Please classify the following X profile bio as "Pro-European" or "Not-Pro-European" according to whether the author of the text (who is from France) holds positive views of the European Union or not. The response should be in the form of a single term with the name of the category: "Pro-European" or "Not-Pro-European": [TEXT OF THE BIO]*

Eurosceptic. *You are an expert in European politics. Please classify the following X profile bio as "Eurosceptic" or "Not-Eurosceptic" according to whether the author of the text (who is from France) holds negative views of the European Union or not. The response should be in the form of a single term with the name of the category: "Eurosceptic" or "Not-Eurosceptic": [TEXT OF THE BIO]*

Conservative. *You are an expert in European politics. Please classify the following X profile bio as "Conservative" or "Not-Conservative" according to whether the author of the text (who is from France) holds conservative views or beliefs, including but not limited to negative views on abortion, gender equality, and same-sex marriage. The response should be in the form of a single term with the name of the category: "Conservative" or "Not-Conservative": [TEXT OF THE BIO]*

Liberal. *You are an expert in European politics. Please classify the following X profile bio as "Liberal" or "Not-Liberal" according to whether the author of the text (who is from France) holds liberal views or beliefs, including but not limited to positive views on abortion, gender equality, and same-sex marriage. The response should be in*

the form of a single term with the name of the category: "Liberal" or "Not-Liberal": [TEXT OF THE BIO]

Liberal immigration. *You are an expert in European politics. Please classify the following X profile bio as "Pro-Immigration" or "Not-Pro-Immigration" according to whether the author of the text (who is from France) holds liberal or positive views on immigration. The response should be in the form of a single term with the name of the category: "Pro-Immigration" or "Not-Pro-Immigration": [TEXT OF THE BIO]*

Restrictive immigration. *You are an expert in European politics. Please classify the following X profile bio as "Anti-Immigration" or "Not-Anti-Immigration" according to whether the author of the text (who is from France) holds conservative or negative views on immigration. The response should be in the form of a single term with the name of the category: "Anti-Immigration" or "Not-Anti-Immigration": [TEXT OF THE BIO]*

Pro environment. *You are an expert in European politics. Please classify the following X profile bio as "Pro-Environment" or "Not-Pro-Environment" according to whether the author of the text (who is from France) supports environmental protection, including but not limited to fighting against climate change, supporting economic de-growth, showing interest in the rights of animals, or adopting dietary restrictions such being vegan or vegetarian. The response should be in the form of a single term with the name of the category: "Pro-Environment" or "Not-Pro-Environment": [TEXT OF THE BIO]*

Climate denialist. *You are an expert in European politics. Please classify the following X profile bio as "Climate-Denialist" or "Not-Climate-Denialist" according to whether the author of the text (who is from France) denies or is skeptical about climate change, climate warning, or their human cause. The response should be in the form of a single term with the name of the category: "Climate-Denialist" or "Not-Climate-Denialist": [TEXT OF THE BIO]*

Economic focus. *You are an expert in European politics. Please classify the following X profile bio as "Economic-Focus" or "Not-Economic-Focus" according to whether the author of the text (who is from France) is interested in the economy, including but not limited to economic growth. The response should be in the form of a single term with the name of the category: "Economic-Focus" or "Not-Economic-Focus": [TEXT OF THE BIO]*

Nationalist. *You are an expert in European politics. Please classify the following X profile bio as "Nationalist" or "Not-Nationalist" according to whether the author of the text (who is from France) is a nationalist or a patriot, including but not limited to opposing multiculturalism, international organizations, or ethnic minority communities. The response should be in the form of a single term with the name of the category: "Nationalist" or "Not-Nationalist": [TEXT OF THE BIO]*

Cosmopolitan. *You are an expert in European politics. Please classify the following X profile bio as "Cosmopolitan" or "Not-Cosmopolitan" according to whether the author of the text (who is from France) is cosmopolitan, including but not limited to supporting multiculturalism, international organizations, international integration, and multilateralism. The response should be in the form of a single term with the name of the category: "Cosmopolitan" or "Not-Cosmopolitan": [TEXT OF THE BIO]*

Table 9: Characteristics of the different parties in the dataset. The number of MPs refers to the presence in our data. The last two columns indicate whether or not the party was included in the CHES surveys.

Acronym	**Name**	**MPs**	**CHES 2019**	**CHES 2023**
RE	Renaissance	193	✓	✓
LR	Les Républicains	172	✓	✓
RN	Rassemblement National	87	✓	✓
PS	Parti Socialiste	84	✓	✓
LFI	La France Insoumise	72	✓	✓
MoDem	Mouvement Démocrate	52	✓	✓
Horizons	Horizons	36	-	-
UDI	Union des Démocrates et Indépendants	33	-	✓
EELV	Europe Ecologie Les Verts	26	✓	✓
PCF	Parti Communiste Français	25	✓	✓
PRV	Parti Radical	14	-	-
LC	Les Centristes	7	-	-
PRG	Parti Radical de Gauche	2	-	✓
DLF	Debout La France	1	✓	✓

Table 10: Number of annotated bios, F1-score, Precision and Recall of the logistic regression models used for the validation. Sorted by decreasing F1-score.

Dimension	Year	Annotator	Label A	Label B	N_A	N_B	F1-score	Precision	Recall
lrgen	2019	human	left	right	1975	1593	0.965	0.976	0.953
lrgen	2019	human	right	left	1593	1975	0.957	0.944	0.971
lrecon	2019	human	left	right	1975	1593	0.957	0.972	0.941
lrecon	2023	human	left	right	1975	1593	0.951	0.977	0.926
lrecon	2019	human	right	left	1593	1975	0.948	0.930	0.967
lrecon	2023	human	right	left	1593	1975	0.942	0.913	0.973
eu_position	2023	human	pro_european	eurosceptic	880	536	0.940	0.962	0.919
eu_position	2019	human	pro_european	eurosceptic	880	536	0.938	0.961	0.916
refugees	2023	human	restrictive_immigration	liberal_immigration	211	161	0.933	0.946	0.919
immigrate_policy	2019	human	restrictive_immigration	liberal_immigration	211	161	0.933	0.946	0.919
galtan	2019	LLM	liberal	conservative	9519	4666	0.920	0.939	0.901
galtan	2023	LLM	liberal	conservative	9519	4666	0.919	0.940	0.898
immigrate_policy	2019	human	liberal_immigration	restrictive_immigration	161	211	0.915	0.898	0.932
refugees	2023	human	liberal_immigration	restrictive_immigration	161	211	0.915	0.898	0.932
sociallifestyle	2019	LLM	liberal	conservative	9519	4666	0.914	0.944	0.887
eu_position	2023	human	eurosceptic	pro_european	536	880	0.907	0.877	0.940
eu_position	2019	human	eurosceptic	pro_european	536	880	0.904	0.872	0.938
antielite_salience	2023	LLM	elite	populist	51070	5265	0.900	0.985	0.828
eu_position	2023	LLM	pro_european	eurosceptic	15846	5422	0.896	0.950	0.847
antielite_salience	2019	LLM	elite	populist	51070	5265	0.894	0.981	0.822
immigrate_policy	2019	LLM	restrictive_immigration	liberal_immigration	5262	3350	0.889	0.918	0.862
eu_position	2019	LLM	pro_european	eurosceptic	15846	5422	0.880	0.949	0.820
antielite_salience	2023	human	elite	populist	416	198	0.875	0.923	0.832
nationalism	2019	LLM	cosmopolitan	nationalist	10345	9300	0.873	0.864	0.882
antielite_salience	2019	human	elite	populist	416	198	0.871	0.902	0.841
lrgen	2019	LLM	left	right	20477	17125	0.870	0.901	0.842
refugees	2023	LLM	restrictive_immigration	liberal_immigration	5262	3350	0.869	0.911	0.831
lrecon	2019	LLM	left	right	20477	17125	0.864	0.910	0.823
lrecon	2023	LLM	left	right	20477	17125	0.859	0.913	0.810
lrgen	2019	LLM	right	left	17125	20477	0.856	0.825	0.889
nationalism	2019	LLM	nationalist	cosmopolitan	9300	10345	0.856	0.866	0.846
lrecon	2019	LLM	right	left	17125	20477	0.854	0.810	0.903
lrecon	2023	LLM	right	left	17125	20477	0.850	0.800	0.907
galtan	2019	LLM	conservative	liberal	4666	9519	0.846	0.813	0.881
galtan	2023	LLM	conservative	liberal	4666	9519	0.845	0.810	0.884
corrupt_salience	2019	LLM	elite	populist	51070	5265	0.841	0.976	0.738
sociallifestyle	2019	LLM	conservative	liberal	4666	9519	0.840	0.794	0.892
immigrate_policy	2019	LLM	liberal_immigration	restrictive_immigration	3350	5262	0.839	0.802	0.879
refugees	2023	LLM	liberal_immigration	restrictive_immigration	3350	5262	0.816	0.767	0.873
people_vs_elite	2019	LLM	elite	populist	51070	5265	0.812	0.961	0.704
antielite_salience	2023	human	populist	elite	198	416	0.773	0.707	0.854
antielite_salience	2019	human	populist	elite	198	416	0.755	0.708	0.808
eu_position	2023	LLM	eurosceptic	pro_european	5422	15846	0.751	0.660	0.870
eu_position	2019	LLM	eurosceptic	pro_european	5422	15846	0.726	0.623	0.870
people_vs_elite	2019	human	elite	populist	416	198	0.722	0.818	0.647
environment	2019	LLM	pro_environment	climate_denialist	28163	107	0.714	0.997	0.556
corrupt_salience	2019	human	elite	populist	416	198	0.699	0.836	0.601
environment	2019	LLM	pro_environment	economic_focus	28163	8387	0.661	0.858	0.538
corrupt_salience	2019	human	populist	elite	198	416	0.581	0.473	0.753
people_vs_elite	2019	human	populist	elite	198	416	0.571	0.484	0.697
antielite_salience	2023	LLM	populist	elite	5265	51070	0.495	0.345	0.878
antielite_salience	2019	LLM	populist	elite	5265	51070	0.474	0.329	0.848
environment	2019	LLM	economic_focus	pro_environment	8387	28163	0.431	0.311	0.700
corrupt_salience	2019	LLM	populist	elite	5265	51070	0.378	0.245	0.825
people_vs_elite	2019	LLM	populist	elite	5265	51070	0.314	0.201	0.722
environment	2019	LLM	climate_denialist	pro_environment	107	28163	0.011	0.005	0.626