The Reddit Politosphere: A Large-Scale Text and Network Resource of Online Political Discourse

Valentin Hofmann,[1,3] **Hinrich Schütze,**[3] **Janet B. Pierrehumbert**[2,1]

[1]Faculty of Linguistics, University of Oxford
[2]Department of Engineering Science, University of Oxford
[3]Center for Information and Language Processing, LMU Munich
valentin.hofmann@ling-phil.ox.ac.uk

Abstract

We introduce the Reddit Politosphere, a large-scale resource of online political discourse covering more than 600 political discussion groups over a period of 12 years. It is to the best of our knowledge the largest and ideologically most comprehensive dataset of its type now available. One key feature of the Reddit Politosphere is that it consists of both text and network data, allowing for methodologically-diverse analyses. We describe in detail how we create the Reddit Politosphere, present descriptive statistics, and sketch potential directions for future research based on the resource.

Introduction

The computational analysis of political discourse has attracted growing interest over the last few years (Garimella et al. 2018; Johnson and Goldwasser 2018; An et al. 2019; Demszky et al. 2019; Grover and Mark 2019; Guimaraes et al. 2019; Soliman, Hafer, and Lemmerich 2019; Davoodi, Waltenburg, and Goldwasser 2020; Jiang et al. 2020; Marchal 2020; Mokhberian et al. 2020; Rajadesingan, Resnick, and Budak 2020; He et al. 2021; Mendelsohn, Budak, and Jurgens 2021; Waller and Anderson 2021), a trend spurred both by the availability of suitable data from social media and by the development of increasingly refined methodological tools. Despite this progress, a comprehensive and easily-accessible resource for research into the online political landscape has so far been missing, resulting in studies mostly based on ad-hoc collected datasets (e.g., tweets crawled for a specific paper). This practice is detrimental in two crucial ways: it hinders replicability as well as comparability, and it is highly inefficient since similar data collections need to be created from scratch repeatedly.

In this paper, we introduce the Reddit Politosphere, a large-scale and diverse resource of online political discourse. The Reddit Politosphere is based on Reddit and covers more than 600 political discussion groups over a period of 12 years. It is to the best of our knowledge the largest and ideologically most comprehensive dataset of its type now available to the computational social sciences. One of the key features of the Reddit Politosphere is that it comprises both text and network data, which means that it is amenable to diverse methodological approaches. To facilitate research, we annotate the data in several ways and release scripts for easy data access. We publish the Reddit Politosphere on Zenodo at https://doi.org/10.5281/zenodo.5851729 and the accompanying collection of scripts in a GitHub repository at https://github.com/valentinhofmann/politosphere.

The remainder of this paper is structured as follows. We first provide an overview of related work, focusing particularly on research based on Reddit, studies on online political discourse, and the complementary roles of text and network data in the computational social sciences. We then present in detail our method of creating the Reddit Politosphere, followed by exploratory analyses that showcase potential avenues for future research. Finally, we provide a short conclusion and discuss ethical aspects of the resource.

Related Work

The Reddit Politosphere builds upon and is inspired by three strands of research: research using Reddit as a data resource, research on online political discourse, and research on jointly modeling text and network data.

Research Based on Reddit

Reddit is an online discussion platform where people can create communities, so-called subreddits, devoted to certain interests or topics. In contrast to other social media sites such as Facebook and Twitter, an almost complete set of Reddit comments is publicly available, the so-called Pushshift Reddit Dataset[1] (PRD; Baumgartner et al. 2020), and the platform has diverse and generally high-quality discussions. These characteristics have made Reddit a popular resource for studies in the computational social sciences (see Medvedev, Lambiotte, and Delvenne (2019) and Baumgartner et al. (2020) for overviews). There has also been a large number of studies in natural language processing based on Reddit (del Tredici and Fernández 2017, 2018; Stewart and Eisenstein 2018; Zayats and Ostendorf 2018; Kim, Kim, and Kim 2019; Hofmann, Pierrehumbert, and Schütze 2020a,b; Hofmann, Schütze, and Pierrehumbert 2020; Tigunova et al. 2020; Bagga, Piper, and Ruths 2021; Hada et al. 2021; Röttger and Pierrehumbert 2021).

[1]https://files.pushshift.io/reddit/comments

Research on Online Political Discourse

Social media platforms have become a key resource for research on political participation and discourse (see Haq et al. (2020) and Theocharis and Jungherr (2021) for overviews). Studies have investigated a wide range of phenomena including political polarization (Adamic and Glance 2005; Conover et al. 2011; Himelboim, McCreery, and Smith 2013; Weber, Garimella, and Batayneh 2013; Garcia et al. 2015; Garimella et al. 2018; Morales et al. 2019), ideological radicalization (Grover and Mark 2019; Crawford, Keen, and Suarez-Tangil 2021; Mamié, Ribeiro, and West 2021), and political framing (Demszky et al. 2019; Mendelsohn, Budak, and Jurgens 2021; Reiter-Haas, Kopeinik, and Lex 2021). Since many subreddits are political discussion groups (e.g., r/politics), sometimes even with explicit ideological orientation (e.g., r/Conservative, r/Liberal), Reddit has become a particularly popular resource for studies on online political discourse (Nithyanand, Schaffner, and Gill 2017; An et al. 2019; Grover and Mark 2019; Guimaraes et al. 2019; Soliman, Hafer, and Lemmerich 2019; Marchal 2020; Rajadesingan, Resnick, and Budak 2020; Waller and Anderson 2021). However, most previous work analyzing political discourse on Reddit has relied on a small number of hand-picked subreddits or external lists of political subreddits, which provides an only incomplete picture of Reddit's multifaceted political landscape.

Research Using Text and Network Data

Texts and networks, particularly social networks, are among the most important data structures for the computational social sciences and allow for complementary perspectives on social phenomena. Methodologically, texts and networks can be analyzed using tools from natural language processing (Jurafsky and Martin 2000) and graph analysis (Newman 2018). While many studies in the computational social sciences focus on only one type of data or analyze texts and networks separately, there has been growing interest in methods allowing for joint modeling (Yang and Eisenstein 2017; del Tredici et al. 2019; Mishra et al. 2019; Hofmann, Pierrehumbert, and Schütze 2021a,b). This approach requires large-scale, structured datasets covering both types of data.

Data Collection

We download all comments between January 2008 and December 2019 from PRD. January 2008 is chosen as the starting point since it is the month seeing the creation of the first political subreddits besides r/politics (specifically, r/Economics, r/obama, and r/ronpaul), a result of changes in Reddit's user policies. December 2019 is chosen as the end point since it constitutes a natural boundary to exclude the marked changes in political discourse due to Covid. This ensures a certain degree of homogeneity of the Reddit Politosphere. For the US (the major focus of political discourse on Reddit), the Reddit Politosphere thus covers the last year of the presidency of George W. Bush, the presidency of Barack Obama, and the first three years of the presidency of Donald Trump. In the following, we describe in detail our methodology to create the Reddit Politosphere based on PRD.

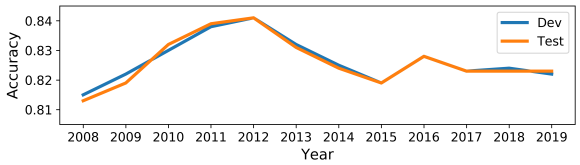


Figure 1: Political comment classification. The figure shows the dev and test accuracies of year-wise multinomial Naive Bayes classifiers trained to distinguish political from non-political comments. The dev set is used to tune the discounting parameter on the 2008 data, which might explain the better dev performance (compared to test) for the first two years. Random performance is 50% for all years.

Identifying Political Subreddits

What kinds of subreddits should be contained within the Reddit Politosphere? *We are looking for subreddits that have a focus on authentic political discourse.* This has two important implications: *focus* implies that a major part of the subreddit threads should be about politics, but subreddits are allowed to cover other topics (e.g., theology) as long as there is overall a clear connection to politics; *authentic* implies that subreddits should host genuine, bona fide political discussions, which excludes subreddits that only mimick such discussions (e.g., political game subreddits). Identifying subreddits that meet these criteria is a pivotal step for the dataset since it determines its coverage; e.g., if we systematically miss subreddits of certain political ideologies, the Reddit Politosphere might misrepresent the space of political discourse on Reddit. Prior studies have typically relied on manually identified political subreddits (An et al. 2019; Grover and Mark 2019; Guimaraes et al. 2019; Soliman, Hafer, and Lemmerich 2019; Marchal 2020) or drawn on a list[2] of political subreddits on r/politics (Nithyanand, Schaffner, and Gill 2017). While the first method does not scale to a larger number of subreddits, the second method has the disadvantage of staleness (e.g., the list on r/politics is no longer maintained and does not contain any recently-created subreddits). Similar to Rajadesingan, Resnick, and Budak (2020), we therefore adopt a machine learning approach to identifying political subreddits.

Specifically, we start by training simple classifiers on PRD to detect political comments. Since political discourse shifts over time, and hence a typical political comment in 2008 is likely to look different from a typical political comment in 2019, we train separate classifiers for each year. As positive examples, we take for each year all comments from r/Anarchism, r/Anarcho_Capitalism, r/Conservative, r/Libertarian, r/Republican, r/democrats, r/progressive, and r/socialism. These subreddits are chosen since they represent different points on the ideological spectrum and thus do not bias the classifiers towards certain political ideologies. As negative examples, we take for each year an equally-sized sample of comments from the default subreddits (a set of 54

[2]https://web.archive.org/web/20190103121805/https://www.reddit.com/r/politics/wiki/relatedsubs

Year	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{W}\rvert$	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{U}\rvert$
2008	1,297,931	47,744,745	9	33,612
2009	1,684,017	75,026,492	14	57,585
2010	3,028,168	140,083,326	25	129,262
2011	5,827,753	276,373,802	56	250,800
2012	9,755,532	463,256,964	86	417,599
2013	9,056,186	444,941,842	110	388,277
2014	8,000,183	396,876,595	135	308,944
2015	12,015,581	559,733,075	170	424,825
2016	51,110,048	1,760,795,288	259	889,656
2017	56,569,660	1,939,378,507	303	1,029,023
2018	59,443,354	2,031,852,970	323	1,172,706
2019	71,044,343	2,395,496,296	416	1,531,980

Table 1: Text data statistics. $\lvert\mathcal{T}\rvert$: number of texts (i.e., comments); $\lvert\mathcal{W}\rvert$: number of word tokens; $\lvert\mathcal{S}\rvert$: number of subreddits; $\lvert\mathcal{U}\rvert$: number of users.

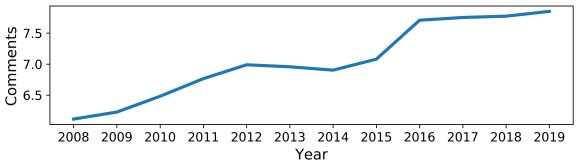


Figure 2: Number of comments in the Reddit Politosphere. The figure shows the number of comments (log scale) for each year of the Reddit Politosphere. The number of comments increases particularly strongly during years with presidential elections (2012 and 2016).

subreddits users used to be subscribed to automatically when joining Reddit). The default subreddits cover a broad range of topics (e.g., r/food, r/gaming, and r/sports) and are representative of the various interests present on Reddit besides politics.[3] We convert the resulting comments for each year (i.e., balanced sets of 50% political and 50% non-political comments) into bags of unigrams and split them into 80% train, 10% dev, and 10% test. We then train year-wise multinomial Naive Bayes classifiers (Manning, Raghavan, and Schütze 2008) with absolute discounting (Ney, Essen, and Kneser 1994) on the train comments and evaluate them on the test comments, using accuracy as the metric. We tune the discounting parameter on the dev comments of 2008 and take the best value for all years. We find that the performance of the classifiers is high for all years and lies between 81.3% in 2008 and 84.1% in 2012 (Figure 1).[4]

The next step is to draw on the trained classifiers (which operate on the level of *comments*) to decide for entire *subreddits* whether they are political or not. To do so, we first predict for all comments from PRD whether they are political or not, using the matching year-wise classifiers (e.g., the 2008 classifier for the 2008 comments). Based on these predictions, we classify a subreddit for a certain year as political if the ratio of political and non-political comments in that year is larger than two, i.e., the classifier predicts at least two thirds of the comments to be political. We choose this hyperparameter, which is more restrictive than in the comparable setup by Rajadesingan, Resnick, and Budak (2020), to ensure a high precision of the identified political subreddits. We further introduce a size threshold and require a political subreddit to have at least 1,000 comments and 100 users for the year in question. This procedure results in a preliminary list of 751 political subreddits.

Identifying subreddits meeting our criteria for political subreddits in an automated way has limits (e.g., we find several political game subreddits on the preliminary list). We therefore conduct a manual cleaning step. Specifically, the first author visits the web presence of each subreddit (or its archived form if the subreddit does not exist anymore) and decides whether it meets our criteria for political subreddits, resulting in the removal of 146 subreddits with either too little focus on politics (e.g., r/Ask_Lawyers) or unauthentic political content (e.g., r/ModelUSGov). We evaluate this step by having an independent annotator relabel 100 randomly selected subreddits from the preliminary list based on our criteria. Cohen's κ indicates high agreement with our decisions (0.78). To check the coverage of the final list of 605 political subreddits, we compare against the above-mentioned list of political subreddits on r/politics (specifically, the partisan subreddits) and find that all subreddits passing the size threshold are contained on the final list.

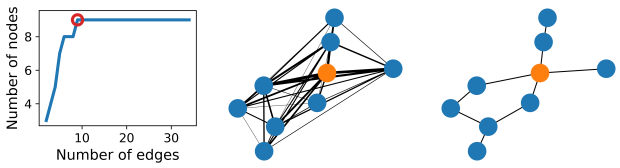
Extracting Text Data

We extract all comments for the 605 political subreddits from the downloaded portion of PRD. Table 1 provides year-wise summary statistics of the resulting text data. The number of extracted comments is growing over time (Figure 2). Notice that we extract comments even for years in which subreddits are not classified as political by the classifiers (because the percentage of political comments is too low, or because they do not pass the size threshold); e.g., we notice that r/The_Donald is classified as political in 2015, the first year of its existence, but as non-political in the years thereafter, which is indicative of its increasing radicalization, leading to its eventual banning in 2020 (Ribeiro et al. 2021). Extracting the comments for all years in such cases allows researchers to track the full evolution of the subreddits.

Besides the comment text and the subreddit, we extract all other data (e.g., exact creation time, score, etc.) contained within PRD; see Baumgartner et al. (2020) for details. To increase the anonymity of the data, we convert the usernames to random five-character pseudonyms and remove data fields that are specific to individual users and hence might be used to identify them. We also convert all mentions of usernames within the text to pseudonyms; we detect mentions of usernames by exploiting the common Reddit practice to prefix usernames with u/ and /u/. Mentions of usernames that are not active within the Reddit Politosphere are converted to a single pseudonym. The fact that the Reddit Politosphere is fully pseudonymized makes it possible to use it in settings

[3] As r/news and r/worldnews also contain political content, we remove them from the set of default subreddits.

[4] Given the generally high performance of the classifiers, we do not consider using n-grams of higher order (e.g., bigrams) or alternative classification methods (e.g., logistic regression).



(a) Edge pruning (b) Weighted (c) Unweighted

Figure 3: Network backboning. The figure illustrates how we create the unweighted network for the example year 2008. (a) plots the number of nodes as a function of the number of edges for the network from 2008 as we iteratively prune edges with the largest p-value. The knee point (red circle) is taken to construct an unweighted network (here corresponding to a significance level of 0.05). (b) shows the original weighted network from 2008, with edge width indicating log edge weight. (c) shows the unweighted network resulting from only keeping edges from (b) that fall below the significance level indicated in (a). The orange node represents r/politics, the largest political subreddit.

Year	$\lvert\mathcal{S}\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	μ_d	μ_π	ρ	Q
2008	9	9	9	2.00	2.31	.250	.309
2009	14	14	17	2.43	2.55	.187	.393
2010	25	21	28	2.67	2.54	.133	.438
2011	56	56	143	5.11	2.70	.093	.479
2012	86	86	271	6.30	2.78	.074	.496
2013	110	108	324	6.00	3.08	.056	.562
2014	135	132	335	5.08	3.86	.039	.673
2015	170	168	493	5.87	3.87	.035	.677
2016	259	254	1,315	10.35	3.13	.041	.593
2017	303	295	1,572	10.66	3.14	.036	.585
2018	323	316	1,604	10.15	3.17	.032	.592
2019	416	412	2,536	12.31	3.20	.030	.604

Table 2: Network data statistics. $\lvert\mathcal{S}\rvert$: number of subreddits; $\lvert\mathcal{V}\rvert$: number of nodes in network backbone; $\lvert\mathcal{E}\rvert$: number of edges; μ_d: average node degree; μ_π: average shortest path length; ρ: density; Q: maximum modularity.

where PRD might breach anonymity protocols (e.g., due to institutional ethics requirements). Since usernames might be valuable for certain research questions, we separately release information about their properties (see below).

To facilitate research using the Reddit Politosphere, we enrich the text data extracted from PRD in two ways. First, to make statistical analyses of the text more straightforward, we clean the comments by lowercasing all words, removing stopwords and URLs, and reducing repetitions of more than three characters (e.g., *niiiice*) to three characters (Han and Baldwin 2011). The cleaned text is included besides the raw text, allowing researchers to choose what they deem more appropriate. Second, we run a probabilistic language detector[5] on all comments and add the predictions to the data. While most comments in the identified political subreddits are English, comments in other languages do occur; the language information can thus be used for filtering.

Extracting Network Data

To capture latent associations between the political subreddits, we construct year-wise networks in which the nodes are subreddits, and the edges are based on patterns of user overlap between the subreddits. Subreddits that have disproportionately many users in common are likely to be ideologically or topically similar (Olson and Neal 2015; Datta, Phelan, and Adar 2017; Kumar et al. 2018), especially since Reddit users tend to be loyal to a small number of subreddits they identify with (Hamilton et al. 2017).

For each year, we first compute for every pair of political subreddits the number of users that posted at least 10 comments in both subreddits, defining a weighted network over the subreddits.[6] We then use backboning methods, specifically the noise-corrected filter[7] (Coscia and Neffke 2017), to transform the weighted networks into unweighted ones. Intuitively, a large weight (i.e., number of shared users) between two *large* subreddits is less indicative of latent associations than a large weight between two *small* subreddits. The noise-corrected filter takes such effects into account by assuming a binomial null model for the edge weight distribution, computing p-values for all edges, and only keeping edges below a signficance level as unweighted edges. To determine the significance level for each year, we follow common practice in network backboning (Serrano, Boguñá, and Vespignani 2009) and measure the ratio of kept edges versus kept nodes while iteratively removing edges with the largest p-value. The optimal significance level manifests itself as a knee point (Figure 3), which we identify by means of the Kneedle algorithm (Satopää et al. 2011). Like for the text data, we also include subreddits for years in which they are not classified as political due to too little political comments, but they need to pass the size threshold (since otherwise user overlap is not a robust measure of latent associations).

Table 2 provides year-wise summary statistics of the resulting networks. Notice that the number of nodes in the networks is slightly smaller than the number of subreddits in several years, which is an effect of network backboning (if all edges incident to a certain node are pruned during backboning, the node is removed from the network). We also measure various network properties such as average node degree, average shortest path length, and density (Newman 2018). To get an impression of the degree of fragmentation exhibited by the community structure of the networks, we use the Louvain method (Blondel et al. 2008) to compute for each year the partition with maximum modularity Q (Newman and Girvan 2004; Newman 2006), a widely employed measure of fragmentation (Waugh et al. 2009; Conover et al. 2011; Kirkland 2013; Soares, Recuero, and Zago 2018; Morán 2020). $Q > 0.3$ for all years, which has been previously used as a cut-off value to determine polarization (Garcia et al. 2015), indicating that the networks of the Reddit Politosphere are highly fragmented.

[5] https://github.com/CLD2Owners/cld2

[6] We exclude bots and automoderators (see below).

[7] https://www.michelecoscia.com/?page_id=287

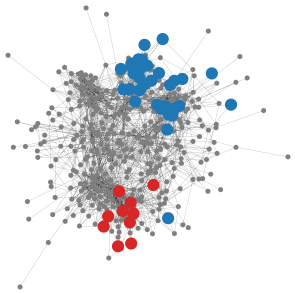


Figure 4: Democratic and republican clusters in the network of 2019. Each point is a subreddit. Blue points represent democratic, red points republican subreddits, respectively. Both groups form visibly distinct clusters in the network, which is confirmed by numerical analyses (see text). The democratic subreddit close to the republican subreddits is r/AskDemocrats, a subreddit explicitly inviting questions from outside the democratic party.

Extracting User Metadata

Usernames can be a valuable source of information for various research questions; e.g., they have been shown to be suggestive of user demographics including gender and ethnicity (Cornetto and Nowak 2006; Jaech and Ostendorf 2015; Wood-Doughty et al. 2018) as well as user ideology (Holt, Freilich, and Chermak 2020). The fact that the Reddit Politosphere is fully pseudonymized means that such information is lost. We therefore release certain properties of the usernames as separate metadata. The properties form sufficiently broad classes to keep pseudonymization intact.

As the first kind of information, many usernames identify a user as male or female by containing a gender-specific given name (e.g., u/john123).[8] To leverage this, we first obtain from the US Social Security Administration[9] lists of the most common male and female names for children born between 1980 and 1999, who constitute the vast majority of Reddit users (Shatz 2017). We extract all names that occur in all years with an absolute frequency of more than 100, resulting in a total of 1,404 names (659 male, 745 female). In the case of gender-neutral names, we only extract the gender with the higher frequency. We then check for each username whether it contains one of the gendered names and tag it accordingly. To prevent a large number of false positives, we only match names between scores and underscores (e.g., u/happy_john), names fitting the camel case pattern if applicable (e.g., u/HappyJohn), and names at the beginning or end of the username, ignoring numbers (e.g., u/happyjohn123). This results in 315,124 gender-tagged usernames (255,292 male, 59,832 female), roughly 10% of all usernames. We notice that the proportion of male and female users is even more skewed toward male users than reported in prior studies (Dou et al. 2015; Shatz 2017), which might be due to the specific topical areas on which the Reddit Politosphere builds; female users also tend to not self-report their gender to avoid online abuse (McLean and Griffiths 2019), which might further decrease the number of female users with clearly female usernames.

Furthermore, usernames often contain lexical elements implying a certain attitude or stance, which can be valuable information for certain research questions. Based on a manual, qualitative inspection of 1,000 randomly sampled usernames, we introduce the following seven categories, indicating the words used for identification:

- Negative attitude: *angry*, *rogue*, *troll*, *wtf* (7,764)
- Overt negation: *anti*, *downvote*, *fuck*, *stop* (9,734)
- Astro: *astro*, *cosm*, *rocket*, *space* (12,003)
- Dangerous: *beast*, *gorilla*, *shark*, *tiger*, *wolf* (14,450)
- Doom: *dead*, *death*, *doom*, *evil*, *zombie* (20,258)
- Military: *c(a)pt*, *colonel*, *commander*, *major*, *sgt* (10,384)
- Nobility: *duke*, *emperor*, *king*, *lord*, *sir* (19,720)

For overt negations, military titles, and titles of nobility, we require the usernames to start with one of the given words in order to minimize false positives. We also notice that there is a large cluster of usernames containing *trump* (10,868); we also indicate this in the metadata.

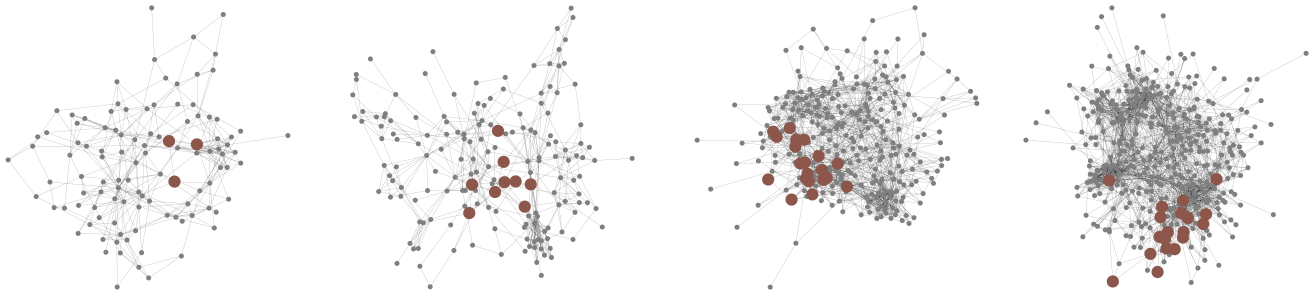
Finally, there are programmed users such as bots (5,803) and automoderators (134). We do not remove them in order to keep the discussion structure intact, but we annotate them in the metadata based on the usernames as well as a list[10] of bots on Reddit. This information can be used for filtering.

Extracting Subreddit Metadata

Besides users, we also provide metadata for the political subreddits. First, we indicate for each subreddit whether it has been banned (38; e.g., r/new_right). Furthermore, we give information about three frequent classes of subreddits: subreddits devoted to a politician (85; e.g., r/AlexandriaOcasio), subreddits explicitly in favor of the democratic (43; e.g., r/VoteBlue) or republican (20; e.g., r/republicans) party (either directly or by endorsing a partisan politician), and subreddits devoted to a certain region (101; e.g., r/AustraliaLeftPolitics). For regional subreddits, we further distinguish between Europe, Canada, Middle East, UK, US, and other parts of the world (mostly Asia and Australia). Since most subreddits in the Reddit Politosphere are about US politics, we only indicate a regional affiliation if a subreddit is dedicated to state-level US politics (e.g., r/NewJerseyuncensored). Finally, there is a cluster of subreddits devoted to the political issue of gun control (13; e.g., r/liberalgunowners); since this is much larger than for other political issues (e.g., there are only three abortion subreddits), we also indicate it in the metadata.

[8] All usernames are made up for demonstration purposes. It is important to note that the binary annotation as male or female ignores the fact that gender exists along a continuum, and intermediate realizations are possible. Thus, special care must be taken when interpreting results based on this part of the user metadata.

[9] https://www.ssa.gov/oact/babynames/limits.html



(a) 2013 (b) 2015 (c) 2017 (d) 2019

Figure 5: Network expansion of radical subreddits between 2013 and 2019. Each point is a subreddit. Brown points represent subreddits that are eventually banned. Radical subreddits form a relatively tight-knit cluster in the network that is expanding over time. Notice that the position within the network is also determined by the overall ideological leaning (see Figure 4): while most radical subreddits are far-right, the separate point in the year 2019 is r/MoreTankieChapo, a far-left subreddit.

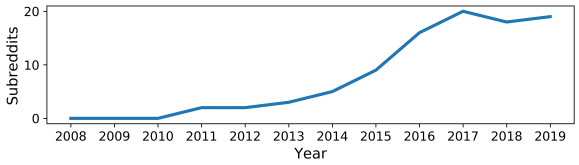


Figure 6: Number of radical subreddits in the Reddit Politosphere. The figure shows for each year the number of subreddits that are eventually banned.

Exploratory Analysis

What kinds of analyses are possible using the Reddit Politosphere? Here, we showcase the potential for future research based on the resource by combining the networks with information from the subreddit metadata.

One natural question to ask is to what extent the networks cluster according to different ideologies; the high modularity values show that the networks have a clustered structure, but it is not clear whether this is due to ideological variables. To analyze this, we draw upon the democratic and republican subreddits and perform two experiments on the largest network from the Reddit Politosphere (2019).

The first experiment focuses on the *internal* clustering of the two groups of subreddits. Specifically, for each group we loop over all nodes and compute (i) the shortest path length to all other nodes in the *same* group and (ii) the shortest path length to nodes in an equally-sized sample drawn randomly from the network. We then compute the average of (i) and (ii) across all nodes in each group. The values for (i) are much *smaller* than the values for (ii) for both groups (democrats: 2.11 versus 3.15, republicans: 1.89 versus 3.09). Both pairwise differences are shown to be highly significant by two-tailed t-tests ($p < 0.001$). Democratic and republican subreddits are much closer to each other in the network than would be expected by chance. The second experiment focuses on the *external* clustering of the two groups of subreddits. Specifically, for each group we loop over all nodes and compute (i) the shortest path length to all nodes in the *other* group and (ii) the shortest path length to nodes in an equally-sized sample drawn randomly from the network. We then compute the average of (i) and (ii) across all nodes in each group. This time, the values for (i) are much *larger* than the values for (ii) for both groups (democrats: 3.44 versus 3.14, republicans: 3.44 versus 3.04). Both pairwise differences are again shown to be highly significant by two-tailed t-tests ($p < 0.001$). Democratic and republican subreddits are further away from each other in the network than would be expected by chance. Thus, democratic and republican subreddits form two cohesive as well as distinct clusters within the network, which can also be seen by inspecting a plot of the network with nodes colored according to party affiliation (Figure 4).

Going one step beyond the democratic-republican polarization, we ask whether the networks of the Reddit Politosphere also reflect ideological radicalization. To examine this, we draw upon the set of subreddits banned by the creation time of the Reddit Politosphere. We first notice that the set of these radical subreddits constantly grew over time, following the overall trend of the Reddit Politosphere (Figure 6). How does this growth look like in terms of the network structure? Specifically, is there a radicalized region in the network that is expanding over time, or do radical subreddits originate at random points in the network?

To answer this question, we conduct a similar experiment as in the last section, but with a temporal perspective: we examine the period from 2013 to 2019 and compute in each year for all newly-originated radical subreddits the shortest path length to each of the already existing radical subreddits. We conduct the same computation for an equally-sized random sample of newly-originated non-radical subreddits. The average across all years is considerably lower for the radical than non-radical subreddits (2.08 versus 3.19), a difference that is found to be highly significant by a two-tailed t-test ($p < 0.001$), indicating that radical subreddits originate close to other radical subreddits in the network. Again, this can also be seen by inspecting plots of the networks

[10]https://www.reddit.com/r/autowikibot/wiki/redditbots

(Figure 5), suggesting the existence of radicalized regions in the networks of the Reddit Politosphere.

While this brief exploration has focused on analyzing the network structure by using information from the subreddit annotations, the rich structure of the Reddit Politosphere allows to address a wide variety of different research questions; e.g., one could easily expand the analysis above by including information from the subreddit comments or features based on the users of the subreddits.

Conclusion

We introduce the Reddit Politosphere, a large-scale resource of online political discourse covering more than 600 political discussion groups over a period of 12 years. The Reddit Politosphere consists of both text and network data, which makes it an attractive resource for approaches operating on both types of data. We hope that the Reddit Politosphere benefits work at the intersection of research on online political discourse, the computational social sciences, network analysis, and natural language processing.

Ethical Statement

This work is entirely based on public data. To protect the identity of individual users, the Reddit Politosphere is fully pseudonymized, which sets it apart from most other datasets based on Reddit, including PRD.

We expect the Reddit Politosphere to be particularly valuable for research on social and political questions, which bears the risk that skews and biases in the data impact the results and conclusions based on them. In this respect, we highlight that the user base of Reddit is not representative of the overall US population. Specifically, it is younger, more male, and more liberal (Shatz 2017). This needs to be kept in mind when using the Reddit Politosphere.

We further notice that we design the Reddit Politosphere to fully adhere to the FAIR principles: it is *findable* as it is made publicly available on a registered and indexed platform (Zenodo); it is *accessible* by its unique and persistent DOI as well as the descriptions provided in the metadata, including this paper; it is *interoperable* as it leverages open and machine-readable data formats (bz2, csv, and json); it is *reusable* as it is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Acknowledgements

This work was funded by the European Research Council (#740516) and the Engineering and Physical Sciences Research Council (EP/T023333/1). The first author was also supported by the German Academic Scholarship Foundation and the Arts and Humanities Research Council.

References

Adamic, L. A.; and Glance, N. 2005. The political blogosphere and the 2004 U.S. Election: Divided they blog. In *International Workshop on Link Discovery (LinkKDD) 3*.

An, J.; Kwak, H.; Posegga, O.; and Jungherr, A. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *International AAAI Conference on Web and Social Media (ICWSM) 13*.

Bagga, S.; Piper, A.; and Ruths, D. 2021. "Are you kidding me?": Detecting unpalatable questions on Reddit. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL) 16*.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift Reddit dataset. In *International AAAI Conference on Web and Social Media (ICWSM) 14*.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10): P10008.

Conover, M.; Ratkiewicz, J.; Francisco, M.; Goncalves, B.; Flammini, A.; and Menczer, F. 2011. Political polarization on Twitter. In *International AAAI Conference on Web and Social Media (ICWSM) 5*.

Cornetto, K. M.; and Nowak, K. L. 2006. Utilizing usernames for sex categorization in computer-mediated communication: Examining perceptions and accuracy. *CyberPsychology & Behavior* 9(4): 377–387.

Coscia, M.; and Neffke, F. M. 2017. Network backboning with noisy data. In *IEEE International Conference on Data Engineering (ICDE) 33*.

Crawford, B.; Keen, F.; and Suarez-Tangil, G. 2021. Memes, radicalisation, and the promotion of violence on Chan sites. In *International AAAI Conference on Web and Social Media (ICWSM) 15*.

Datta, S.; Phelan, C.; and Adar, E. 2017. Identifying misaligned inter-group links and communities. *Proceedings of the ACM on Human-Computer Interaction* 1: 1–23.

Davoodi, M.; Waltenburg, E.; and Goldwasser, D. 2020. Understanding the language of political agreement and disagreement in legislative texts. In *Annual Meeting of the Association for Computational Linguistics (ACL) 58*.

del Tredici, M.; and Fernández, R. 2017. Semantic variation in online communities of practice. In *International Conference on Computational Semantics (IWCS) 12*.

del Tredici, M.; and Fernández, R. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *International Conference on Computational Linguistics (COLING) 27*.

del Tredici, M.; Marcheggiani, D.; Schulte im Walde, S.; and Fernández, R. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in NLP. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019*.

Demszky, D.; Garg, N.; Voigt, R.; Zou, J.; Gentzkow, M.; Shapiro, J.; and Jurafsky, D. 2019. Analyzing polarization in social media: Method and application to Tweets on 21

mass shootings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*.

Dou, W.; Cho, I.; ElTayeby, O.; Choo, J.; Wang, X.; and Ribarsky, W. 2015. DemographicVis: Analyzing demographic information based on user generated content. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*.

Garcia, D.; Abisheva, A.; Schweighofer, S.; Serdült, U.; and Schweitzer, F. 2015. Ideological and temporal components of network polarization in online political participatory media. *Policy and Internet* 7(1): 46–79.

Garimella, K.; Morales, G. D. F.; Gionis, A.; and Mathioudakis, M. 2018. Quantifying controversy on social media. *ACM Transactions on Social Computing* 1(1): 1–27.

Grover, T.; and Mark, G. 2019. Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In *International AAAI Conference on Web and Social Media (ICWSM) 13*.

Guimaraes, A.; Balalau, O.; Terolli, E.; and Weikum, G. 2019. Analyzing the traits and anomalies of political discussions on Reddit. In *International AAAI Conference on Web and Social Media (ICWSM) 13*.

Hada, R.; Sudhir, S.; Mishra, P.; Yannakoudakis, H.; Mohammad, S. M.; and Shutova, E. 2021. Ruddit: Norms of offensiveness for English Reddit comments. In *Annual Meeting of the Association for Computational Linguistics (ACL) 59*.

Hamilton, W.; Zhang, J.; Danescu-Niculescu-Mizil, C.; Jurafsky, D.; and Leskovec, J. 2017. Loyalty in online communities. In *International AAAI Conference on Web and Social Media (ICWSM) 11*.

Han, B.; and Baldwin, T. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Annual Meeting of the Association for Computational Linguistics (ACL) 49*.

Haq, E. U.; Braud, T.; Kwon, Y. D.; and Hui, P. 2020. A survey on computational politics. *IEEE Access* 8: 197379–197406.

He, Z.; Mokhberian, N.; Câmara, A.; Abeliuk, A.; and Lerman, K. 2021. Detecting polarized topics in COVID-19 news using partisanship-aware contextualized topic embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Himelboim, I.; McCreery, S.; and Smith, M. 2013. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication* 18(2): 40–60.

Hofmann, V.; Pierrehumbert, J. B.; and Schütze, H. 2020a. DagoBERT: Generating derivational morphology with a pretrained language model. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*.

Hofmann, V.; Pierrehumbert, J. B.; and Schütze, H. 2020b. Predicting the growth of morphological families from social and linguistic factors. In *Annual Meeting of the Association for Computational Linguistics (ACL) 58*.

Hofmann, V.; Pierrehumbert, J. B.; and Schütze, H. 2021a. Dynamic contextualized word embeddings. In *Annual Meeting of the Association for Computational Linguistics (ACL) 59*.

Hofmann, V.; Pierrehumbert, J. B.; and Schütze, H. 2021b. Modeling ideological agenda setting and framing in polarized online groups with graph neural networks and structured sparsity. In *arXiv 2104.08829*.

Hofmann, V.; Schütze, H.; and Pierrehumbert, J. B. 2020. A graph auto-encoder model of derivational morphology. In *Annual Meeting of the Association for Computational Linguistics (ACL) 58*.

Holt, T. J.; Freilich, J. D.; and Chermak, S. M. 2020. Examining the online expression of ideology among far-right extremist forum users. *Terrorism and Political Violence* 1–21. ISSN 0954-6553.

Jaech, A.; and Ostendorf, M. 2015. What your username says about you. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2015*.

Jiang, J.; Chen, E.; Lerman, K.; and Ferrara, E. 2020. Political polarization drives online conversations about COVID-19 in the United States. *Human Behavior and Emerging Technologies* .

Johnson, K.; and Goldwasser, D. 2018. Classification of moral foundations in microblog political discourse. In *Annual Meeting of the Association for Computational Linguistics (ACL) 56*.

Jurafsky, D.; and Martin, J. H. 2000. *Speech and language processing*. Upper Saddle River, NJ: Prentice Hall.

Kim, B.; Kim, H.; and Kim, G. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*.

Kirkland, J. H. 2013. Hypothesis testing for group structure in legislative networks. *State Politics & Policy Quarterly* 13(2): 225–243.

Kumar, S.; Hamilton, W.; Leskovec, J.; and Jurafsky, D. 2018. Community interaction and conflict on the web. In *The Web Conference (WWW) 27*.

Mamié, R.; Ribeiro, M. H.; and West, R. 2021. Are anti-feminist communities gateways to the far right? Evidence from Reddit and YouTube. In *ACM Conference on Web Science (WebSci) 13*.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press.

Marchal, N. 2020. The polarizing potential of intergroup affect in online political discussions: Evidence from Reddit r/politics. In *SSRN: 3671497*.

McLean, L.; and Griffiths, M. D. 2019. Female gamers' experience of online harassment and social support in online gaming: A qualitative study. *International Journal of Mental Health and Addiction* 17(4): 970–994.

Medvedev, A. N.; Lambiotte, R.; and Delvenne, J.-C. 2019. The anatomy of Reddit: An overview of academic research. In *Dynamics On and Of Complex Networks (DOOCN) III*.

Mendelsohn, J.; Budak, C.; and Jurgens, D. 2021. Modeling framing in immigration discourse on social media. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2021*.

Mishra, P.; del Tredici, M.; Yannakoudakis, H.; and Shutova, E. 2019. Abusive language detection with graph convolutional networks. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*.

Mokhberian, N.; Abeliuk, A.; Cummings, P.; and Lerman, K. 2020. Moral framing and ideological bias of news. In *International Conference on Social Informatics (SocInfo) 12*.

Morales, A. J.; Dong, X.; Bar-Yam, Y.; and Pentland, A. 2019. Segregation and polarization in urban areas. *Royal Society Open Science* 6(10): 190573.

Morán, C. L. 2020. Cooperation and polarization in a presidential congress: Policy networks in the Chilean Lower House 2006–2017. *Politics* 40(2): 227–244.

Newman, M. 2006. Modularity and community structure in networks. *PNAS* 103(23): 8577–8582.

Newman, M. 2018. *Networks*. Oxford, UK: Oxford University Press.

Newman, M.; and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical Review E* 69: 026113.

Ney, H.; Essen, U.; and Kneser, R. 1994. On structuring probabilistic dependencies in stochastic language modelling. *Computer Speech and Language* 8: 1–38.

Nithyanand, R.; Schaffner, B.; and Gill, P. 2017. Online political discourse in the Trump era. In *arXiv 1711.05303*.

Olson, R. S.; and Neal, Z. P. 2015. Navigating the massive world of Reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science* .

Rajadesingan, A.; Resnick, P.; and Budak, C. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *International AAAI Conference on Web and Social Media (ICWSM) 14*.

Reiter-Haas, M.; Kopeinik, S.; and Lex, E. 2021. Studying moral-based differences in the framing of political tweets. In *International AAAI Conference on Web and Social Media (ICWSM) 15*.

Ribeiro, M. H.; Jhaver, S.; Zannettou, S.; Blackburn, J.; de Cristofaro, E.; Stringhini, G.; and West, R. 2021. Does platform migration compromise content moderation? Evidence from r/The_Donald and r/Incels. In *ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW) 24*.

Röttger, P.; and Pierrehumbert, J. B. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Satopää, V.; Albrecht, J.; Irwin, D.; and Raghavan, B. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *International Conference on Distributed Computing Systems (ICDCS) 31*.

Serrano, M. Á.; Boguñá, M.; and Vespignani, A. 2009. Extracting the multiscale backbone of complex weighted networks. *PNAS* 106(16): 6483–6488.

Shatz, I. 2017. Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review* 35(4): 537–549.

Soares, F. B.; Recuero, R.; and Zago, G. 2018. Influencers in polarized political networks on Twitter. In *International Conference on Social Media and Society (SMSociety) 9*.

Soliman, A.; Hafer, J.; and Lemmerich, F. 2019. A characterization of political communities on Reddit. In *ACM Conference on Hypertext and Social Media (HT) 30*.

Stewart, I.; and Eisenstein, J. 2018. Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018*.

Theocharis, Y.; and Jungherr, A. 2021. Computational social science and the study of political communication. *Political Communication* 38(1-2): 1–22.

Tigunova, A.; Yates, A.; Mirza, P.; and Weikum, G. 2020. RedDust: a large reusable dataset of Reddit user traits. In *International Conference on Language Resources and Evaluation (LREC) 12*.

Waller, I.; and Anderson, A. 2021. Quantifying social organization and political polarization in online platforms. *Nature* 600(7888): 264–268.

Waugh, A. S.; Pei, L.; Fowler, J. H.; Mucha, P. J.; and Porter, M. A. 2009. Party polarization in congress: A social networks approach. In *arXiv 0907.3509*.

Weber, I.; Garimella, K.; and Batayneh, A. 2013. Secular vs. islamist polarization in Egypt on Twitter. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2013*.

Wood-Doughty, Z.; Andrews, N.; Marvin, R.; and Dredze, M. 2018. Predicting Twitter user demographics from names alone. In *Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES) 2*.

Yang, Y.; and Eisenstein, J. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics* 5: 295–307.

Zayats, V.; and Ostendorf, M. 2018. Conversation modeling on Reddit using a graph-structured LSTM. *Transactions of the Association for Computational Linguistics* 6: 121–132.