Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

Political Discussion is Abundant in Non-political Subreddits (and Less Toxic)

Ashwin Rajadesingan, Ceren Budak, Paul Resnick
School of Information
University of Michigan, Ann Arbor
arajades,cbudak,presnick@umich.edu

Abstract

Research on online political communication has primarily focused on content in explicitly political spaces. In this work, we set out to determine the amount of political talk missed using this approach. Focusing on Reddit, we estimate that nearly half of all political talk takes place in subreddits that host political content less than 25% of the time. In other words, cumulatively, political talk in non-political spaces is abundant. We further examine the nature of political talk and show that political conversations are less toxic in non-political subreddits. Indeed, the average toxicity of political comments replying to a out-partisan in non-political subreddits is even less than the toxicity of co-partisan replies in explicitly political subreddits.

1 Introduction

Casual everyday political conversations are central to a vibrant deliberative democracy. Through these conversations, individuals learn new perspectives, form informed opinions and update their preferences (Kim and Kim 2008). These interactions may take place in explicitly political spaces such as city townhalls and civic committees but also in seemingly non-political spaces such as book readings, workplaces and social gatherings (Conover and Miller 2018). Importantly, this kind of everyday political talk is significantly correlated with opinion quality and political participation which are central to forming a well-informed electorate (Wyatt, Katz, and Kim 2000). In this work, we explore this phenomenon online, particularly studying political discussions in communities on Reddit that are not explicitly political.

Most research on political discussions has primarily focused on explicitly political spaces, examining communities around political news groups, figures or ideologies (Soliman, Hafer, and Lemmerich 2019; Himelboim, Gleave, and Smith 2009; An et al. 2019). However, survey research suggests that most people encounter political content online not in explicitly political spaces but in hobby and leisure groups where politics is incidental to the conversation (Wojcieszak and Mutz 2009). Further, recent years have seen increased political engagement among the electorate perhaps due to high levels of partisanship (Huddy, Mason, and Aarøe 2015) and growing social movements (Bosi, Giugni, and Uba 2016) such as Black Lives Matter. This heightened level of political engagement can also be observed online. For example, on Reddit, many communities that would not be typically construed as being 'political' such as r/EDM and r/MaleFashionAdvice protested against the platform's hate speech policies and police brutality in the US.[1] Further, in recent years, scholars have observed increasing politicization of typically non-political spaces (Dagnes 2019). The most prominent example is the politicization of emerging science and technology where inherent uncertainties are harnessed by political actors to cast doubt on the existence of scientific consensus (Bolsen and Druckman 2015). On Reddit, this phenomenon manifests in the formation of multiple communities on the same topic along partisan lines, for example, r/China_Flu and r/Coronavirus (Zhang et al. 2021). Thus, these developments call for expanding analysis of political discussions outside of typical political communities to communities that aren't explicitly political.

It is important to note that expanding the study of political discussions to include non-political spaces does not merely increase the volume of discussions for analysis. The dynamics of political discussions in these spaces may also be fundamentally different. Political discussions in these spaces may be moderated not by partisan identity but by participants' shared non-political interests and identities that drew them to the same community in the first place (Gaertner and Dovidio 2011). Thus, we might expect political conversations in non-political spaces, including cross-partisan ones, to be less toxic. However, shared non-political group identity may fail to offset, and might even exacerbate, the animosity generated by partisan identity (Klar 2018). Further, norms in these non-political spaces may not be designed to foster political discourse. Indeed, there may be norms against having political conversations at all, and thus when they occur, they may be even more toxic.

In this work, we focus on two primary questions: (i) What is the prevalence of political discussions in communities that are not explicitly political? (ii) Are cross-partisan political discussions in these spaces less toxic than ones in explicitly political spaces? We estimate that 49.26% $\pm 3.59\%$ of all political discussions on Reddit takes place in communi-

[1] https://www.theverge.com/2020/6/3/21279601/reddit-dark-subreddits-protest-police-violence-racism-hate-speech-policies

ties that host political discussions less than 25% of the time. This finding is not simply the result of a few very large non-political communities hosting some political content. It is instead due to a long tail of small communities that host some political content each. Our toxicity analysis reveals that political conversations in non-political spaces, including cross-partisan political interactions, are indeed less toxic than such interactions in political spaces. Interestingly, we find that there is an uptick in toxicity levels when talking politics, but even with this increase, the toxicity levels in non-political subreddits are still much lower than the toxicity in political subreddits.

2 Background

Political scientists have long highlighted the presence and importance of casual political talk in everyday social interactions taking place in spaces that are not explicitly political (see (Conover and Miller 2018) for a review). In fact, research suggests that most political conversations take place at work or with neighbors, with more than 70% of American survey respondents reporting that they have never or only rarely even attended public meetings explicitly designed for political discussions (Conover, Searing, and Crewe 2002). Similarly online, early survey research suggests that most people encounter political talk in message boards and chatrooms designed not for political discussions but for hobby and leisure related discussions (Wojcieszak and Mutz 2009). Thus, research limited to studying only political discussion spaces may overlook other spaces where a significant amount of such interactions may be taking place. Such everyday political talk, although not always deliberative and conducive to rational-critical argumentation, have important positive outcomes such as increased political knowledge (Pattie and Johnston 2008), political participation (Searing et al. 2007), refined opinions (Kim and Kim 2008) and higher tolerance (Pattie and Johnston 2008; Mutz 2006).

Recently, scholars have analyzed political discussions taking place in online "third spaces" a term derived from sociologist Ray Oldenburg's conceptualization of the 'third place', referring to public spaces outside of work and home such as cafes, parks and libraries where people meet and interact informally, fostering community ties and political participation (Wright 2012; Oldenburg 2001). Graham et al. (2015) found that political discussions in the three UK-based non-political forums they analyzed were as likely to emerge from non-political, personally-oriented discussions as from discussions that were about politics from the start, with users explicitly linking their personal experiences to public policy. In contrast to discussions in political spaces, they found that the discursive culture in these discussions centers around help and support rather than being competitive and combative. Yan et al. (2018) found that the political arguments made on transnational online cricket forums were typically short, unsubstantiated by external sources and occasionally uncivil. However, there was high exposure to cross-cutting political discussions with some engagement with opposing views in the form of question exchange and mutual acknowledgement. Analyzing a reality television discussion forum, Graham found that most political exchanges were driven by users' life experiences representing a more "lifestyle-oriented, personal form of politics" (Graham 2012). While exhibiting deliberative features such as the exchange of reasoned claims (as opposed to assertions) and reciprocity, participants also employed affirming, supportive and empathetic communicative practices fostering genuineness and civility in the discussions.

Political discussions in these non-political spaces may also be more civil and social compared to discussions in explicitly political communities. A significant factor contributing to hostility commonly observed in online political discussions is the increased levels of affective polarization (Hutchens, Hmielowski, and Beam 2019), the tendency of partisans to view opposing partisans negatively and co-partisans positively (Iyengar and Westwood 2015). This increased out-party animosity is explained by Social Identity Theory which argues that by merely categorizing individuals into groups (here, Republicans and Democrats), group identities are activated, creating an 'us' versus 'them' group dynamic (Tajfel et al. 1979). Crucially, unlike race, gender and other protected attributes where group-related behaviors are mediated by strong social norms (and laws) against discrimination, there are no norms that temper hostility towards out-partisans (Iyengar et al. 2019). In fact, the open hostility displayed by political elites towards their political opponents demonstrates that such behavior is appropriate (Banda and Cluverius 2018). Given the social identity underpinnings of affective polarization, researchers have explored ways to offset partisan identity drawing from prior research on intergroup conflict. One successful approach to reducing out-partisan animosity is by priming a superordinate identity. Based on the Common Ingroup Identity Model (Gaertner and Dovidio 2011), Levendusky showed that priming Republicans and Democrats to think of each other as Americans rather than outgroup members recategorized them as being part of the same common ingroup, resulting in reduced animosity and warmer attitudes towards each other (Levendusky 2018). Although our study is not a direct test of this theory, we expect that interactions in non-political subreddits likely increase the salience of shared common non-political group memberships. This may mediate how cross-partisan political discussions are conducted in these spaces.

Though not conclusive, the prior literature provides two compelling arguments: (i) political conversations are abundant in non-political spaces. (ii) quality of discourse in these conversations may be different and in some cases, better than political conversations in explicitly political communities. In this work, we assess these claims empirically in the context of Reddit. First, we quantify the relative contribution of non-political subreddits to the overall political content on Reddit. In this aspect, our work is similar to Munson et al.'s work on estimating the prevalence of political content in non-political blogs. Analyzing a sample of blogs from Blogger.com, they found that "25% of all political posts are from blogs that post about politics less than 20% of the time" (Munson and Resnick 2011). Second, we examine a specific marker of conversation quality: toxicity in cross-partisan political interactions. Scholars have suggested that political talk in these third spaces are likely to

be less polarized, since users participate in these spaces because of shared interests such as a soccer team or fast fashion which are not aligned politically (Wojcieszak and Mutz 2009; Wright, Graham, and Jackson 2015). Thus, mediated by shared non-political identities, these spaces could facilitate respectful and civil cross-partisan interactions. In this work, we examine this hypothesis by quantifying the toxicity levels of cross-partisan political discussions in non-political spaces and comparing them to toxicity levels in other settings on Reddit.

3 Reddit Data

Reddit, a collection of communities of varied and diverse topics, provides us with an ideal platform to examine the prevalence of political content in non-political spaces. We use the PushShift Reddit dataset (Baumgartner et al. 2020) to perform our analysis on comments posted from 2016 to 2019. We exclude comments from subreddits that have hosted less than 1000 comments over the four years. We also remove comments from known bots and moderators from the analysis. To allow for robust estimation of political prevalence, we only consider comments which are 50 characters or more in length in this analysis. In total, we examined 2.8 billion comments posted in 30,899 subreddits.

4 Estimating the Prevalence of Political Content in Non-political Spaces

Our basic approach to estimating the prevalence of political content is to train a classifier that yields, for each comment, a predicted probability that it would be judged as political by a panel of three MTurk raters. If the classifier's outputs are properly calibrated, the average of those outputs for all the comments in a subreddit is an estimate of the prevalence of political content in the subreddit.

Our training data consists of a sample of 10,000 comments, each rated by three people on MTurk as either political or not. We do not use these labels to directly train a classifier that predicts what how each item will be labeled. Following the quantification approach (Forman 2006; González et al. 2017), if the goal is to estimate prevalence rather than to correctly classify individual comments, it can be more effective to use ground-truth data to perform calibration on a crudely trained classifier than to use up the training data on improving the classifier.

Section 4.1 describes a process for training a classifier to distinguish between comments from two baskets of subreddits, one of which consists of a hand-selected set of subreddits that are overtly political. This text-based classifier outputs a probability that the comment is from one of the political subreddits. The middle of Figure 1 shows the distribution of classifier outputs for all comments from the two baskets of subreddits. Some comments from the political subreddits contain phrases that are more common in the other subreddits. Such comments get a low score from the classifier. However, most comments originating in the political subreddits get higher scores (the blue distribution, on top) and most comments originating in the other subreddits get lower scores (the orange distribution, below).

The classifier does not have perfect accuracy. Moreover, it may make different kinds of errors on content from different subreddits. So we do not directly use it to classify and count political comments in each subreddit. The second step, as described in section 4.2, is to build two calibration curves for the classifier, one based on human ratings of a sample of comments from the set of known political subreddits and the other based on human ratings for comments from other subreddits. Each calibrator provides a mapping from a classifier output stratum (e.g., 0.5-0.6) to a calibrated forecast, the fraction of comments that are political when the classifier gives an output in that stratum. The right side of Figure 1 shows the two calibration curves. For all classifier outputs below 0.9, comments originating in the political subreddits were more likely to be judged as political, as indicated by the gap between the blue calibration curve and the orange one.

Section 4.3 then describes a process for selecting a calibrator to use for a particular subreddit. Section 4.4 describes how to use the calibrator to generate an estimate of the fraction of political content in that subreddit. Finally, section 4.5 describes how we combine the estimates for individual subreddits to yield overall prevalence.

4.1 Training a Classifier

We built an L1-regularized logistic regression classifier trained on bigrams and trigrams from a random sample of 500,000 comments from known political subreddits (positive or "politics" class) and 500,000 comments from all other subreddits (negative or "not politics" class). We used a list of 277 political subreddits provided by (Rajadesingan, Resnick, and Budak 2020) and updated the list to include more recently created subreddits supporting Democratic primary candidates such as r/YangForPresidentHQ and r/JoeBiden before the 2020 US presidential election.

Assessed through 5-fold cross-validation, we obtained an accuracy of 81.56% with a false positive rate of 14.41% and a false negative rate of 22.45%. Note that the false positive and false negative rates are for predicting the source of a comment, not whether the comment itself is truly political. While this classifier performed reasonably well in identifying content from political subreddits, it was trained not on political and non-political comments but on *comments from political and non-political subreddits*. This is particularly problematic since our goal is to estimate the fraction of *political comments* in non-political subreddits. If the classifier were perfectly accurate at distinguishing between content from the two types of subreddits, and we used it as if it were distinguishing political comments, it would tell us that there were zero political comments in non-political subreddits, which might or might not be the case. Thus, a further calibration step is needed in order to estimate the error rates of this classifier at predicting whether a comment is truly political, and then adjust for those error rates in our prevalence estimates.

4.2 Building Calibrators

The right side of Figure 1 outlines the calibration process. We use the classifier to produce a probability estimate for

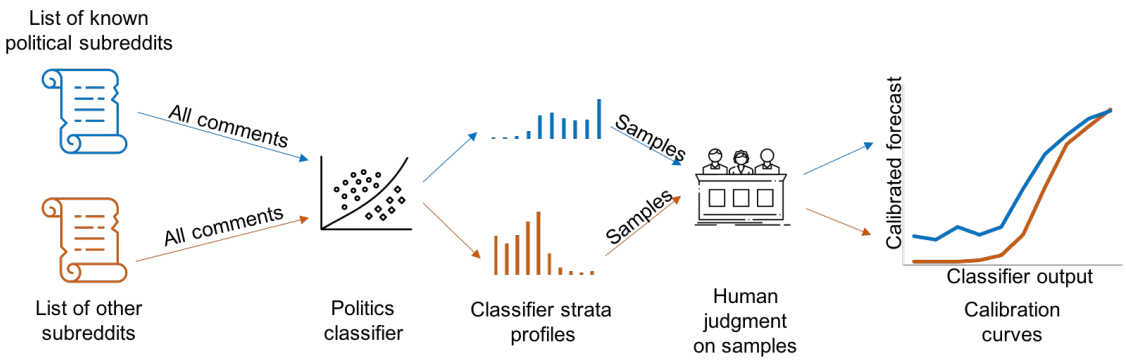

Figure 1: Training a classifier that distinguishes between comments from political and non-political subreddits, then calibrating it to produce predictions of whether comments are political.

Strata	Political subreddits				Non-political subreddits			
	Pop. proportion $W_{k,\text{pol}}$	Samples	Political% $p_{k,\text{pol}}$	Std dev. $s_{k,pol}$	Pop. proportion $W_{k,\text{nonpol}}$	Samples	Political% $p_{k,\text{nonpol}}$	Std dev. $s_{k,nonpol}$
1	0.004	50	0.180	0.054	0.148	615	0.021	0.006
2	0.007	50	0.160	0.052	0.117	797	0.024	0.005
3	0.017	50	0.240	0.060	0.150	1239	0.023	0.004
4	0.047	107	0.187	0.038	0.199	1810	0.031	0.004
5	0.145	346	0.237	0.023	0.242	2296	0.057	0.005
6	0.165	394	0.475	0.025	0.083	788	0.189	0.014
7	0.129	295	0.695	0.027	0.026	237	0.485	0.032
8	0.116	241	0.821	0.025	0.013	107	0.757	0.041
9	0.119	204	0.917	0.019	0.008	61	0.869	0.043
10	0.252	263	0.970	0.011	0.012	50	0.980	0.020

Table 1: Population proportions, Neyman samples, percent of samples identified as political by human judgment and standard deviation per strata for political and non-political subreddits. The standard deviations for both political and non-political subreddits are lower in the strata where their population proportions are higher. This effect is by design (using Neyman allocation) to ensure that the confidence intervals for the prevalence estimates are lower.

each comment. Then, we allocate comments into ten strata, 0-10%, 10-20%, etc., based on the classifier outputs. That yields a profile of classifier strata: the proportion of comments that fall into each stratum. We compute two separate classifier strata profiles, one based on comments from the list of known political subreddits, the other based on comments from other subreddits. As might be expected, the classifier assigns many more comments from the political subreddits to the higher strata (higher probability of being political).

Then, we calibrate the classifier outputs against human judgments of the comments, separately for comments from each source. Below we first describe the human judgment process and then explain the rationale and details behind each of the steps in the calibration process.

Human judgments When asked to identify topics they considered political from a list, Fitzgerald (2013) found systematic demographic differences with partisans, liberals and men identifying significantly more topics as political compared to non-partisans, conservatives and women respectively. In order to reduce such differences in interpretation, Fitzgerald suggests providing human raters with an explicit definition to follow.

For this work, we modify (Moy and Gastil 2006)'s political discussion definition, which is predominantly based on political issues, to also include references to political figures, parties and institutions. We consider a comment to be political if it is about (i) political figures, parties and institutions, (ii) Broad cultural and social issues (e.g., civil rights, moral values, and the environment), (iii) National issues (e.g., healthcare, welfare policy, and foreign affairs), (iv) Local and state concerns (e.g., school board disputes and sales taxes) or (v) neighborhood and community affairs (e.g., decisions about a neighborhood watch crime prevention program).

Even with an explicit definition, however, whether a particular comment is political or not is open to interpretation. Conceptually, we take the ground truth classification of a comment to be the label that the majority of people who ever read online comments would apply to that comment, if they all were asked to judge it according to the explicit definition. Of course, this ground truth is a counterfactual; no such survey of all readers of comments can ever be conducted for any comment. Instead, we rely on a proxy for this ground truth, a survey of three raters on Amazon Mechanical Turk. To elicit

high quality labels, we limit the task to crowdworkers with high performance in prior tasks [2] who also correctly labeled at least 4 out of 5 items in a qualification task where raters were shown sample comments and were asked to identify if the comment was political or not. Such qualification tasks are shown to improve crowdsourcing label quality (Budak, Goel, and Rao 2016).

The inter-rater agreement score as computed by Krippendorff's alpha was 0.55. While an alpha score of 0.55 is just below the threshold used for conventional content analysis, this agreement is relatively higher compared to other cases of crowd coding (e.g. (Lind, Gruber, and Boomgaarden 2017)). The most common outcome (66.85%) was for a comment to be unanimously labeled as not political. An additional 10.81% of comments were unanimously labeled as political. The remainder were split decisions: 7.75% were labels as political by two of three raters and 14.59% by one of three raters.

Following common practice in treatment of crowd labels, our primary analysis treats a comment as truly political if two or three of the raters label it as such. Given the relatively low agreement among raters, for robustness, we also report analyses in the Appendix that treat a comment as political if any of the three label it as political, or only if all three label it as political. The different aggregation strategies produced largely similar results and provide additional informative bounds on our estimates.

Classifier strata We group comments into ten strata based on the classifier probability output. For example, stratum 1 consists of all comments whose classifier output is between 0 and 0.1; stratum 10 consists of all comments whose classifier output is between 0.9 and 1. By stratifying comments into multiple relatively homogeneous groups based on classifier probability, we require fewer samples to estimate true prevalence per stratum as within-group variance reduces in more homogeneous groups (Cochran 1977).

One calibration for each subreddit type We expect the per-stratum prevalence estimates to be different in different subreddits. That is, comments in the 60-70% stratum in political subreddits could be judged 70% of time to be political subreddits, while this number can be, say, 45% in non-political ones. This would not be a concern if we were trying to estimate the overall prevalence of political comments, since we could just estimate the classifier's error rates on a random sample of comments. However, our task demands accurate estimates of political prevalence in each subreddit; if the classifier is more prone to err on content from the stratum that originates in one subreddit than another, it would throw off our cumulative prevalence estimate.

Estimating separate per-stratum prevalence rates for each subreddit is practically infeasible as it would require human judgments for samples from each subreddit. Instead, we compute per-stratum error rates separately for each *subreddit type*: a sample of comments from known political subreddits and a sample of comments from other subreddits (same as those used to train the classifier).

Optimum strata sampling for human judgments For each subreddit type, we must sample comments from each stratum for human judgments to obtain stratum specific prevalence estimates. The more comments we sample from a stratum, the less variance there will be in our estimate of the true prevalence of political comments in that stratum. Intuitively, fewer samples should be taken from a stratum with very few comments for calibration. A high variance in our estimated prevalence of political comments in such a stratum will not affect our overall estimate very much because it affects very few comments. Formally, for a fixed number of comments that we can afford to send for human rating, Neyman allocation provides the optimal allocation strategy which minimizes the variance of the overall prevalence estimate. Under Neyman allocation, the number of samples allocated to each stratum is given by:

$$n_k = n \frac{W_k * S_k}{\sum_{i=1}^{K} W_i * S_i}$$

- n is the total number of comments to be rated
- K is the number of strata (10 in our case)
- n_k is the number of comments to sample from the k-th stratum
- W_k is the weight of the k-th stratum in the classifier strata profile, i.e. the fraction of comments that are in that stratum
- S_k is the standard deviation of stratum k.

Note that $S_k = \sqrt{P_k(1-P_k)}$ is unknown before sampling, where P_k is the political prevalence in stratum k. Instead, we use our best estimate, the mean of the range limits of each stratum to calculate the approximate standard deviation expected in each stratum ($P_k = 0.05$ for Stratum 1, $P_k = 0.15$ for Stratum 2 and so on). Since our aim is to accurately estimate prevalence for each subreddit type, we perform separate stratified samplings, choosing two different W_k for each stratum k to match the overall comment proportions over strata (classifier strata profiles) for the two subreddit types.

We modified the Neyman allocation to include a minimum threshold to sample at least 50 comments in each stratum. We added this threshold to reflect the relative uncertainty in our initial estimates of W_k and S_k. We had a total budget for rating 10,000 comments. We used $n = 2000$ comments from political and $n = 8000$ for other subreddits. The rationale behind the uneven breakdown is that there are likely fewer political comments in non-political subreddits, meaning that a similar sized confidence interval for both estimates would result in significantly larger levels of relative uncertainty for prevalence estimates in non-political spaces. The fraction of comments that fall into each stratum (the classifier strata profile) are shown in the "Pop. proportions" columns of Table 1 and graphically as a histograms in Figure 1. The number of comments selected per stratum are reported in the "Samples" columns of Table 1.

[2] Raters must be based in the US, previously completed 1000 tasks and have at least 98% acceptance rate on the tasks that they have previously completed.

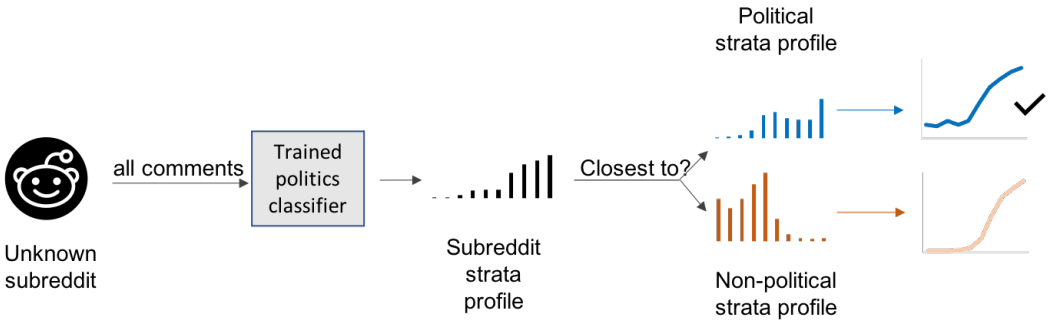

Figure 2: Selecting the political or nonpoltical calibrator depending on if the subreddit strata profile is similar to the political or non-political strata profile

Results of Stratum-specific prevalence estimation using human judgments The prevalence estimates per stratum from labeling comments for the two subreddit types are shown in Table 1 under "Political %" ($p_{k,\text{pol}}$ and $p_{k,\text{nonpol}}$). They are also shown graphically in the calibration curves in Figure 1, where x-axis is the stratum (classifier output) and y-axis is the calibrated prevalence in that stratum.

4.3 Selecting a Calibrator

Given that the prevalence estimates for the same strata are quite different for the two subreddit types, it is important to determine which set of prevalence estimates to use for each subreddit. Figure 2 outlines this process. For each subreddit, we obtain the classifier probabilities of all its comments to build its strata profile. If the subreddit strata profile is more similar to the known political strata profile than to the other subreddits strata profile, we use the calibration curve of the known political subreddits, else we use the other calibration curve. We use the Jenson-Shannon divergence (JSD) to make these comparisons between profiles. Lower JSD values imply higher similarity between distributions.

$$\text{diff}_{subr} = JSD(D_{subr}||D_{pol}) - JSD(D_{subr}||D_{nonpol})$$

$$f(subr) = \begin{cases} \text{political calibrator}, & \text{if diff}_{subr} \leq 0 \\ \text{non-political calibrator}, & \text{otherwise} \end{cases}$$

where D_{subr}, D_{pol} and D_{nonpol} are strata profiles of subreddit $subr$, political and non-political subreddits respectively.

The political and non-political strata profiles shown in Figure 2 correspond to the actual prevalence estimates reported in Table 1. Comparing the two profiles, we expect that subreddits with strata profiles that are either uniformly distributed or are peaked at around the middle strata will have smaller diff_{subr} scores, leading to potentially incorrectly assigning calibrators for those subreddits. Reassuringly, we find that less than 2.5% of all subreddits that we analyze have an absolute diff_{subr} less than 0.1, for comparison, the distance between the two strata profiles ($JSD(D_{pol}||D_{nonpol})$) is 0.40. We include the heatmap of subreddits based on $JSD(D_{subr}||D_{pol})$ and $JSD(D_{subr}||D_{nonpol})$ scores in Figure 5 in the Appendix. Further, since in this work it is especially important to not overestimate the prevalence of political content in non-political subreddits, we experimented with a more conservative approach of assigning subreddits to the non-political calibrator (detailed in the Appendix section B); we did not find a major difference in the prevalence estimates using this more conservative approach.

4.4 Corrected "Classify and Count" Estimation

We quantify the prevalence of political content in each subreddit $subr$ according to the proportion of content in each stratum and the corresponding forecast from the calibration curve. We calculate the estimated prevalence of political content in a subreddit (subr) as:

$$p_{\text{subr}} = \sum_{k=1}^{K} W_{k,\text{subr}} * p_{k,f(subr)}$$

where $W_{k,\text{subr}}$ is the proportion of total comments in stratum k for the subreddit and $p_{k,f(subr)}$ is prevalence estimate of the k-th stratum of the calibrator selected by $f(subr)$.

4.5 Estimates of Cumulative Counts of Political Comments

We estimate the total prevalence of political content on Reddit as the weighted sum of the prevalence in each subreddit.

$$p = \sum \left(\frac{N_{\text{subr}}}{N} \right) * p_{\text{subr}}$$

where N_{subr} is the total comments in subreddit $subr$ and N is the total comments across all of the subreddits.

We can estimate the variance of this prevalence estimate by combining the variance estimates across strata. The weight for each stratum is computed from the fraction of comments that the classifier assigns to that stratum. For political subreddits:

$$s^2_{\text{pol}} = \sum_{k=1}^{K} \left(\frac{N_{k,pol}}{N} \right)^2 * s^2_{k,pol}$$

where

$$N_{k,pol} = \sum_{\text{subr} \in \text{pol}} N_{k,\text{subr}}$$

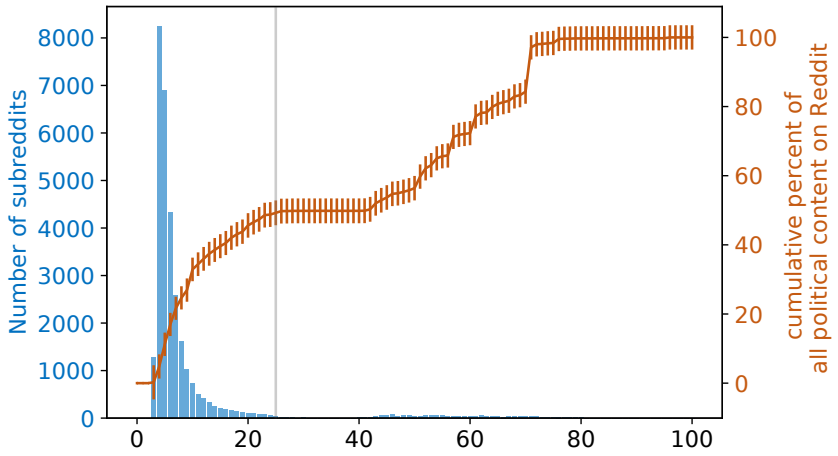

Figure 3: The line graph shows the cumulative percentage of all political comments posted in subreddits that host political comments less than $x\%$ of the time. The bar graph shows the number of subreddits that host political comments $x\%$ of the time. The grey line marks the 25% threshold we use to identify subreddits as political or non-political.

$N_{k,pol}$ is the sum of the comments in each stratum k for subreddits similar to the political strata profile. $s^2_{k,pol}$ is the variance estimated for political strata profiles from Table 1. Similarly, we calculate s^2_{nonpol} for non-political subreddits. The overall variance is just the sum.

$$s^2 = s^2_{\text{pol}} + s^2_{\text{nonpol}}$$

Finally, we define subreddits that are not explicitly political as those that host fewer than some threshold y percentage of political content and calculate the aggregate prevalence and variance of political content in all subreddits that host less than y% of political content. A higher cutoff of y will, of course, treat more subreddits as non-political and thus yield a higher estimate of the proportion of all political content that is in non-political subreddits.

4.6 Prevalence Estimation Results

In total, we estimate that 12.84% $\pm$0.45% of all comments on Reddit are political. To study the prevalence of political content in subreddits that are not explicitly political we construct Figure 3. The blue histogram shows the frequency of subreddits with x-coordinate percentage of political content. Of the 1399 subreddits whose classifier strata profile was closer to the profile for known political subreddits, almost all (99.71%) were estimated to have 40% or more political content. Of the 29,500 subreddits that were closer to other classifier strata profile, almost all (99.79%) were estimated to have less than 25% political content.

Each point on the orange line graph represents the cumulative percent of all political content on Reddit contributed by subreddits that host political comments less than x-coordinate percent of time. We find that 49.26% $\pm$3.59% of all political content on Reddit are from subreddits that host political content less than 25% of the time. Most subreddits on Reddit host very little political content, but cumulatively the non-political subreddits contribute nearly half of all political comments. This could be driven by the few most popular non-political subreddits having far more comments overall than the political subreddits. We examine this possibility by removing the top 10 non-political subreddits (see Table 3 in the Appendix) that contribute the most political comments. After removing these subreddits, we find that, similar to our original estimates, about 44.82% $\pm$3.42% of all political content on Reddit are from subreddits that are not explicitly political. These results suggest that the large fraction of political content in non-political subreddits is primarily driven by a large number of relatively small subreddits that each host a small percentage of political content. Robustness checks using different human judgment aggregation strategies and calibrator selection approaches yield similar estimates (see Appendix sections A and B).

5 Quantifying Toxicity of Cross-partisan Political Discussions in Non-political Spaces

Our main goal in this section is to identify the toxicity levels of cross-partisan discussions on political topics in non-political spaces. Our secondary goal is to compare that toxicity to toxicity observed in other settings to better contextualize our findings. To do so, we determine how toxicity on Reddit varies according to the following attributes: (i) political leaning of the users participating in the discussion, (ii) nature of the discussion, and (iii) type of the subreddit where the conversation is taking place.

For (i), we define a cross-partisan discussion as a left leaning user replying to a right leaning user or vice-versa. Our analysis is focused on the *reply* comments for each parent-reply discussion pair as the parent comment could be directed at a co-partisan or may not be directed at anyone if it is a top-level comment. For (ii), we rely on our calibrated classifier to determine the probability of a reply being political. Finally, for (iii), we classify any subreddit that contains

fewer than 25% political content as not being explicitly political as per Section 4.6.

5.1 Identifying Political Leaning of Users

To identify political leaning of users, we adopt a simple heuristic similar to ones that have been used in prior Reddit political studies (An et al. 2019; Soliman, Hafer, and Lemmerich 2019). First, we identify the well known subreddits r/politics, r/Liberal, r/progressive as left-leaning and r/The_Donald, r/Conservative, r/Republican as right-leaning. Then, we identify a user as left leaning only if all three of the following conditions are satisfied:

1. They post more comments in left-leaning subreddits than right-leaning subreddits.
2. The mean karma points score of their comments in left-leaning subreddits is higher than their mean score in right-leaning subreddits.
3. Their mean karma score in left-leaning subreddits is greater than 1. 1 is the default score that any comment receives on Reddit. So, a higher than 1 karma score implies that the comment has met net approval by the community.

Similarly, we identify right leaning users. Among users who posted at least once in these subreddits, we have 1,223,229 left leaning and 367,363 right leaning users. We cannot identify political leanings of other users and do not include them in this analysis.

5.2 Quantifying Toxicity of Replies

We use the Perspective toxicity classifier to identify toxicity of a comment. The classifier provides the probability of a comment being toxic, defined as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion" (Wulczyn, Thain, and Dixon 2017). The Perspective classifier has been used in prior Reddit studies (Mittos et al. 2020; Xia et al. 2020). Research evaluating its performance on comments from political communities shows that, on average, its toxicity classification is comparable to a single human judgment of toxicity (Rajadesingan, Resnick, and Budak 2020). We have toxicity classifier probabilities only for comments posted in 2016 and 2017, so we limit this analysis to comments posted in that time interval [3].

We calculate the probability of a reply comment r being toxic and political TP_r as:

$$TP_r = toxicity(r) * political(r)$$

where $toxicity(r)$ is the toxicity probability given by the Perspective classifier and $political(r)$ is the probability that the comment is political, which is calculated using the calibrated classifier. Similarly, we calculate the probability of the reply comment r being toxic and not political TNP_r as:

$$TNP_r = toxicity(r) * (1 - political(r))$$

[3]We use the 5th version of the Perspective classifier

5.3 Comparing Toxicity Between Discussion Pairs

Our aim is to compare the mean toxicity levels of cross-partisan political interactions to toxicity levels in other settings. Replies from the same subreddit are clustered to perform semi-pooling. We conduct mixed effects logistic regression using the $lme4$ package (Bates et al. 2015) modeling the toxicity of replies with a random effect for subreddits. The count of toxic replies is modeled as the number of *successes* and the total replies as the number of Bernoulli *trials* in a binomial distribution. We estimate the following 3-way interaction model:

$$\begin{aligned} T_{s,polreply,cross} &= Binomial(P(toxicity), \\ &\quad N_{s,polreply,cross}) \\ P(toxicity) &= logit(\alpha_s + \beta_1 polsub + \beta_2 polreply \\ &\quad + \beta_3 cross + \beta_4 polsub * polreply \\ &\quad + \beta_5 polsub * cross \\ &\quad + \beta_6 polreply * cross \\ &\quad + \beta_7 polsub * polreply * cross) \end{aligned}$$

where, $polsub$ is an indicator variable for whether the subreddit s is political, $polreply$ denotes whether the reply is political, $cross$ represents whether the reply is directed at an out-partisan. For each subreddit s, we identify the total number of replies ($N_{s,polreply,cross}$) for each $(polreply, cross)$ combination and the number of replies in $N_{s,polreply,cross}$ that are toxic ($T_{s,polreply,cross}$). We quantify the total political cross-partisan replies and number of such replies that are toxic in subreddit s as:

$$N_{s,polreply=1,cross=1} = \sum_{r \in XR_s} political(r)$$

$$T_{s,polreply=1,cross=1} = \sum_{r \in XR_s} TP_r$$

where XR_s is the set of all cross-partisan replies in subreddit s. Similarly, we quantify the non-political co-partisan replies and number of such replies that are toxic in s as:

$$N_{s,polreply=0,cross=0} = \sum_{r \in CR_s} (1 - political(r))$$

$$T_{s,polreply=0,cross=0} = \sum_{r \in CR_s} TNP_r$$

where CR_s is the set of all copartisan replies in subreddit s. Similarly, we calculate $N_{s,polreply,cross}$ and $T_{s,polreply,cross}$ for all other combinations of $(polreply, cross)$.

Table 2 shows the number of replies for each $(polreply, polsub, cross)$ combination used to estimate the binomial model. Upon estimating the model, we are most interested in (i) comparing the average toxicity levels of cross-partisan political discussions in non-political spaces to such discussions in political spaces. We further (ii) compare the average toxicity levels of political and non-political cross-partisan interactions in non-political spaces, since any

Interaction type	Conversation type	Political subreddits	Non-political subreddits
Copartisan	Non-political	13.83M	57.55M
	Political	21.48M	6.52M
Cross-partisan	Non-political	3.72M	20.51M
	Political	6.03M	3.17M

Table 2: Discussion data (in millions) for model estimation.

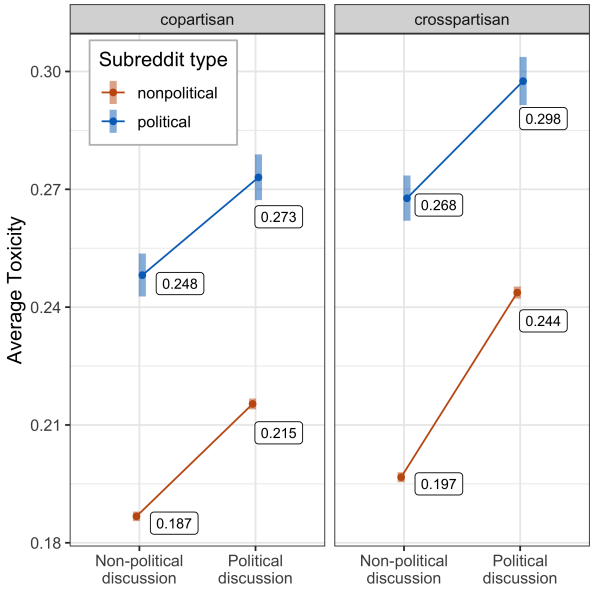

Figure 4: Interaction plot modeling the toxicity of discussions. All observed pairwise differences are statistically significant at .05 level.

large increase in toxicity levels when talking politics has important implications for the health of non-political communities. Finally, we (iii) compare average toxicity levels of cross-partisan and co-partisan interactions as a sanity check. We would expect to see a higher level of toxicity in cross-partisan interactions than in co-partisan ones.

5.4 Toxicity Analysis Results

Since the coefficients in a three-way interaction are hard to interpret, we present our results in the form of interaction plots. In Figure 4, y-axis is the average toxicity of replies and x-axis is the indicator variable on whether the discussion is political or not. Orange and blue lines represent the toxicity levels in non-political and political subreddits respectively. Left and right subgraphs represent discussions between co-partisans and cross-partisans respectively.

First, we focus on the plot on the right side of Figure 4, considering only cross-partisan replies. Answering our primary research question, we find that cross-partisan replies are significantly less toxic in non-political subreddits (24.4% toxic) compared to such interactions in political subreddits (29.8% toxic). This holds both for political and non-political content. There is also a main effect of content type, with political discussions being more toxic than non-political ones. Finally, there is an interaction effect: the difference in toxicity between political and non-political discussion is significantly larger in non-political subreddits. Still, cross-partisan political replies in non-political subreddits are less toxic than even co-partisan ones, in political subreddits. A similar pattern holds for copartisan replies. However, co-partisan replies are less toxic than cross-partisan replies in all settings.

6 Discussion

Political discussions on Reddit, much like face-to-face discussions, appear to crop up incidentally in various social settings or communities (Conover and Miller 2018). We found that political discussions, while by definition uncommon in non-political subreddits, are cumulatively abundant. There are many more non-political subreddits than political ones. Due to their sheer number, non-political subreddits *cumulatively* host nearly half of all political comments on Reddit.

This result suggests the need to diversify where researchers are looking when they try to understand the nature of political discussion online. Importantly, political discussions were not limited to only a few big non-political communities. Rather, we found a large number of small-sized non-political subreddits that had occasional political comments. This surely poses an important challenge. Given the sheer scale of content in non-political communities, this necessarily requires building classifiers to accurately identify political content across a wide variety of communities. As is evident from our calibration exercise, commonly used classifiers generally include bias, making this task daunting.

Political conversations in non-political spaces not only add to the volume of total political discourse but also are qualitatively different from conversations in political communities. We find compelling evidence to support the theory that cross-partisan political discussions are indeed much less toxic in non-political spaces than such discussions are in political spaces (Wright, Graham, and Jackson 2015). There are multiple potential explanations for this finding. First, political discussions in non-political communities are more likely to be moderated by shared group identity (Levendusky 2018) and social ties (Birchall 2020) instead of partisan identity which may reduce cross-partisan animosity (Gaertner and Dovidio 2011). Second, in general, the toxicity levels observed in non-political communities are much lower and it is likely that these low toxicity norms moderate and temper the tendency to indulge in harsh rhetoric in cross-partisan interactions (Iyengar and Westwood 2015). Regardless of the cause, our findings pose an important caution for researchers: simply aggregating political discussions from political and non-political communities may obfuscate the differences in the types of conversations in these spaces.

There is one important nuance in our toxicity findings. While cross-partisan political discourse is indeed less toxic in non-political spaces, it is significantly more toxic than non-political discourse in the same non-political spaces. Thus, these conversations may have adverse effects on non-political communities. More research is required to understand the consequences of political interactions in these

spaces. Further, we observe a larger increase in toxicity levels when talking politics in these spaces than when talking politics in political spaces. We speculate that the norms, rules and the style of moderation in place to foster conducive topic-specific conversations in non-political communities may not be as effective in handling toxicity stemming from cross-partisan political discussions. Further, a political comment in a non-political space can also be seen as a norm violation, leading to aggression from other community members. Alternately, the smaller increase in toxicity levels in political subreddits could indicate a ceiling effect; the toxicity levels of non-political discussions in political subreddits may already be so high that they are near the upper limits of how toxic the discussions can be.

Finally, there are important open questions regarding how these political conversations in atypical spaces fit into the "deliberative system" and how they ought to be studied. The deliberative system consists of both formal spaces such as legislatures and townhalls as well as informal spaces such as social gatherings and online political discussions in social media sites (Parkinson and Mansbridge 2012). Recently, deliberation theorists have highlighted the importance of everyday talk as a web that interconnects these diverse deliberation sites, urging empirical researchers to study discussions wherever they happen (Mansbridge 1999; Conover and Miller 2018). Future work on how ideas, frames and narratives transition from these spaces to more explicitly political deliberation sites both online and offline will provide important insights on the role and importance of these conversations in non-political spaces.

7 Limitations and Future Work

The approach we followed to estimate prevalence is an improvement over a conventional classify and count approach in two important ways, but is still imperfect. The first improvement is that we employ a calibration process to map probabilistic outputs of the classifier into calibrated forecasts of the frequency of political comments. The second improvement is that, rather than assuming that the classifier performs equally well on comments originating in different subreddits, we separately calibrate the classifier on two samples of comments, one from known political subreddits and one from other subreddits. Indeed, we do find that the same classifier score is much more likely to indicate a political comment when the comment comes from a political subreddit, and this dual calibrator approach allows us to appropriately lower estimates of the prevalence of political comments in non-political subreddits.

The approach, however, is still imperfect. First, while creating two calibrators is better than one, there could be more than two types of subreddits, with the classifier having a different error profile for each. Second, our process for selecting the calibrator for each subreddit, by comparing its classifier strata profile to that of the known political subreddits and that of other subreddits, may itself be error prone. We have taken a conservative approach, with more subreddits using the calibrator that yields lower counts than the number of subreddits that are eventually classified as non-political based on their counts. This avoids overestimating the political content in non-political subreddits, but may undercount the political content in political subreddits.

In the first step, we use a simple n-grams based logistic regression classifier as opposed to using word-embeddings or deep learning techniques. Developing a more accurate classifier generally will improve the effectiveness of the stratified sampling since each stratum is likely to be more homogeneous, leading to smaller confidence intervals for the overall prevalence estimate (see (Kumar and Raj 2018) for a similar argument using a simulation analysis). In our particular case, since we train the classifier on comments from political subreddits rather than on political comments, the gains from using a more accurate classifier are likely tempered by the extent to which comments from political subreddits accurately approximate political comments. Future research examining the gains of using more accurate classifiers in combination with calibrators will refine prevalence estimation techniques. Finally, our robustness checks suggest that the relatively low levels of agreement between raters did not majorly affect our prevalence estimates. Yet, raters disagreeing frequently on what is political indicates scope for improvement in the labeling process, perhaps by providing training examples and exercises.

In the toxicity analysis, we also did not perform a similar calibration process of the Perspective API. While a previous study showed that it was reasonably accurate on content from political subreddits (Rajadesingan, Resnick, and Budak 2020), there was insufficient data provided for calibration, and we do not know whether the error profile of the Perspective API is different on content originating in political vs. non-political subreddits. Numerous other factors, in addition to the type of subreddit, political nature of the comment, and partisanship, can dictate toxicity of responses on Reddit (e.g. toxicity of the parent comment). In our analysis, we used a simplified model to only account for the select attributes of interest to our study. Finally, while Reddit is a popular online forum for political discussions, it surely is not the only one. Future work that determines the role non-political communities play in driving political discourse on other platforms can help political communication scholars better identify spaces to pay attention to.

8 Conclusion

The subreddits where political comments are uncommon cumulatively produce almost 50% of all political comments. This is true even when we estimate prevalence based on conservative classifier calibrations. This large cumulative prevalence is not because of the volume of political comments in a few large non-political subreddits; instead, it is driven by a large number of non-political subreddits that host occasional political conversations. Importantly, political comments in non-political spaces seem to be less toxic on average. Thus, scholars looking at the promise and perils of online political deliberation would do well to focus their attention on political discussions that occur in venues that are not primarily organized to encourage political discussion.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1717688. Ashwin Rajadesingan is supported by a Facebook Fellowship. We thank the ARC-TS team at Michigan for Cavium-Thunderx Hadoop cluster support. We also thank the anonymous reviewers for their invaluable feedback on this work.

Appendix

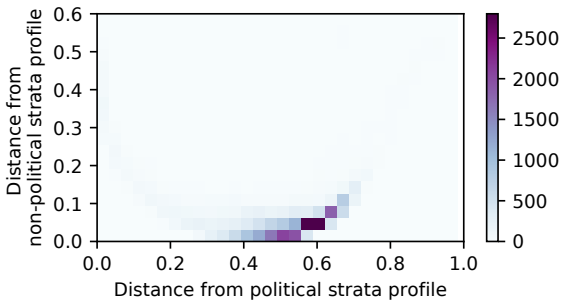

Figure 5: Density of subreddits by $JSD(D_{subr}||D_{pol})$ and $JSD(D_{subr}||D_{nonpol})$ scores. Few subreddits are equidistant from both strata profiles; most are close to the non-political strata profile.

Subreddit	Political percent	Political comments
AskReddit	9.97	15,151,228
pics	19.15	2,979,427
todayilearned	16.57	2,537,532
unpopularopinion	22.49	1,995,731
videos	12.01	1,632,210
funny	10.03	1,614,550
nba	5.73	1,481,271
nfl	6.32	1,373,542
soccer	6.71	1,223,417
AdviceAnimals	20.94	1,080,184

Table 3: Non political subreddits that contain the most political comments

A Robustness of Prevalence Estimates Varying Label Aggregation Strategy

We estimate prevalence with two other aggregating strategies: (i) 'any one' strategy: comment is political if at least one rater labels it as political. (ii) 'all three' strategy: comment is political if all three raters label it as political. We estimate that 54.56% $\pm$3.28% and 42.28% $\pm$3.95% of the overall political content are from subreddits that are not explicitly political, based on the 'any one' and 'all three' strategies respectively.

B Robustness of Prevalence Estimates Varying Political Subreddit Identification Strategy

To ensure robustness of our prevalence estimates, we employ another identification strategy. We relax the diff_{subr} criterion such that subreddits near the decision boundary will be calibrated using the political calibration curve. With this strategy, the prevalence estimates in political subreddits are expected to be higher.:

$$f(s) = \begin{cases} \text{political calibration curve,} & \text{if } \text{diff}_{subr} \leq 0.1 \\ \text{non-political calibration curve,} & \text{otherwise} \end{cases}$$

From this, we find that 43.29 $\pm$3.40% of overall political content are from subreddits that are not explicitly political.

References

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*, volume 13.

Banda, K. K.; and Cluverius, J. 2018. Elite polarization, party extremity, and affective polarization. *Electoral Studies* 56: 90–101.

Bates, D.; Mächler, M.; Bolker, B.; and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles* 67(1).

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*.

Birchall, C. 2020. Trying not to fall out: the importance of non-political social ties in online political conversation. *Information, Communication & Society* 23(7): 963–979.

Bolsen, T.; and Druckman, J. N. 2015. Counteracting the politicization of science. *Journal of Communication* 65(5).

Bosi, L.; Giugni, M.; and Uba, K. 2016. *The consequences of social movements*. Cambridge University Press.

Budak, C.; Goel, S.; and Rao, J. M. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly* 80(S1): 250–271.

Cochran, W. G. 1977. *Sampling Techniques, 3rd Edition*. John Wiley & Sons, 3rd edition edition.

Conover, P. J.; and Miller, P. R. 2018. Taking everyday political talk seriously. *The Oxford handbook of deliberative democracy* .

Conover, P. J.; Searing, D. D.; and Crewe, I. M. 2002. The deliberative potential of political discussion. *British journal of political science* 21–62.

Dagnes, A. 2019. Us vs. Them: Political Polarization and the Politicization of Everything. In *Super Mad at Everything All the Time*, 119–165. Springer.

Fitzgerald, J. 2013. What does "political" mean to you? *Political Behavior* 35(3): 453–479.

Forman, G. 2006. Quantifying trends accurately despite classifier error and class imbalance. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 157–166.

Gaertner, S. L.; and Dovidio, J. F. 2011. Common ingroup identity model. *The encyclopedia of peace psychology* .

González, P.; Castaño, A.; Chawla, N. V.; and Coz, J. J. D. 2017. A review on quantification learning. *ACM Computing Surveys (CSUR)* 50(5): 1–40.

Graham, T. 2012. Beyond "political" communicative spaces: Talking politics on the Wife Swap discussion forum. *Journal of Information Technology & Politics* 9(1): 31–45.

Graham, T.; Jackson, D.; and Wright, S. 2015. From everyday conversation to political action: Talking austerity in online 'third spaces'. *European Journal of Communication* .

Himelboim, I.; Gleave, E.; and Smith, M. 2009. Discussion catalysts in online political discussions: Content importers and conversation starters. *Journal of computer-mediated communication* 14(4): 771–789.

Huddy, L.; Mason, L.; and Aarøe, L. 2015. Expressive partisanship: Campaign involvement, political emotion, and partisan identity. *American Political Science Review* 109(1).

Hutchens, M. J.; Hmielowski, J. D.; and Beam, M. A. 2019. Reinforcing spirals of political discussion and affective polarization. *Communication Monographs* 86(3): 357–376.

Iyengar, S.; Lelkes, Y.; Levendusky, M.; Malhotra, N.; and Westwood, S. J. 2019. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science* 22: 129–146.

Iyengar, S.; and Westwood, S. J. 2015. Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science* 59(3): 690–707.

Kim, J.; and Kim, E. J. 2008. Theorizing dialogic deliberation: Everyday political talk as communicative action and dialogue. *Communication Theory* 18(1): 51–70.

Klar, S. 2018. When common identities decrease trust: An experimental study of Partisan women. *American Journal of Political Science* 62(3): 610–622.

Kumar, A.; and Raj, B. 2018. Classifier risk estimation under limited labeling resources. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 3–15. Springer.

Levendusky, M. S. 2018. Americans, not partisans: Can priming American national identity reduce affective polarization? *The Journal of Politics* 80(1): 59–70.

Lind, F.; Gruber, M.; and Boomgaarden, H. G. 2017. Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication methods and measures* 11(3): 191–209.

Mansbridge, J. 1999. Everyday Talk in the Deliberative System. In *Deliberative Politics: Essays on Democracy and Disagreement*, 1–211. Oxford University Press.

Mittos, A.; Zannettou, S.; Blackburn, J.; and Cristofaro, E. D. 2020. Analyzing Genetic Testing Discourse on the Web Through the Lens of Twitter, Reddit, and 4chan. *ACM Transactions on the Web (TWEB)* 14(4): 1–38.

Moy, P.; and Gastil, J. 2006. Predicting deliberative conversation: The impact of discussion networks, media use, and political cognitions. *Political Communication* 23(4).

Munson, S.; and Resnick, P. 2011. The prevalence of political discourse in non-political blogs. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Mutz, D. C. 2006. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.

Oldenburg, R. 2001. *Celebrating the third place: Inspiring stories about the great good places at the heart of our communities*. Da Capo Press.

Parkinson, J.; and Mansbridge, J. 2012. *Deliberative systems: Deliberative democracy at the large scale*. Cambridge University Press.

Pattie, C. J.; and Johnston, R. J. 2008. It's good to talk: Talk, disagreement and tolerance. *British Journal of Political Science* 38(4): 677–698.

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*.

Searing, D. D.; Solt, F.; Conover, P. J.; and Crewe, I. 2007. Public discussion in the deliberative system: does it make better citizens? *British Journal of Political Science* .

Soliman, A.; Hafer, J.; and Lemmerich, F. 2019. A characterization of political communities on reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*.

Tajfel, H.; Turner, J. C.; Austin, W. G.; and Worchel, S. 1979. An integrative theory of intergroup conflict. *Organizational identity: A reader* 56(65): 9780203505984–16.

Wojcieszak, M.; and Mutz, D. 2009. Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement? *Journal of communication* .

Wright, S. 2012. From "third place" to "third space": Everyday political talk in non-political online spaces. *Javnost-the public* 19(3): 5–20.

Wright, S.; Graham, T.; and Jackson, D. 2015. Third space, social media, and everyday political talk. In *The Routledge companion to social media and politics*, 74–88. Routledge.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399.

Wyatt, R. O.; Katz, E.; and Kim, J. 2000. Bridging the spheres: Political and personal conversation in public and private spaces. *Journal of communication* 50(1): 71–92.

Xia, Y.; Zhu, H.; Lu, T.; Zhang, P.; and Gu, N. 2020. Exploring Antecedents and Consequences of Toxicity in Online Discussions: A Case Study on Reddit. *Proceedings of the ACM on Human-Computer Interaction* .

Yan, W.; Sivakumar, G.; and Xenos, M. A. 2018. It's not cricket: examining political discussion in nonpolitical online space. *Information, Communication & Society* 21(11).

Zhang, J. S.; Keegan, B. C.; Lv, Q.; and Tan, C. 2021. Understanding the Diverging User Trajectories in Highly-related Online Communities during the COVID-19 Pandemic. *International AAAI Conference on Web and Social Media* .