Comparing Apples and Oranges: Recognizing Political Heterogeneity on Reddit and Its Implications for Behavioral Analysis

Anonymous ACL submission

Abstract

Reddit is home to a broad spectrum of political activity and users signal their political affiliations in multiple ways—from self-declarations to community participation. Commonly, political studies have assumed political users are a single bloc, both in developing models to infer political leaning and in studying political behavior. Here, test this model assumption of political users. We show that a variety of commonly-used political-inference approaches models do not generalize, indicating heterogeneous types of political users, and remains imprecise at best for most users, regardless of which sources of data or methods are used. Across a 14-year longitudinal analysis, we demonstrate that the choice in definition of a political user has significant implications for behavioral analysis. Controlling for multiple factors, political users are more toxic on the platform and inter-party interactions are even more toxic—but not all political users behave this way. Last, we identify a subset of political users who repeatedly flip affiliations, showing that these users are the most controversial of all, acting as provocateurs by more frequently bringing up politics, and are more likely to be banned, suspended, or deleted.

1 Introduction

Individuals readily engage in political behavior online, sharing content and forming communities with like-minded individuals. Scholars study these active political communities to understand partisanship (Leong et al., 2020), polarization (Morales et al., 2021; Hofmann et al., 2021), and voting behaviors (Gayo-Avello, 2012).

Many studies of political behavior in social media have the underlying assumption that political leanings can be reliably identified. Prior work has shown that partisan leaning can be inferred from a diverse set of behavioral characteristics such as text (Volkova et al., 2014), social networks (Lindamood et al., 2009; Barberá, 2015), and even community participation (An et al., 2019). Yet, inferring political leaning is known to be a challenging problem (Cohen and Ruths, 2013), in particular for centrist or apolitical users who infrequently express political beliefs. Further, inference models typically used a single source of political affiliation without examining whether this source generalizes to all types of users. This methodology fails to account for the disparate types of political users and introduces sampling bias downstream. Here, we re-examine inferring political behavior for these diverse groups in a unified setting to understand the consequences our data have on results.

This paper tests what effect current assumptions of social media users' political affiliations have on our ability to model political users and their behaviors. The first part of the paper tests how different definitions of political users generalize to other users' behaviors and to inferring political leaning. Using 574K political users on Reddit, we show that the common definitions of a political user (e.g., those making self-declarations of affiliation) result in behaviorally diverse types of users. Further, we demonstrate that multiple computational approaches for political inference do not generalize across these political users types; our results show that political inference on Reddit is challenging, with our best model for inference only attaining a 0.60 AUC score across all users.

The second part of the paper tests whether the choice in which type of political user influences the outcomes of political analyses. We show that controlling for multiple factors, political users are generally more toxic on the platform and that cross-affiliation interactions are *even more* toxic—with liberal-to-conservative interactions being most toxic. However, not all types of political users are equally toxic, highlighting the importance of how studies define political users. In addition, we identify a small set of users who near-simultaneously declare differing political affiliations. These users act provocateurs and have substantially more controversial comments—with the most active eventually becoming banned. Together, our study has substantial implications for future work on political behavior on Reddit and highlights the need to account for different types of political users.

2 Political Affiliation Online

Online communities are active spaces for political discussions and cross-community engagement. Researchers have examined how these political spaces, and the users therein, influence real-life politics (Zhuravskaya et al., 2020), forecast future political outcomes, (Swamy et al., 2017), increase political engagement offline (Lane et al., 2017) and even polarize opinions (Settle, 2018). Such research depends on knowing the political affiliation of users.

People online may express their political affiliations explicitly or implicitly, and not all users reveal their affiliations (Haq et al., 2020). This lack of data potentially limits large studies of political engagement. As a result, substantial work has focused on inferring affiliation to increase the data representativeness (e.g., Rao et al., 2010; Al Zamal et al., 2012; Gentzkow et al., 2016; Preoţiuc-Pietro et al., 2017; Tatman et al., 2017). However, political inference is known to be challenging, and prior work has shown methods often fail to generalize to users outside the narrow range of political orientation on which they were trained (Cohen and Ruths, 2013). Moreover, the majority of work uses only a single source of ground truth—when multiple are available—without testing the implications of which type of user makes the political declaration, and whether those users are representative at large. This study tests this underlying assumption of generalizability of how political users are defined and what effect this has on affiliation inference and behavioral studies of political users.

3 Identifying Political Affiliation

Individuals signal their political beliefs in multiple ways from self-declarations to participation in partisan communities. These different sources of information offer complementary ways of recognizing beliefs—and defining who exactly is a “political user.” Prior works have varied significantly in which of these signals they use (e.g., Beller et al., 2014; Shen and Rose). Following, we define political affiliation and describe different sources of political identification.

Defining Political Affiliation Political affiliation is a complex description based on a person’s values and special interests (Conover and Feldman, 1984). Multiple studies have attempted to simplify affiliation to a single dimension (Poole and Rosenthal, 1985; Clinton et al., 2004; Shor and McCarty, 2011), with the most common being a continuous ideal point value along a conservative-liberal spectrum. Prior work has largely adopted binary affiliation labels (e.g., An et al., 2019; Shen and Rose), though some work has attempted to infer continuous values (Preoţiuc-Pietro et al., 2017). Due to the sparsity of information and the need to support non-American affiliations, we adopt binary conservative and liberal labels.

Metadata Affiliations (Flair) Multiple Reddit communities allow users to have a piece of flair displayed with their username (Tigunova et al., 2020; He, 2021); several political communities follow this practice, allowing us to extract precise affiliations for users based on their self-declared identity. For example, a user posting in the r/Conservative subreddit may select a “Reagan Republican” or “Trump Supporter” flair, both indicating a conservative political leaning. In total, we used 70 known flairs and iterated through the 169 months of comments and posts which resulted in 16,451 unique users with a political flair.

Self-declarations In conversation, individuals will sometimes make self-declarations about their identity (Bergsma and Van Durme, 2013; Beller et al., 2014). Therefore, we capture politically-related self-declarations using a limited set of regular expressions; for example, a user who commented “I only vote Republican” would be labeled as a conservative. Matched comments were further filtered to remove posts from known bots, quotations and hypothetical statements, and statements indicating a past affiliation that does not imply a present one. Appendix §A.1 describes the regular expressions and filtering.

To verify the accuracy of the extracted labels, three annotators labeled a sample of 100 instances, labeling users as liberal, conservative, ambiguous, or neither. Annotators attained a Krippendorf’s α=0.82; this agreement is substantially higher than seen for annotating general user statements (cf. Shen and Rose) because the text focuses on political self-declarations. Among 31 pairs of disagreeing annotations from three annotators, 29 of them have at least one annotator labeling it as *can’t*

Dataset	Conservatives	Liberals	*Total*
Flair	12,185	4,266	16,451
Self-declaration	12,542	17,961	30,503
Community data	343,773	183,102	526,875

Table 1: Dataset sizes based on source of ground truth

Source One		Source Two		
		Flair	Self-Declaration	Community
	Flair	-	0.014	0.025
	Self-Declaration	0.461	-	0.063
	Community	0.443	0.034	-

Table 2: Overlap in the percent of users in Source One users who are in Source Two

tell or *neither* — suggesting most disagreements were due to vagueness in the comment.

Community Participation Reddit has multiple communities associated with political ideologies (Weninger et al., 2013; Soliman et al., 2019). Participating in these communities can thus serve as an implicit signal of affiliation. For example, if a user frequently posts in `r/Conservative`, they can be assigned as a conservative user. Prior work has used participation in these communities as a proxy for affiliation (e.g., An et al., 2019; Shen and Rose). We intentionally exclude (i) quasi-political communities such as r/the_donald, which though affiliated, attracts a broader set of users, and (ii) political communities with mixed affiliations to maximize the precision of the ground truth. Some users participate in multiple communities across the political spectrum; we exclude these from the dataset. Using a list of 24 political communities (see Appendix Table 5), we identify 343,773 conservative and 183,102 liberal users.

Data Summary The dataset is collected from Reddit and consists of all English comments and posts from December 2005 until December 2019. We identified 573,829 political affiliations as seen in Table 1. The community labels are the largest source of affiliation, providing ∼17x more data than self-declarations from the comments. These datasets show two important trends. First, surprisingly, few users had more than one source of affiliation, shown in Table 2; a little under half the users who self-declare (44%) or have user flair (46%) also actively participate in politically-affiliated communities. This difference suggests these sources of ground truth are relatively distinct.

Second, the datasets differ in their skew towards one affiliation, with flair- and community-based affiliations heavily skewed towards conservative users. Given Reddit's reputation for having a liberal bias (Vogels et al., 2021), this skew has an important implication on downstream studies of these users alone. Our results suggest that conservative users are more likely to be more active in overtly partisan communities and identify their politics more clearly than liberal users.

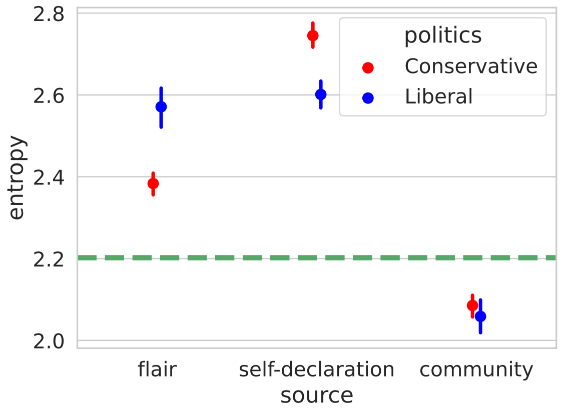


Figure 1: Subreddit entropy by data source and politics, showing community partisans have the least-diverse behavior. The dashed line shows the mean entropy of non-political users.

4 Characterizing Political Behavior

Do users who declare their political beliefs in different ways also *behave* differently? Here, we test for behavioral differences between user categories.

Behavior By Data Source We analyzed general behavioral differences between political users and non-political users using a random sample of 10K users from each category and 10K non-political users. For each category, we measure (i) how old their accounts are as the time between first and last comment and (ii) diversity in community participation as entropy over subreddits.

Substantial variation was seen between the groups. Users with no declared affiliation had accounts nearly twice as old (μ=94 months) as political users (μ=46), and for every political data source, conservatives have a shorter lifespan of activity. The median longevity for conservative users is a full year less than their liberal counterparts. Conservatives in the flair dataset have the shortest overall lifespan with a median of 31 months. As Reddit's user base has grown substantially since its beginning—particularly with an influx of political users around the 2016 U.S. election—our results point to the need to recognize political and non-political users as heterogeneous groups.

Political users varied in how widely they com-

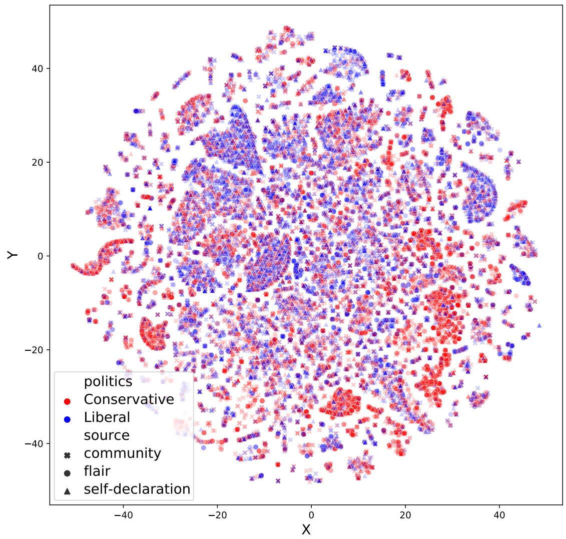


Figure 2: A t-SNE embedding of a sample of political Reddit users according to their commenting behavior reveals partisan clusters mixed by source as well as single-source clusters, indicating heterogeneous types of political users.

ment across communities, with users who self-declare and those with flairs participating in more communities on average. Figure 1 shows the mean entropy for user type, revealing users in two groups participate more broadly than those whose affiliation is derived from participating in partisan communities ($p<0.01$). The entropy is calculated from the probability that user u_j posts or comments in a subreddit $s_i \in S_{u_j}$ across all of their activities: $-\sum_{s_i \in S_{u_j}}(p(s_i) * log(p(s_i)))$. High entropy indicates the user visits many communities with equal frequency; low entropy indicates that they visit a few communities more often. We average these per-user entropies across all users in a data source. Political discussion on Reddit is known to be common outside of political subreddits (Rajadesingan et al., 2021) and our work suggests that this behavior is driven by certain types of political users.

Conservatives vs Liberals Conservatives and liberals are known to operate in different bubbles online (Adamic and Glance, 2005; Bakshy et al., 2015). Here, we test whether the different groups within an affiliation have separate bubbles themselves. Political users are represented by their commenting frequencies across subreddits. PCA is applied to identify latent variations in where users are active (see Appendix C.1. Figure 2 shows the t-SNE projection of these political users colored by affiliation with shapes for each user type; closeness in this plot indicates users are active in the same communities. This projection shows three trends. First, as expected, some conservative and liberal users participate in bubble-like spaces with users of primarily one affiliation. Second, surprisingly, some clusters exhibit strongly-mixed affiliation, indicating that Reddit is not entirely polarized and some users do regularly interact across affiliations. Third, some politically affiliated clusters are primarily made of one user type (see Appendix Figure 7 for this plot with points colored by type). This result suggests that several micro-bubbles exist where users may not interact with others of their affiliation. As a result, computational studies using only one source of data may incorrectly estimate how information spreads between users or the norms of political users in a community.

5 Inferring Political Affiliation

Multiple methods have been proposed for inferring political affiliation. However, these methods have typically used only a single source of information as ground truth (e.g., community membership). Given the behavioral differences between observed users from different sources of information, we test how well a broad set of approaches identifies political affiliation and to what degree does an approach and source of ground truth generalize to inferring the affiliation for other types of users. Additional details on the hyperparameter settings for each model are detailed in Appendix §C.

Username Classifier Usernames can reveal aspects of identity (Wood-Doughty et al., 2018; Wang and Jurgens, 2018), e.g., `Hillary4Prez` reveals a liberal leaning. To predict affiliation from names, we follow Wang and Jurgens (2018) and train a bidirectional character-based LSTM.

Text Classifier Some topics are politically oriented and can potentially reveal a user's leaning, e.g., discussing interests in gun rights. To infer affiliation from such statements, we train a RoBERTa (Liu et al., 2019b) model over comments made from each user, excluding any statements they make that explicitly self-identify their affiliation. The model predicts each comment, and we aggregate the model outputs by taking the mean of predictions of selected comments associated with a user as the final label.

Behavioral Classifier User behavior can be a strong indicator of affiliation as individuals participate in political or politically adjacent communities (e.g., environmentalism). Prior work has shown

	Username				Text-based				Behavioral			
Training	Flair	Self-Decl.	Comm.	All	Flair	Self-Decl.	Comm.	All	Flair	Self-Decl.	Comm.	All
Flair	52.06	39.93	61.11	49.47	**70.76**	60.63	55.08	55.46	58.10	50.51	54.52	54.30
Self-Declaration	50.45	43.55	**66.67**	46.58	61.20	**61.06**	54.87	55.70	48.66	52.54	60.76	59.61
Community	47.23	51.88	40.74	45.79	67.08	60.68	60.34	**60.43**	50.58	47.35	50.83	50.52

Table 3: Classifier performances (AUC) at predicting user political affiliation relative which dataset a model is trained (row) and tested on (column). The best performing system (method + data) on each test set is **bolded**.

that modeling user engagement across subreddits using community2vec (Martin, 2017) can identify subreddit-specific affiliations (Waller and Anderson, 2020). This process is analogous to training a word2vec model with separate user u and subreddit s embeddings which learn parameters to maximize $\sigma(u_i \cdot s_j)$=1 if the user participates in the subreddit or 0 if not. We extend this approach to use semi-supervised training in a multi-task setup: the traditional user2community model is retained and a separate linear layer is used to predict political affiliation from the user embedding if that user's affiliation is known. This semi-supervised setup provides structure to the user embeddings, ideally infusing all users with information on their affiliation based on subreddit commenting behavior. Unlike the text-based classifier, the behavioral one captures user engagement in politically-affiliated communities, even if the user never explicitly declares their affiliation in comments.

5.1 Experimental Setup

All users with a political affiliation were merged into one set and then randomly divided the dataset into train (80%), development (10%), and test (10%) sets. For every classifier, we trained a model on users from each data source and then evaluated users from each source.

For the text classifier, we select at most 20 comments (chosen randomly) for each user. To train the behavior model, we generate a bipartite network between users and subreddits weighted by how often a user posts in a subreddit. The network was restricted to the top 1000 subreddits and users were required to have a minimum of 10 posts, following Waller and Anderson (2020). We randomly sampled non-political users (5x political users) and introduced their subreddit frequencies into the bipartite network (Appendix §C).

5.2 Results

Classifier performance, shown in Table 3, reveals stark contrasts in generalizability between the different data sources—with no model performing highly accurately (F1 scores shown in Appendix Table 7). In general, text-based classifiers perform better than classifiers inferring affiliation from a username or where a user comments. When generalizing to all data, username models performed worse than random, suggesting this approach is unsuitable for generalizing to the broader population. In most cases, models perform best on users whose affiliation was determined in the same way as training users (e.g., train and test on flair users).

Could models still be effective if limited to their high-confidence predictions? We plotted the precision-recall curve of each classifier across all user types, shown in Appendix Figure 13, which reveals models' precisions are moderate at best, with no model offering substantially higher precision at the expense of recall. Model predictions were moderately correlated with each other (mean Pearson's r=0.36), indicating they capture complementary information about a user.

The relatively low performance of the text-based classifier aligns with recent insights from Shen and Rose which found annotators have a hard time perceiving ideology from text alone. Our work offers complementary insight with Cohen and Ruths (2013), who found political inference in Twitter is easy for sharply partisan users but challenging for moderates and apolitical users; our results show that even for openly political users, models typically perform poorly, though mostly above chance.

6 Political Interactions and Engagement

Political discussions are known to be heated (Iyengar et al., 2019) and online discussions of political topics are more uncivil and aggressive than non-political topics (Coe et al., 2014; Barnidge, 2017). In part, political topics have become increasingly moralized (Finkel et al., 2020), where discussions are more connected to a person's identity. Here, we examine the interactions between political users to probe the mechanisms behind this toxicity. Reddit allows communities to discuss political topics with like-minded individuals, but also allows common spaces for both political and non-political topics for all (Rajadesingan et al., 2021). As a result, we test whether these discussions become more uncivil due to political persons or the topic itself. Further, given the clear differences seen between our groups of political users, we test whether these users behave differently to test for potential confounds from only studying one group.

6.1 Experimental Setup

To test for affiliation-based hostility, we construct a mixed-effect linear regression model to estimate the toxicity of a reply to a comment. We include a random effect for the subreddit in which a discussion takes place, which controls for the relative levels of toxicity in different subreddits (Rajadesingan et al., 2020). Categorical variables are used for the political affiliation of the parent comment's user and the replying user, setting the reference to `Unknown` for all users without ground truth. We include fixed effects for which type of source is used to determine the political affiliation as a way of estimating whether these sources reflect different groups of users with distinct behaviors. Comments by flair-based users provide an explicit signal of affiliation that may attract more hostility; therefore we include a fixed effect for whether the parent comment's user's political affiliation is visible in the subreddit. We add a factor for whether the discussion is in one of 187 political subreddits (Appendix §E) to test whether discussions around political topics are more contentious, which cover news, regions, ideologies, politicians, and activism. Finally, as toxic conversations may lead to more toxicity, we include a linear factor for the parent comment's toxicity.

We select comments where at least one of the comment's user and replying user appears in all political users we identified from all comments in our dataset. We also sample some interaction comments from non-political users to non-political users (`Unknown` to `Unknown`). In this way, we collected 6,099,866 interaction comments.

Toxicity is defined as messages which include insults, threats, or containing profane language (Wulczyn et al., 2017). We follow the approach of previous work studying political toxicity on Reddit (Rajadesingan et al., 2020) for our regression settings. To measure toxicity, we fine-tune a BERT (Devlin et al., 2019) model on the Offensive Language Identification Dataset (Zampieri et al., 2019) This dataset collects comments from Twitter, which are shorter on average but are similar in style and register. Our toxicity model follows the setup of the top-performing SemEval system on the same data (Liu et al., 2019a) and attained an F1 of 82.3, which is close to their reported F1 of 82.9. We validate the toxicity scores on the Reddit data by evaluating the model on 150 manually annotated comments, which resulted in a 0.88 weighted-F1, indicating the model generalizes to toxicity on Reddit. The model assigns each comment a toxicity score between 0 and 1.

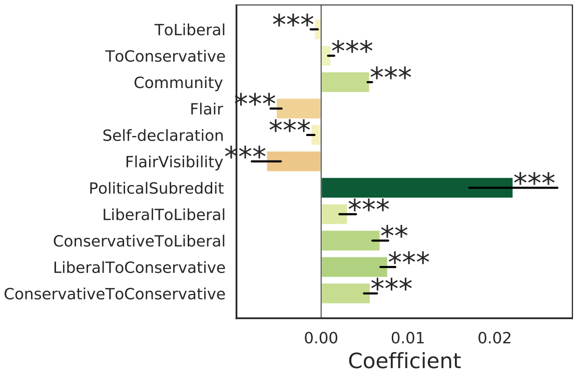


Figure 3: Significant regression coefficients for explaining the toxicity of a reply relative to the political affiliation of the users. Full coefficients are in Table 8.

6.2 Results

Regressing on the factors contributing to toxicity in replies shows three main findings (Figure 3; full regression in Appendix Table 8). First, consistent with prior work, we find that controlling for subreddit-specific levels of toxicity, discussion in political communities is much more toxic, suggesting that these topics are a primary source of increased hostility.

Second, we find substantial affiliation-based toxicity, with increased toxicity particularly for interactions between cross-affiliation users. While conservative users receive more toxic replies, such users are more toxic when replying to liberal users than liberal-to-conservative replies. Surprisingly, this increased toxicity is not due to an explicit flair signal; when wheres are commenting in a community where the flair is visible—which can include mixed-affiliation subreddits—users receive less toxic replies.

Third, our results point to clear behavioral differences between the three different sets of users. Across all of Reddit, individuals who actively participate in politically affiliated subreddits for one

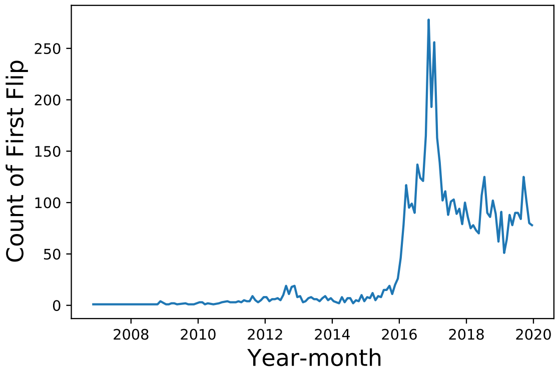


Figure 4: Temporal frequencies of two-faced actors declaring different political affiliations within a 90-day window. Large number of two-faced actors showed up in Reddit around the 2016 U.S. presidential election.

party are substantially more toxic in their interactions; in contrast, those who participate in flair-based communities or who have declared their affiliation in a comment (but do not participate in political communities or have flair) are much less toxic. Our result points to the importance of explicitly recognizing and modeling differences of how users self-affiliate, as these choices have significant downstream implications for behavioral studies.

7 Two-Faced Actors

Our data identified a small percent of political users who declare different political affiliations within a short period. Given the rise in trolls and other malicious actors on social media (Zannettou et al., 2019; Im et al., 2020), we ask whether these users, who we refer to as *two-faced actors*, behave differently than other types of political users.

Experimental Step To identify two-faced actors, we analyze explicit political declarations made in the flair and self-declarations user sets. Users are filtered to identify those that declare different affiliations within a 90 day period.A total of 5,524 users match these criteria in our data, which we refer to as *two-faced actors*. The total number of two-faced actors under different time constraints can be seen in Appendix Figure 11.

Analysis Two-faced actors are substantially more active than regular political users and comment 266 times per month, compared to a baseline of 82. These users are late arrivals to Reddit's political sphere and only begin showing up after the November 2016 US presidential election (Figure 4). Their comments are frequently judged more controversial (Figure 5), as measured by Reddit's controversial score which measures the split between upvotes and downvotes. Two-faced actors' average controversial score is 3.4 times higher than the average political user and nearly 10 times higher than non-political users.

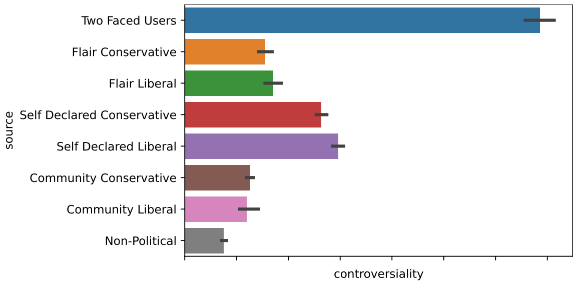


Figure 5: Reddit controversial score by cohort. Two faced actors are far more controversial

To better understand where two-faced actors are active, we plot their commenting behavior by subreddit relative normal political, calculating the log-odds with a Dirichlet prior for the two groups (Monroe et al., 2008), shown in Figure 6. Two-faced actors frequently participate in more contentious subreddits such as r/the_donald, r/ChapoTrapHouse, r/Gamingcirclejerk, and r/genderskeptical; these subreddits cross the political divide and some have been later quarantined by Reddit for being sources of trollish or abusive behavior (Copland, 2020). This behavior suggests that two-faced users are likely not acting in good faith and are behaving as provocateurs on Reddit. This argument is supposed by the fact that 28.92% of their accounts have been either suspended or deleted. In contrast, for other political users studied in §8 had only 17.52% of their accounts suspended or deleted. This study shows that researchers should be aware of the two-faced users when analyzing the political behaviors of users online, as these users form a distinct group that may bias downstream analyses.

8 Changing Political Beliefs

In American politics, political beliefs have shifted closer to closer-held ideological beliefs (Finkel et al., 2020). However, some individuals *do* change affiliation. Are users of one type more likely to switch parties? Here, we test whether affiliation changes can be predicted from prior behavior.

Analysis To identify changing behavior, we use a time constraint that requires the change in political affiliation to occur at least one year following the user's original political declaration. A total of 2,076 affiliation-changing users were identified

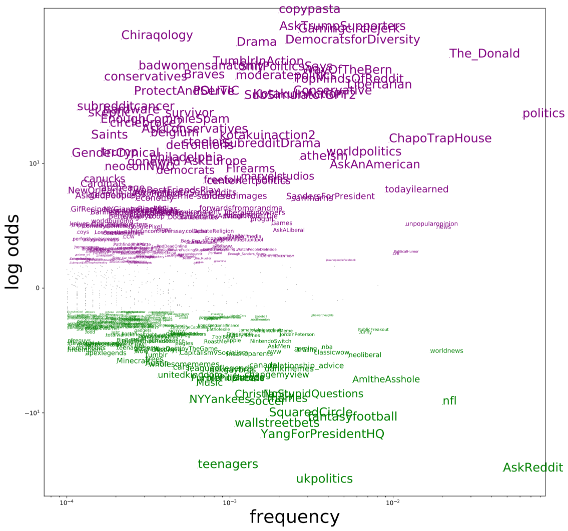


Figure 6: Log odds of subreddits. The upper part is for two-faced actors and the other side is for normal partisan users. Contentious subreddits like r/The_Donald, r/ChapoTrapHouse appear on the two-faced actor side.

through this process. The changing of affiliation was split roughly evenly, with 56% of flips going from conservative to liberal, and 44% going from liberal to conservative. Flips are not unique to one type of user, but users do differ in their probability of flipping, with self-declared and flair users being an order of magnitude more likely to flip (2.8% and 1.6%, relatively) than community-based-users (0.2%). For users who change affiliations, only 17.5% of their accounts have been suspended or deleted, which is significantly lower than the two-faced actor rate and lower than the mean rate for non-political user (21.0%), suggesting these affiliation changes are likely done in good faith.

Experimental Setup To control for confounds from behavioral differences, we created our dataset using coarsened matching to pair users who change their affiliations with political users who do not. Two users are paired by having the same initial politics, closest comment count, and activity lifespan.

For each matched pair of users, we collect six months of features prior to the change of political affiliation. The feature set includes the data source of the users, their original political declarations, and participation in popular and political subreddits. A complete list of the model features can be found in Appendix §C.3. We train a Logistic Regression model to predict whether a user will change their political affiliation. We evaluate separate models for conservatives who became liberals, liberals who become conservatives, and a combined affiliation-independent model to analyze components of change regardless of party.

Results Our results indicate that models can predict changing political affiliation, with the affiliation-independent model attaining an F1 of 64.8, relative to the random baseline of 0.5. Surprisingly, party-specific models had lower performance; The model predicting conservative's change of affiliations resulted in a Macro F1 score of 45.35—worse than random; similarly, the model predicting liberal's changes of affiliations had an F1 score of 48.05. The higher performance of the affiliation-independent model suggests the existence of common signals for intent to change one's political beliefs, independent of party.

Are some types of users less predictable? Separating test results by user type shows the model has substantially higher performance at predicting flips for self-declaration users (77.6 Macro F1) in comparison to flair users (68.9 Macro F1) and community users (57.3 Macro F1). Together with the differences in relative rates of users changing affiliations, our results again point to fundamental differences in behavior for each group of users and, again, the importance of modeling this diversity.

9 Conclusion

Social media is rife with political activity and research on these political spaces depends on accurate measurement of political users. We examine political users on Reddit and show that the choice in how political users are defined—the evidence used to establish ground truth—has substantial consequences for downstream models and analyses. In particular, user groups from different definitions behave differently (§4 and §6) and models trained on one type of user do not necessarily generalize to other groups (§3). In three studies of political users, we show that (i) political users themselves drive hostility on the platform—with conservative users being the recipients of more toxicity, (ii) a small-but-very-active group of provocateurs simultaneously declare different affiliations and are a notable source of toxicity and controversiality on the platform, and (iii) changing political affiliation can be predicted, but performance varies considerably by user type. Across all three studies, we show that the *type* of political user matters, with different types having substantially different behavior. Models, data, and code for this study will be released as *anonymized*.

10 Ethical Considerations

The models in §5 make inferences about the political affiliations of users. Given the increasing importance of political identity in American society (Finkel et al., 2020) and inter-party hostility (Miller and Conover, 2015), these models could come with some risk if a user is mislabeled with an affiliation they do not have, e.g., a public mislabeled political identity could cause a user to be socially ostracized for their supposed political beliefs. However, as we demonstrate, these inference models offer moderate performance *at best* and are not likely to be reliable in practice. As a result, we hope our models discourage future use of such inference on Reddit, mitigating the potential risk.

Annotators were a part of the study team and were not additionally compensated for their annotations.

References

Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43.

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6.

Jisun An, Haewoon Kwak, Oliver Posegga, and Andreas Jungherr. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 68–79.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91.

Matthew Barnidge. 2017. Exposure to political disagreement in social media versus face-to-face and anonymous online settings. *Political communication*, 34(2):302–321.

Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. 2014. I'm a belieber: Social roles via self-identification and conceptual attributes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 181–186, Baltimore, Maryland. Association for Computational Linguistics.

Shane Bergsma and Benjamin Van Durme. 2013. Using conceptual class attributes to characterize social media users. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720, Sofia, Bulgaria. Association for Computational Linguistics.

Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review*, pages 355–370.

Kevin Coe, Kate Kenski, and Stephen A Rains. 2014. Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4):658–679.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on twitter: It's not easy! In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.

Pamela Johnston Conover and Stanley Feldman. 1984. Group identification, values, and the nature of political beliefs. *American Politics Quarterly*, 12(2):151–175.

Simon Copland. 2020. Reddit quarantined: Can changing platform affordances reduce hateful material online? *Internet Policy Review*, 9(4):1–26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eli J Finkel, Christopher A Bail, Mina Cikara, Peter H Ditto, Shanto Iyengar, Samara Klar, Lilliana Mason, Mary C McGrath, Brendan Nyhan, David G Rand, et al. 2020. Political sectarianism in america. *Science*, 370(6516):533–536.

Daniel Gayo-Avello. 2012. "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"–A Balanced Survey on Election Prediction using Twitter Data. *ArXiv preprint*, abs/1204.6441.

Matthew Gentzkow, Jesse M Shapiro, Matt Taddy, et al. 2016. *Measuring polarization in high-dimensional data: Method and application to congressional speech*. National Bureau of Economic Research.

Ehsan Ul Haq, Tristan Braud, Young D Kwon, and Pan Hui. 2020. A survey on computational politics. *IEEE Access*, 8:197379–197406.

Qinglai He. 2021. *Online Platform Policy and User Engagement*. Ph.D. thesis, Arizona State University.

Valentin Hofmann, Janet B Pierrehumbert, and Hinrich Schütze. 2021. Modeling ideological agenda setting and framing in polarized online groups with graph neural networks and structured sparsity. *ArXiv preprint*, abs/2104.08829.

Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. 2020. Still out there: Modeling and identifying russian troll accounts on twitter. In *12th ACM Conference on Web Science*, pages 1–10.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22:129–146.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Daniel S Lane, Dam Hee Kim, Slgi S Lee, Brian E Weeks, and Nojin Kwak. 2017. From online disagreement to offline action: How diverse motivations for using social media can increase political information sharing and catalyze offline political participation. *Social Media+ Society*, 3(3):2056305117716274.

Yuan Chang Leong, Janice Chen, Robb Willer, and Jamil Zaki. 2020. Conservative and liberal attitudes drive polarized neural responses to political content. *Proceedings of the National Academy of Sciences*, 117(44):27731–27739.

Jack Lindamood, Raymond Heatherly, Murat Kantarcioglu, and Bhavani M. Thuraisingham. 2009. Inferring private information using social network data. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 1145–1146. ACM.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Trevor Martin. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 27–31, Vancouver, Canada. Association for Computational Linguistics.

Patrick R Miller and Pamela Johnston Conover. 2015. Red and blue states of mind: Partisan hostility and voting in the united states. *Political Research Quarterly*, 68(2):225–239.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Gianmarco De Francisci Morales, Corrado Monti, and Michele Starnini. 2021. No echo in the chambers of political interactions on reddit. *Scientific Reports*, 11(1):1–12.

Keith T Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 729–740, Vancouver, Canada. Association for Computational Linguistics.

Ashwin Rajadesingan, Ceren Budak, and Paul Resnick. 2021. Political discussion is abundant in nonpolitical subreddits (and less toxic). *International Conference on Web and Social Media (ICWSM).*

Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 557–568.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44.

Jaime E Settle. 2018. *Frenemies: How social media polarizes America*. Cambridge University Press.

Qinlan Shen and Carolyn Rose. what sounds "right" to me? experiential factors in the perception of political ideology. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1762–1771, Online. Association for Computational Linguistics.

Boris Shor and Nolan McCarty. 2011. The ideological mapping of american legislatures. *American Political Science Review*, pages 530–551.

Ahmed Soliman, Jan Hafer, and Florian Lemmerich. 2019. A characterization of political communities on reddit. In *Proceedings of the 30th ACM conference on hypertext and Social Media*, pages 259–263.

Sandesh Swamy, Alan Ritter, and Marie-Catherine de Marneffe. 2017. "i have a feeling trump will win..................": Forecasting winners and losers from user predictions on Twitter. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1583–1592, Copenhagen, Denmark. Association for Computational Linguistics.

Rachael Tatman, Leo Stewart, Amandalynne Paullada, and Emma Spiro. 2017. Non-lexical features encode political affiliation on Twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 63–67, Vancouver, Canada. Association for Computational Linguistics.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2020. RedDust: a large reusable dataset of Reddit user traits. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6118–6126, Marseille, France. European Language Resources Association.

Emily A. Vogels, Brooke Auxier, and Monica Anderson. 2021. Partisan differences in social media use show up for some platforms, but not facebook. "https://www.pewresearch.org/fact-tank/2021/04/07/partisan-differences-in-social-media-use-show-up-for-some-platforms-but-not-facebook/".

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 186–196, Baltimore, Maryland. Association for Computational Linguistics.

Isaac Waller and Ashton Anderson. 2020. Community embeddings reveal large-scale cultural organization of online platforms. *ArXiv preprint*, abs/2010.00590.

Zijian Wang and David Jurgens. 2018. It's going to be okay: Measuring access to support in online communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 33–45, Brussels, Belgium. Association for Computational Linguistics.

Tim Weninger, Xihao Avi Zhu, and Jiawei Han. 2013. An exploration of discussion threads in social news sites: a case study of the reddit community. In *Advances in Social Networks Analysis and Mining 2013, ASONAM '13, Niagara, ON, Canada - August 25 - 29, 2013*, pages 579–583. ACM.

Zach Wood-Doughty, Nicholas Andrews, Rebecca Marvin, and Mark Dredze. 2018. Predicting Twitter user demographics from names alone. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 105–111, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1391–1399. ACM.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th acm conference on web science*, pages 353–362.

Ekaterina Zhuravskaya, Maria Petrova, and Ruben Enikolopov. 2020. Political effects of the internet and social media. *Annual Review of Economics*, 12:415–438.

A Indentifying Political Affiliation

A.1 Self-declaration

Table 4 shows the regular expressions we use to find out political users and corresponding examples. To find self-declaration users, we pick users in any comment or post whose text has a substring that will match with these regular expressions.

A.2 Community users

Table 5 shows the 24 community subreddits we used to find community political users. Users who post in communities in this list but which have different political labels are excluded.

B Characterizing Political Behavior

Figure 7 is the t-SNE plot of the same set of users as Figure 2 where we use different colors to separate data sources and shapes to separate politics. Comparing Figure 2 and Figure 7 plot shows how some users have highly similar behavior but declare their political affiliations in different ways, while other users, particularly those that self-declare, form distinct source-homogeneous clusters indicating they do not participate in communities where they encounter others of their own party but who declare in different ways.

C Additional Training Details

C.1 t-SNE training

For the details of t-SNE plots shown in Figure 2 and Figure 7, we calculate a matrix of 101,959 users by commenting frequencies in across the 359,432 subreddits they comment in. We then decompose the matrix into 20 dimensions by standard PCA. We only include users that comment at least 5 times.

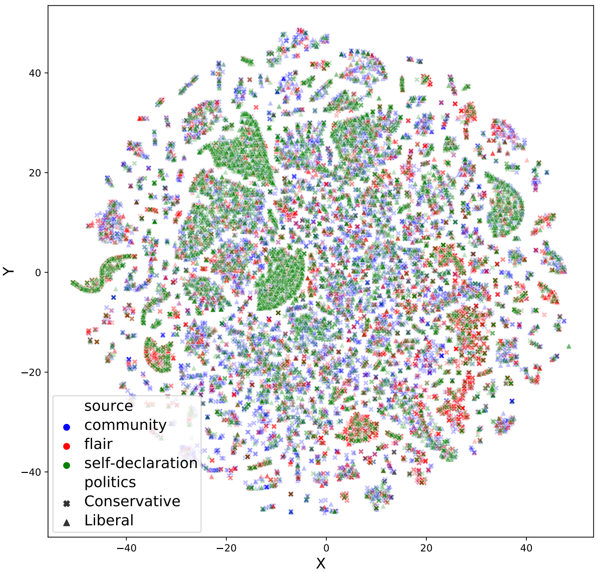


Figure 7: The t-SNE plot of a sample of political Reddit users accord to their commenting behavior (which subreddits they are active in), using colors to separate data source and shapes to separate politics (cf. with Main Paper Figure 2, which shows the same layout but colored by politics).

Then we trained it on t-SNE to 2-dimensions. The perplexity is 60 and the verbose is 4.

C.2 Political Affiliation Classifiers

For the username classifier, the embedding dimension of the bi-LSTM is 15. The hidden dimension is 256. The number of layer is 2. The dropout rate is 0.2. For the text-based classifier, we use the original setting of Roberta model with pre-trained parameters. For the behavioral classifier, the embedding dimension is 50. The dropout is 0.5 applied before the political layer. Table 6 shows other hyper parameters of our classifiers.

Regular Expression	Example
(i am | i’m) a (democrat | liberal)	i am a liberal and i don’t think that the government is more trustworthy
i vote[d]?(for | for a)? (democrat | hillary | biden | obama | blue)	i voted for hillary on the hopes that trump’s rhetoric hadn’t fooled that many of my fellow americans.
(i am | i’m) a (conservative | republican)	i am a republican, and i do think climate change is a real thing
i vote[d]?(for | for a)? (republican | conservative | trump | romney | mcconell)	i voted for trump for his stance on immigration and economy.

Table 4: Regular expressions we used to find self-declaration users and their corresponding examples.

Subreddit	Political Label
r/alltheleft	Liberal
r/Capitalism	Conservative
r/Conservative	Conservative
r/conservatives	Conservative
r/demsocialist	Liberal
r/democrats	Liberal
r/GreenParty	Liberal
r/Liberal	Liberal
r/Libertarian	Conservative
r/LibertarianLeft	Liberal
r/LibertarianSocialism	Liberal
r/Marxism	Liberal
r/neoprogs	Liberal
r/new_right	Conservative
r/progressive	Liberal
r/Republican	Conservative
r/republicanism	Conservative
r/republicans	Conservative
r/socialdemocracy	Liberal
r/socialism	Liberal
r/tea_party	Conservative
r/occupywallstreet	Liberal
r/hillaryclinton	Liberal

Table 5: 24 community subreddits used to find community political users

C.3 Changing Political Beliefs

For the feature of logistic regression, we select the number of comments a user have in the top 100 most frequent subreddits and 24 political subreddits shown in Table 5. Besides that, we include a user's Reddit-specific feature including controversiality, awards, score, gilded. For the related counts of users, we include morning, afternoon, evening post count and total comments count. We also include the account age (in months) and source of a user (how they were defined as political). We randomly split the users into training (80%) and test (20%) sets with fixed random seed (42) across all experiments; no hyperparameter tuning was performed.

D Inferring Political Affiliation

Are different political inference models capturing the same information or complementary information? To test this, we examine the correlations in predictions between classifiers. Figure 8 is a bi-variate plot of predictions on self-declaration users from two text-based models. One is trained by self-declaration and the other is trained by flair users. The diagonal shaped figure showed that our text-based models are able to transfer across different sources in some level. Figure 9 is a similar plot while the training sources are self-declaration and flair users and the evaluation targets are self-declaration users. Figure 10 is a similar plot while the training sources are self-declaration and flair users and the evaluation targets are all users. Future work may try to leverage these complementary sources to improve overall prediction accuracy.

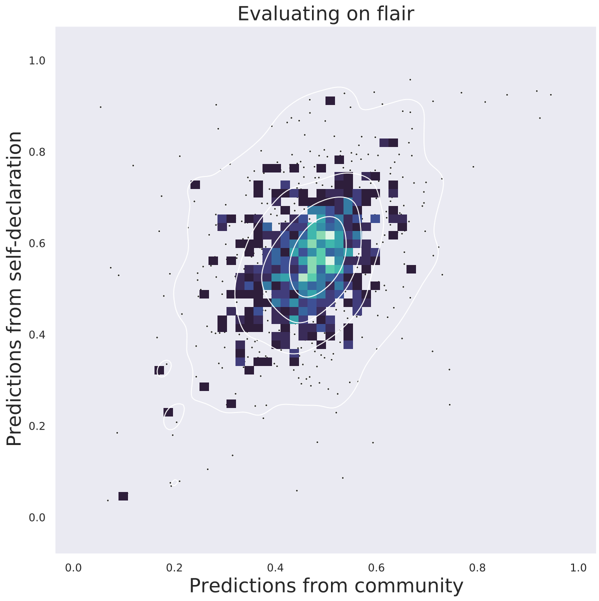


Figure 8: A bi-variate plot of predictions on flair users from two text-based models. One is trained by self-declaration users and the other is trained by community users.

Table 7 is the Macro-F1 performance of each classifiers. Figure 13 is the precision-recall curve of each classifier across all user types.

E Political Interactions and Engagement

The full list of 187 political subreddits was obtained from the curated list at `https://www.reddit.com/r/redditlists/comments/josdr/list_of_political_subreddits/`.

Table 8 is an overall summary of the regression coefficients of variables at predicting toxicity in a reply to a user.

The final regression predictors include:

- FromPolitics: The political affiliation of the replying user of a comment. It can be 'Liberal'

	hyper-parameters				
Classifier	epoch	optimizer	learning rate	loss function	batch size
Username	10	Adam (Kingma and Ba, 2015)	1e-3	BCE	128
Text-based	10	Adam	1e-5	Cross entropy	64
Behavioral	10	Adam	1e-4	BCE	512

Table 6: Username classifier performance (Macro-F1) at predicting user political affiliation relative which dataset the model is trained and tested on.

	Username				**Text-based**				**Behavioral**			
Training	Flair	Self-Decl.	Comm.	All	Flair	Self-Decl.	Comm.	All	Flair	Self-Decl.	Comm.	All
Flair	42.73	29.21	25.14	36.91	**59.70**	**57.83**	50.02	51.07	56.26	51.84	51.85	52.01
Self-Declaration	31.91	37.41	41.17	34.71	57.19	54.65	53.41	53.89	48.94	51.84	**56.80**	**56.08**
Community	20.25	37.05	39.96	29.38	51.02	55.62	48.62	49.43	48.30	45.39	50.87	50.30
Majority Class	42.60	38.08	39.69	39.14	42.60	38.08	39.69	39.14	42.60	38.08	39.69	39.14
Random	47.01	49.18	48.69	48.97	47.01	49.18	48.69	48.97	47.01	49.18	48.69	48.97

Table 7: Classifier performances (Macro-F1) at predicting user political affiliation relative which dataset a model is trained (row) and tested on (column). The best performing system (method + data) on each test set is **bolded**. Note that the random and majority baselines are the same across all classifiers.

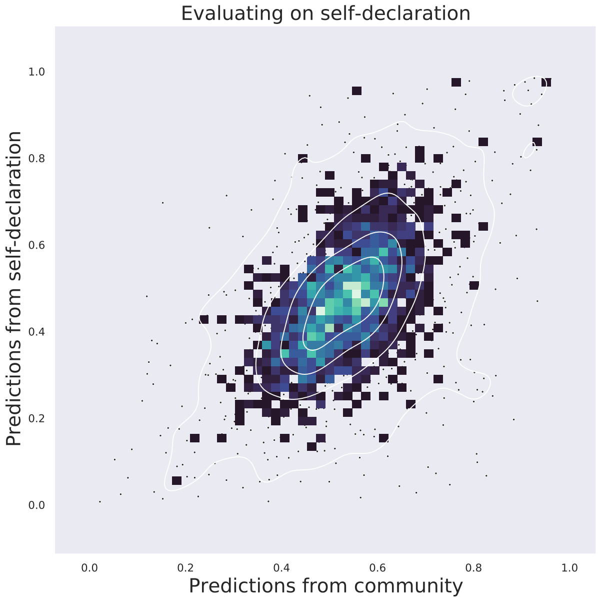


Figure 9: A bi-variate plot of predictions on self-declaration users from two text-based models. One is trained by self-declaration and the other is trained by flair users.

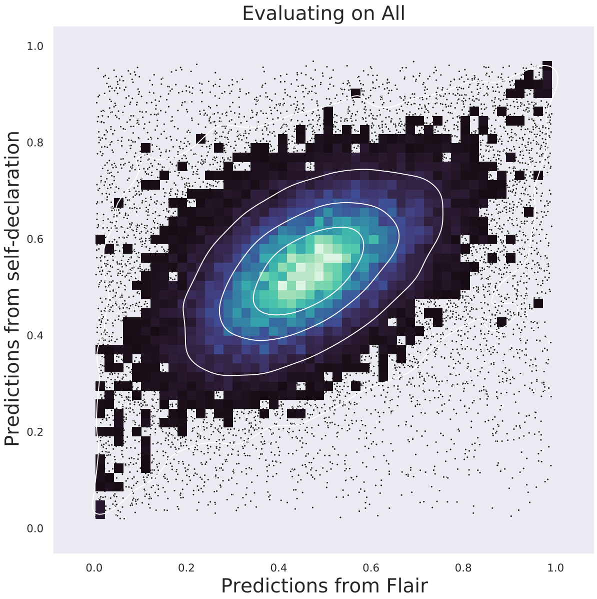


Figure 10: A bi-variate plot of predictions on all users from two text-based models. One is trained by self-declaration and the other is trained by flair users.

or ’Conservative’.

- ToPolitics: The political affiliation of the parent user of a comment. It can be ‘Liberal’ or ’Conservative’.
- source: The source dataset of the replying user. It can be ‘community’, ’flair’, or ‘Self-declaration’.
- Flair Visibility: Boolean variable indicating if the flair is visible to the replying user.
- Political Subreddit: Boolean variable indicating if the comment is in a political subreddit.
- Parent Toxicity: A floating number indicating the toxicity of the parent comment.
- FromPolitics:ToPolitics: Composition affiliations of the replying and parent user.

F Two-Faced Actors

Figure 11 shows the number of Two-Faced actors with varying time constraints between when two declarations of different political affiliations would be considered suspect. In the main paper, we opt for the conservative estimate of 90 days under the assumption that most individuals would not publicly declare opposing political beliefs within a three month period.

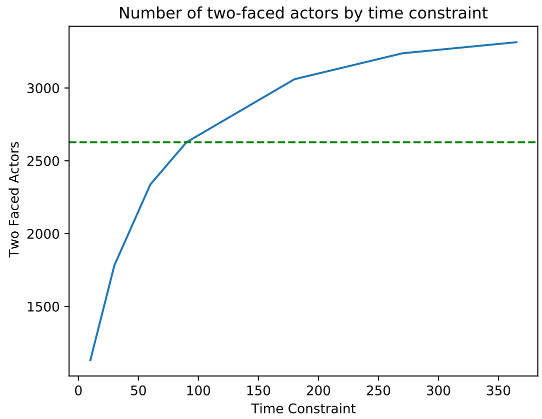


Figure 11: Two-Faced actors count based on time constraint i.e. min days between flips. Dashed line represents the number of two-faced actors at 90 days.

Figure 12 shows the top 30 subreddits where Two-Faced actors post and comment. The x-axis is the log of the real counts.

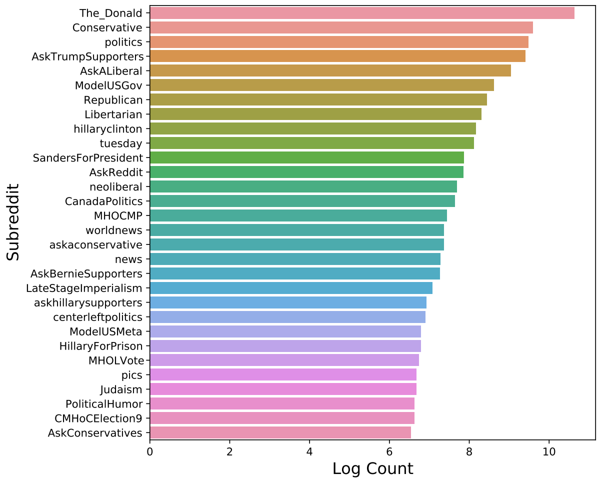
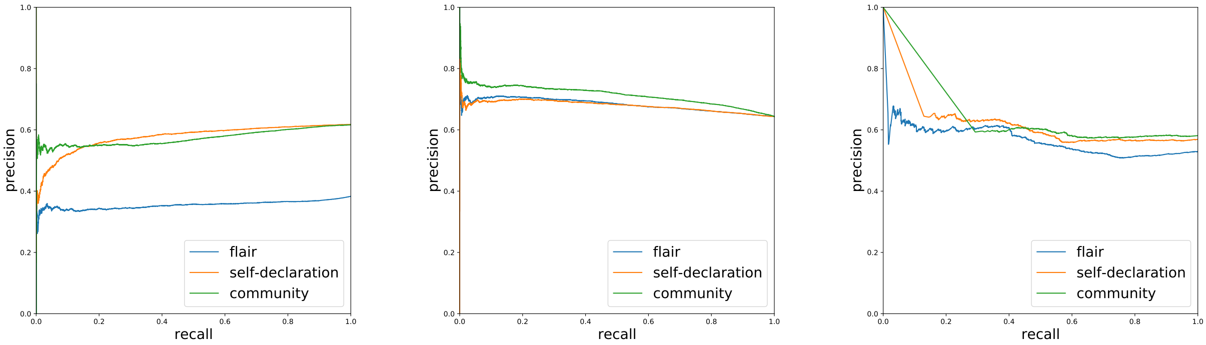


Figure 12: The frequency of the subreddits in which two-faced actors post and comment (log-scaled), highlighting their activity in controversial subreddits.

	Dependent variable:
	toxicity
C(FromPolitics)Liberal	0.000183
	(0.000386)
C(FromPolitics)Conservative	0.000199
	(0.000355)
C(ToPolitics)Liberal	-0.000791^{*}
	(0.000419)
C(ToPolitics)Conservative	0.001154^{***}
	(0.000360)
C(source)community	0.005610^{***}
	(0.000264)
C(source)flair	-0.005189^{***}
	(0.000643)
C(source)Self-declaration	-0.001189^{***}
	(0.000424)
Flair Visibility	-0.006314^{***}
	(0.001676)
Political Subreddit	0.022141^{***}
	(0.005082)
Parent Toxicity	0.164355^{***}
	(0.000398)
C(FromPolitics)Liberal:C(ToPolitics)Liberal	0.003064^{***}
	(0.000959)
C(FromPolitics)Conservative:C(ToPolitics)Liberal	0.006817^{***}
	(0.000910)
C(FromPolitics)Liberal:C(ToPolitics)Conservative	0.007690^{***}
	(0.000860)
C(FromPolitics)Conservative:C(ToPolitics)Conservative	0.005688^{***}
	(0.000752)
Constant	0.257223^{***}
	(0.000861)
Observations	6,099,866
Log Likelihood	−64,981.070000
Akaike Inf. Crit.	129,996.100000
Bayesian Inf. Crit.	130,227.700000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 8: Regression coefficients for predicting the toxicity of a reply relative to political, social, and toxicity factors. A plot of the coefficients for the significant terms is shown in Figure 3 in the main paper.



(a) Username Classifier

(b) Text-based Classifier

(c) Behavioral Classifier

Figure 13: The precision-recall curve of each classifier.