Reading Between the Tweets: Deciphering Ideological Stances of Interconnected Mixed-Ideology Communities

Anonymous ARR submission

Abstract

Online communities provide a platform for sharing ideas and information. Recent advances in NLP have driven interest in understanding the ideological nuances between these communities. Existing research has focused on probing the views of liberals and conservatives, treating them as separate groups. However, this fails to account for the nuanced views of the organically formed online communities and the connections between them. In this paper, we use discussions of the 2020 U.S. election on Twitter to identify complex interacting communities. Capitalizing on this interconnectedness, we introduce a novel approach that harnesses message passing in finetuning language models to probe the nuanced ideologies of these communities. Extensive experiments demonstrate that our proposed method consistently outperform existing baselines, highlighting the potential of using language models in revealing complex ideologies within and across online communities.

1 Introduction

Social media platforms like Twitter and Facebook connected people worldwide within digital town squares, transforming how they share information and exchange ideas. However, mass connectivity, has created new vulnerabilities, including rampant misinformation, the formation of echo chambers that confirm people's pre-existing beliefs (Cinelli et al., 2021; Rao et al., 2022), and the fragmentation of society into polarized factions that disagree with and distrust each other (Iyengar et al., 2019). These developments intensify societal conflicts and undermine trust in democratic institutions (Kingzette et al., 2021; Whitt et al., 2021).

Given these challenges, understanding the ideological nuances within online communities is essential. Existing works provide insights into political ideologies of online groups (Webson et al., 2020; Jiang et al., 2022); however, they treat ideology as a liberal/conservative binary and cannot capture the spectrum of ideologies that may organically emerge in interconnected online communities.

To bridge this gap, we describe a methodology to uncover interacting communities in political discourse on Twitter that are not merely liberal or conservative, but possess a complex mixture of political ideologies. To reveal communities' ideologies, we adapt GPT-2 language models to the language of communities by finetuning on tweets they generate. This finetuning, enriched by message passing techniques inspired by Graph Convolutional Networks (Kipf and Welling, 2016), leverages the interconnected nature of these communities, allowing for a more robust representation of their ideological stances. With generative language models, we can then probe the stances of the communities towards various targets, including different political figures and social groups, by looking at the sentiment of generated responses. This way we can measure 1) for each target, which communities are more in favor of or against it (target-specific community ranking), and 2) for each community, which targets it favors more and which it is against (community-specific target ranking). Our method, when benchmarked against existing baselines, consistently outperforms them on these tasks, validating its effectiveness in capturing the political ideology of interconnected online communities.

Our work highlights the potential of leveraging social media data to reveal the nuanced ideological stances of organically-formed, interconnected online communities. Such insights pave the way for a more informed understanding of the dynamics and shifts in digital attitudes.

2 Related Work

Sociolinguistics and Online Communities. Existing research examined language change and social dynamics of online communities from a number of perspectives. Danescu-Niculescu-Mizil et al.

(2013) analyzed linguistic change in two online communities of beer enthusiasts. They identified strong patterns within the lifecycle of users within online communities determined by their receptivity to community language norms. Eisenstein et al. (2014) identified geographic differences in the use of language on Twitter and tracked diffusion of linguistic changes across United States. They showed that demographically similar communities were more likely to adopt new language norms.

Framing and Ideology. Political speech uses framing to make certain aspects of the message salient (Lakoff, 2014). By highlighting these aspects, the message can implicitly manipulate the understanding, without explicitly biased argument. Polarized language allows partisans to talk about the same issues using different words to elicit different mental and emotional frames: e.g., "tax relief" creates an impression that taxes are an affliction, while talking about "illegal aliens" instead of "undocumented workers" makes the same group appear threatening (Webson et al., 2020). Word embeddings provide a technology to automatically detect frames in polarized text (Kozlowski et al., 2019). Milbauer et al. (2021) trained word embeddings on 32 communities from Reddit and discovered multifaceted ideological and worldview characteristics of community pairs, beyond the predetermined "left" vs. "right" dichotomy of U.S. politics. By using machine translation, KhudaBukhsh et al. (2021) studied the political polarization and demonstrated that liberal and conservatives use different expressions as two languages. He et al. (2021) explore the stances of bipartisan news media towards various topics using contextualized word embeddings. Relevant work also showed different patterns of moral framing among liberals and conservative in the partisan news headlines (Mokhberian et al., 2020) and rhetoric of political elites such as speeches given on the floor of the House and Senate (Wang and Inbar, 2021).

Probing Community Ideologies with LMs. There is growing interest in adapting language models (LMs) to probe the ideologies of human communities. Chu et al. (2023) predicted public opinions from language models by finetuning the models to online news, TV broadcast, and raido shows. By conditioning GPT-3 on socio-demographic backstories from real human participants, Argyle et al. (2022) demonstrated that the information contained in GPT-3 goes beyond surface similarity and reflects the nuanced and multifaceted nature of human attitudes. Feng et al. (2023) studied politically biased LMs by left and right news and Reddit corpora on hate speech and misinformation detection, and revealed that pretrained LMs reinforce the polarization present in the pretraining corpora. Jiang et al. (2022) finetuned two language models on tweets from Democratic and Republican communities and probed the ideological stances of the two communities from the models using language prompts that elicit opinions. However, they focus on two manually-defined Democrat/Republican communities and ignore the interactions between them.

3 Data

2020 U.S. Election Twitter Data. We use a public Twitter dataset about the 2020 U.S. presidential election (Chen et al., 2021). The data was collected by tracking specific user mentions and accounts tied to the official or personal accounts of candidates, ranging from December 2019 to June 2021. We limit tweets to the time period before April 10 2020, the time of the ANES survey, which we use as ground truth. This way, the dataset does not leak information beyond this date.

We identify online communities based on the news co-sharing activities (§4). We only keep users who authored at least one tweet containing a URL to a news article and extract the domain of the URL. The domain represents a news outlet. We identify a total of 996 news outlets in this dataset, with the top 10 most shared outlets being *nytimes*, *foxnews*, *washingtonpost*, *cnn*, *breitbart*, *thehill*, *politico*, *nypost*, *cnbc*, *businessinsider*. After processing, we are left with 41M tweets from 135K users.

ANES Survey. Following Jiang et al. (2022), we use the 2020 Exploratory Testing Survey[1] from the American National Election Studies (ANES), which provide ground truth data for evaluating ideological stances predicted by language models. This survey was conducted in April 2020 with a sample of 3,080 US adults. We use the 30 questions from the *Feeling Thermometers* section, which asked participants to rate a target—a person or a group—on a scale from 0 to 100. A higher score indicates a warmer, more positive attitude towards the target, and a lower score indicates a cooler, more negative

[1] https://electionstudies.org/data-center/2020-exploratory-testing-survey

attitude. For each target, the bipartisan ground-truth ratings are the average across all scores from liberals and conservatives respectively. Please refer to Appendix A for the 30 studied targets.

4 Exploring Ad-hoc Online Communities

Communities in News Co-sharing Network. We represent the structure of the information ecosystem as a *news co-sharing network* (Faris et al., 2017; Mosleh and Rand, 2022; Starbird, 2017) and discover communities in it. Utilizing community detection on a *news co-sharing network* is instrumental in discerning the underlying patterns of information dissemination and consumption. By analyzing these communities, we can comprehend how users cluster based on their news-sharing behaviors, offering insights into the sources they prioritize and trust. Such an approach aids in capturing the nuanced dynamics of news engagement, revealing potentially shared interests, regional relevance, or the impact of influential figures.

We construct a bipartite *news co-sharing network* $G_{co} = (U, V, E)$, where U is the set of users, V the set of news outlets (specified by their domains), and E the weighted edges between them. An edge's weight represents the number of times a user u ($u \in U$) shared links to news stories from this outlet v ($v \in V$) in their tweets. We use Louvain algorithm (Blondel et al., 2008) to identify communities on G_{co}[2]. As a result, each community $C = (U^C, V^C)$ consists of a set of users U^C and news outlets V^C. The method identifies 42 communities. We keep the 20 largest communities, and the users from these communities cover more than 99% of tweets in the dataset. The statistics and the most shared news outlets in these top 20 communities are shown in Table 1.

Mixed Ideologies of Online Communities. To investigate the ideological leaning of online communities, we first need to identify the partisanship of its constituents. Previous works have leveraged on cues in tweet text (Rao et al., 2021; Cinelli et al., 2021), follower relationships (Barberá, 2015) and retweet interactions (Conover et al., 2011; Badawy et al., 2018) to quantify user ideology. In this study, we rely on methods discussed in (Rao et al., 2021) to identify user ideology. Specifically, this method extracts ideological cues from tweet text and URLs embedded in them to classify ideology as liberal (0) or conservative (1).

Using this approach, they estimate the ideology of a subset of users in a COVID-19 Twitter dataset (Chen et al., 2020). The COVID-19 dataset is contemporary with the 2020 U.S. Election Twitter dataset that we use in this paper, and has a significant overlap of users. We adopt their identified ideology scores. Of the 135K users in our sample, we identify 71K as liberals and 46K as conservatives, and the rest users do not have an identified political ideology. The liberal users authored 17M tweets and conservative authored 20M tweets.

For each community, we quantify the ratio of liberal users and liberal tweets in it in Table 1. It is important to note that these 20 communities span the political spectrum, evident by the varying ratios of liberals present within them. This wide range is evident even in the largest, most conservative-leaning community (Community 1) which still includes 9% liberal members. The detected communities collectively demonstrate the diversity and variability of media consumption patterns in the online space. Each community appears to represent a unique intersection of political leanings, topical interests, and geography. For instance, some communities, such as Community 1, gravitate towards conservative news outlets, while others lean towards more liberal sources, as seen with Community 2 and 3. Another layer of differentiation comes from the specific interests or focus areas, with Community 5 showing a preference for business and Community 16 for celebrity and health-related news. Geography also play a role in news consumption, as demonstrated by outlets associated with local television news sources, like fox5ny (Community 15) and *ktla* (Community 20). Overall, these differences underscore the multifaceted nature of information consumption and sharing within different communities in an online ecosystem. These observations point out the limitations of conventional methods to probe community ideologies, which rely on a predetermined binary political division *left* vs *right* of communities, which does not conform to the organic formalization of communities.

Interactions between Online Communities. Previous works focus on isolated communities, ignoring the interactions between them (Jiang et al., 2020; He et al., 2021; Webson et al., 2020). However, retweeting is a popular user activity on Twit-

[2] We set the resolution to 1, and find that using different resolution values barely change the top 20 detected communities.

comm.	#users	#tweets	%lib. tweets	top-5 shared news outlets
1	38.9K	19.3M	9	*foxnews, breitbart, nypost, washingtonexaminer, wsj*
2	19.4k	3.9M	85	*nytimes, washingtonpost, time, wapo.st, bostonglobe*
3	15.8k	3.9M	78	*thehill, nbcnews, theguardian, vox, latimes*
4	11.5K	2.9M	87	*rawstory, huffpost, apnews, thedailybeast, politicususa*
5	10.2K	2.4M	80	*politico, businessinsider, newsweek, theatlantic, bloomberg*
6	7.5K	1.5M	69	*npr.org, forbes, reuters, msn, bbc*
7	7.1K	1.4M	86	*cnn, politico.eu, irishtimes, baltimoresun, ccn*
8	5.2K	1.1M	79	*usatoday, politifact, snopes, factcheck.org, military*
9	3.2K	0.8M	76	*abcnews.go, markets.businessinsider, c-span.org, cs.pn, sfchronicle*
10	3.0K	0.7M	31	*cnbc, nj, abc.net.au, kansascity, mcall*
11	2.1K	0.4M	75	*apple.news, sun-sentinel, seattletimes, local10, Salon*
12	1.8K	0.3M	80	*abcn.ws, reut.rs, bbc.co.uk, sacbee, azcentral*
13	1.3K	0.4M	39	*dailymail.co.uk, spectator.us, mercurynews, thewrap, nejm.org*
14	1.2K	0.3M	45	*axios, warroom.org, bostonherald, ajc, minnesota.cbslocal*
15	1.1K	0.3M	33	*politi.co, tampabay, calmatters.org, fox5ny, americamagazine.org*
16	1.1K	0.3M	52	*cbsnews, hollywoodreporter, postandcourier, modernhealthcare, the-sun*
17	1.0K	0.2M	59	*news.yahoo, christianpost, sfgate, taskandpurpose, mashable*
18	1.0K	0.2M	47	*reason, detroitnews, freep, statnews, mlive*
19	0.8K	0.2M	90	*citylab, cbs7, thestreet, palmbeachpost, houstonchronicle*
20	0.5K	0.1M	61	*miamiherald, reviewjournal, ktla, kvue, on.ktla*

Table 1: Statistics of the 20 largest communities in the *news co-sharing network* of the 2020 Elections Twitter data. Five most popular news outlets are listed for each community. The liberal and liberal-leaning news outlets are highlighted in blue, and the conservative and conservative-leaning outlets are highlighted in red. Outlets with no overt political bias are shown in black.

ter. By retweeting, users endorse the message conveyed in the original tweets (Jiang et al., 2023; Barberá, 2015). In our dataset, ~80% tweets are either retweets or quoted tweets[3], and we only focus the former that are more likely to signify endorsement. Therefore, utilizing messages that have been widely retweeted by a given community helps understand what information the community's members consume, including messages posted by users in other communities.

To study the interactions between communities, we construct a *community retweet network* among the 20 communities. For a retweet by a user a of a user b's message, we add an edge from the community to which user a belongs to the community where user b is a member. Self-loops are allowed in the network, where a user is retweeting another user in the same community. The edges are weighted, representing the frequency that the retweeting activities happened. For each community, we normalize the weights of its out-edges by its total out-degree. The retweet network is shown in Figure 1, where edges with weights lower than 0.05 are not shown.

From the retweet network we observe the following key takeaways: 1) Interconnectedness matters: The frequent retweets among communities highlight the importance of network interactions in understanding their ideologies. 2) Echo chamber phenomenon: Community 1's prevalent self-retweets (as indicated by the large weight of its self-loop) suggest a strong echo chamber effect, indicating certain conservative groups might be more ideologically isolated than their liberal counterparts. 3) Diverse news consumption: The different media outlets preferred by each community show that even communities with similar ideologies can have varied news consumption patterns, shaping their individual ideologies. 4) Comparative inclusivity of liberal communities: Communities 2 and 3 engage more with external content compared to Community 1, hinting at potentially broader information

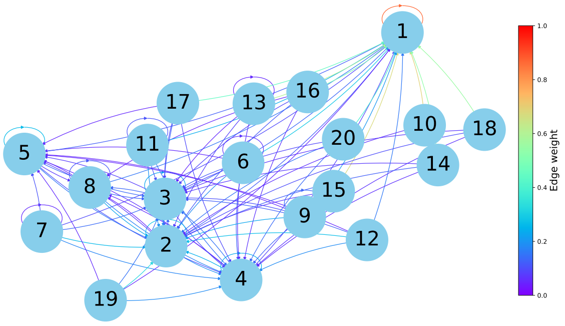


Figure 1: *Community retweet network*. The source node of an edge is the retweeting community, and the target node is the retweeted community. For each community, the weights of its out edges are normalized by its out degree. Edge colors represent the edge weights. The edges whose weights are lower than 0.05 are not shown.

[3] A quoted tweet is a retweet that has been made with a comment

consumption.

5 Probing Stances of Online Communities

To study the different opinions and stances of different communities, we delineate each community with a large language model finetuned on this community's corpus. We further use the massage passing technique to account for the information and opinion shared between communities. Finally, to verify that our models indeed capture communities' political ideology, we test it against multiple baselines on stance prediction toward 30 politically salient entities or groups. The results show the outstanding performance of our method.

5.1 Methodology

Finetuning Language Model. A community's corpus D consists of tweets made by all users within the community. For each community, we finetune a generative language model GPT-2 (Radford et al., 2019) on the corpus using the causal language modeling task. During finetuning, the language model mines insights from the community (Jiang et al., 2022).

Message Passing between Community Corpora. Given the established interconnected nature of communities in the *community retweet network*, it becomes paramount to consider these connections when fine-tuning individual language models for different communities. Drawing inspirations from Graph Neural Networks (GNNs) where nodes exchange information with their neighbors (message passing), we propose to finetune the community language models using message passing between their corpora. The intuition is that if a community C_i retweets another community C_j, then C_i is likely to share similar ideologies as C_j (Jiang et al., 2023; Barberá, 2015).

In our method, we represent the corpus of community C_i as $D_i = (t_1^i, t_2^i, ..., t_{|D_i|}^i)$, where each t_k denotes a specific tweet. The outgoing neighbors of C_i are represented as $N^+(C_i)$. The normalized edge weight, representing the strength of connection between two communities C_i and C_j, is denoted by w_{ij}. In the *community retweet network*, $N^+(C_i)$ signifies the communities that have been retweeted by C_i. It is important to note that C_i itself can be included in $N^+(C_i)$ as a community can retweet itself.

The language model of each community C_i is fine-tuned on its corresponding corpora D_i over a total of x steps, with message passing performed in intervals of y ($y < x$). During each message passing step, C_i exchanges information with its neighboring communities. This is achieved by updating its corpus to D_i':

$$D_i' \Leftarrow \sum_{C_j \in N^+(C_i)} \text{sample}(D_j, w_{ij} * |D_i|), \quad (1)$$

where D_j is the corpus of C_j, and sample(D_i, k) represents a sub-corpus comprising k tweets, randomly sampled from D_i. Note that the updated corpus D_i' is of the same size as D_i. The sum of two corpora implies their merging. This method of using message passing introduces minimal computational overhead and is highly scalable. Notably, it does not necessitate collective fine-tuning of multiple language models, which allows for more flexible and efficient training.

Utilizing message passing, we ensure that the learning process of one community-specific model benefits from the insights and nuances found in its interconnected neighbors. This approach acknowledges the reality that no community exists in isolation; they frequently influence and are influenced by their surrounding communities. By allowing the exchange of information between these models during the fine-tuning process, we harness the collective intelligence of the entire network.

5.2 Evaluation Protocol

Community Response Generation. For each finetuned community language model, we use four prompts (Jiang et al., 2022) to probe its attitude towards a target X, which represents one of 30 politically salient entities or groups (Appendix A): (1) "X", (2) "X is/are", (3) "X is/are a", (4) "X is/are the". For each target, the model generates r responses using each prompt.

Community Stance Aggregation. Following Jiang et al. (2022), we calculate the sentiment of the response and use it as a proxy of the community's stance towards the target. We use Twitter sentiment classifier *cardiffnlp/roberta-base-sentiment-latest* (Barbieri et al., 2020; Loureiro et al., 2022) to measure sentiment: negative (-1), neutral (0), or positive (1). The average sentiment score $\hat{s}_{i \to j}$ over all generated responses is a measure of community C_i's attitude towards the target t_j. Please refer to Appendix B for the reasoning behind using sentiment analysis as a proxy of stance detection.

Community Stance Reweighting. The ANES survey reports the liberal rating toward the target t_j (averaged over all liberal participants) as s_j^l, and the conservative rating (averaged over all conservative participants) as s_j^c. As we demonstrate in §4, every ad-hoc community has a mixed ideology with users from both sides. Thus, delineating the ideology of these communities entails taking into account such mixture of ideologies. As a result, we use the weighted average of the two-sided ratings from the survey by the ratios of liberal tweets and conservative tweets in the community as the ground truth score of a target. Specifically, we denote the rating (i.e., ground truth stance score) of community C_i towards the target t_j as $s_{i\rightarrow j} = r_i^l * s_j^l + r_i^c * s_j^c$, where r_i^l and r_i^c represent the ratios of liberal and conservative tweets respectively in community C_i and $r_i^l + r_i^c = 1$.

Target-specific Community Ranking. Given a target, we try to capture the stances of different communities towards it, i.e., identify which communities favor the target and which are against it. Specifically, for target t_j, we compare two lists of sentiment scores from N communities towards it: one from the model prediction $\hat{S}_{t_j} = \{\hat{s}_{0\rightarrow j}, \hat{s}_{1\rightarrow j}, ..., \hat{s}_{N\rightarrow j}\}$, and the other from the reweighted ground truth $S_{t_j} = \{s_{0\rightarrow j}, s_{1\rightarrow j}, ..., s_{N\rightarrow j}\}$. The correlation between them is measured by a ranking coefficient $\text{rank\_corr}_{t_j}(\hat{S}_{t_j}, S_{t_j})$, which varies between -1 and 1 with 0 implying no correlation. The final target-specific community ranking coefficient is averaged over all M targets, as $\frac{1}{M}\sum_{j=1}^{M} \text{rank\_corr}_{t_j}(\hat{S}_{t_j}, S_{t_j})$.

Community-specific Target Ranking. Given a community C_i, we also want to measure which targets the community favors more and which it is against. Given two lists of sentiment scores from the language models and reweighted ground truth of community C_i towards M targets, the ranking coefficient between them is $\text{rank\_corr}_{C_i}(\hat{S}_{C_i}, S_{C_i})$ The final community-specific target ranking coefficient is averaged over all N communities, as $\frac{1}{N}\sum_{i=1}^{N} \text{rank\_corr}_{C_i}(\hat{S}_{C_i}, S_{C_i})$.

5.3 Baselines

We compare our finetuned language model with message passing between corpora to the following baselines.

Pretrained GPT-2 (Radford et al., 2019). The vanilla pretrained GPT-2. To adapt the model to different communities with varying ratios of liberals and conservatives, when generating responses we append a context to the prompt: "As an independent who agrees with Democrats x% percent of the time and Republicans y% percent of the time, I think," where x and y represent the ratios of liberal and conservative tweets in that community.

Pretrained GPT-3 (Brown et al., 2020). The original GPT-3 Ada. The same context is used for generating responses as for the pretrained GPT-2. The generations are obtained by querying the API. We do not use ChatGPT because it refuses to generate personal opinions or beliefs.

Finetuned GPT-2 (Jiang et al., 2020). GPT-2 finetuned on each community corpus independently, without using interactions between communities.

5.4 Experimental Setup

Tweet Processing. We removed mentions, hashtags, emojis, and URLs (after constructing the *news co-sharing network*) from the tweet texts. For tweets that are cut off by an ellipsis due to exceeding the max length in querying the Twitter API, we removed the ellipsis as well as the characters preceding it.

Backend Language Model. Following Jiang et al. (2020), we pick GPT-2 as our backend generative language model. We do not use a bigger open-sourced language model like LLaMA (Touvron et al., 2023) for the following reasons. First, our goal is to proactively predict opinions towards people or groups. Therefore, for fair evaluation, the language model should be pretrained on data generated before April 2020 when the ANES survey was conducted. However, recent large language models are pretrained using data after this time. Second, we argue that our method to finetune language models with corpora message passing to probe community ideologies is highly portable and can be used with any backend language model. By demonstrating its effectiveness on GPT-2, we believe that it will generalize to larger language models.

Model Finetuning. Please refer to Appendix C.

Evaluation. For a finetuned GPT-2 model on a community, it generates 1,000 responses for a target using each prompt with greedy decoding. We run the generations for 5 times with different random seeds. The average performance over different runs are reported. For the GPT-3 Ada model, we only query it once with 1,000 responses due to the cost. We use Spearman's rank correlation coeffi-

	Pretrained GPT-3		Pretrained GPT-2		Finetuned GPT-2		Finetuned GPT-2 + MP	
	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall
P1	0.10*	0.06	0.12+-0.06*	0.09+-0.04*	0.19+-0.01*	0.14±0.01*	**0.25±0.04***	**0.20±0.03***
P2	0.08	0.06	0.21±0.05*	0.16±0.04*	0.22±0.03*	0.17±0.02*	**0.23±0.01***	**0.19±0.01***
P3	0.08	0.05	0.19±0.03*	0.14±0.02*	0.19±0.02*	0.14±0.01*	**0.20±0.01***	**0.17±0.01***
P4	0.08	0.06	0.18±0.05*	0.13±0.04*	0.16±0.02*	0.12±0.02*	**0.24±0.01***	**0.19±0.01***

(a) Results on target-specific community ranking. Reported correlations are averaged over all targets.

	Pretrained GPT-3		Pretrained GPT-2		Finetuned GPT-2		Finetuned GPT-2 + MP	
	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall
P1	0.02	0.01	0.02±0.03	0.01±0.02	-0.03±0.04	-0.02±0.03	**0.06±0.01**	**0.06±0.01**
P2	-0.03	-0.02	0.02±0.01	0.01±0.01	0.01±0.03	0.01±0.02	**0.10±0.01***	**0.09±0.01***
P3	0.04	0.02	0.06±0.01	0.04±0.008	0.04±0.02	0.03±0.02	**0.10±0.02***	**0.10±0.01***
P4	-0.08	-0.06	0.00±0.04	-0.01±0.03	0.05±0.02	0.03±0.01	**0.13±0.02***	**0.11±0.02***

(b) Results on community-specific target ranking. Reported correlations are averaged over all communities.

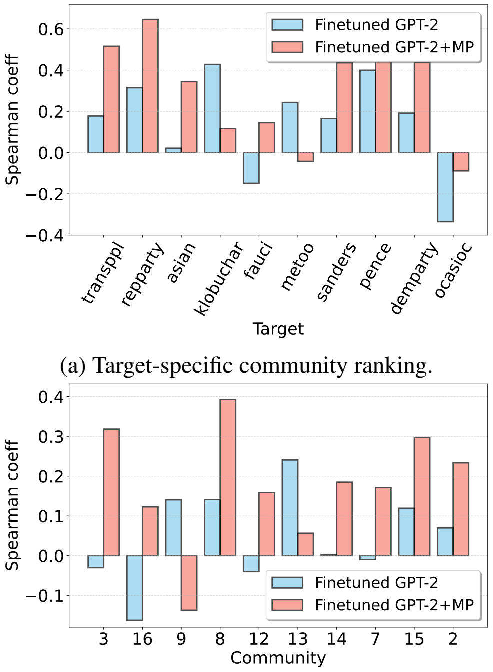
Table 2: Spearman and Kendall tau rank correlation coefficients on two ranking tasks. The targets are entities and groups in the ANES survey. The coefficients measure the correlation of the ranking of model's predictions of community's stances towards the targets to the ground truth ranking obtained from the ANES survey. P1 through P4 stand for the four prompts used to query the model: (1)"X", (2)"X is/are", (3) "X is/are a", and (4) "X is/are the". "MP" stands for message passing. The best results using different prompts on Spearman correlation and Kendall tau are highlighted in bold. * indicates statistical significance at the $p < 0.05$ level.

cient and Kendall's tau as the metrics for evaluating the two ranking tasks.

5.5 Results

The overall results on target-specific community ranking and community-specific target ranking are shown in Table 2a and 2b. First, using messaging passing between community corpora (our method) achieves state-of-the-art performance, consistently outperforming all baselines on all prompts and metrics. It is worth noting that in contrast to Jiang et al. (2020), who use classification task to decide which of the two communities favors a target more, the ranking tasks we use to evaluate performance over multiple communities and targets are much more challenging. Second, finetuned GPT-2 does not have non-trivial performance improvement compared to the original pretrained GPT-2. We argue that this is because the tweet corpora of most communities are relatively small and insufficient for finetuning a language model on the causal language modeling task. However, our method allows the language models to make use of more text from other communities sharing similar beliefs, creating a more prominent gain over the baseline. Moreover, pretrained GPT-3 Ada, with a more sophisticated architecture and trained on more data, underperforms the much simper pretrained GPT-2. Reviewing the retrieved responses from the API, we found that GPT-3 produced shorter and more factual texts, containing less personal opinions towards targets. We hypothesize that GPT-3 was tuned this way due to safety concerns. Finally, out of the two ranking tasks, community-specific target ranking is a harder task, where the model needs to capture the intrinsic differences in attitudes within a community towards the targets. Here no baseline can capture any correlation with the ground-truth (Table 2b). This is even more challenging when one community barely mentions the target, providing the language model little information to learn about it. However, our method allows the language model to learn about the target from the neighboring communities which the community retweets. This improves the learned community insights, increasing the correlations in Table 2b from 0 to 0.1.

In-depth Analysis. Figure 2a shows the Spearman coefficients with largest differences on the target-specific community ranking task using Prompt 4 for ten communities, between the finetuned GPT-2 baseline and our method using message passing. Similarly, Figure 2b shows coefficients with largest differences on the community-specific target ranking task. We observe that for most targets and communities, message passing leads tos a higher correlation score. Notably, on "Asians", Community 3, 12, 14, and 17, the correlation score improves from 0 a positive, suggesting that message passing helps the language model learn community's attitudes towards this target. However, our method underperforms on a few targets, such as "#MeToo movement". The MeToo Movement touches upon deeply rooted societal issues and experiences. The richness and complexity of sentiments associated



(a) Target-specific community ranking.

(b) Community-specific target ranking.

Figure 2: Spearman's rank correlation coefficients using Prompt 4 for 10 targets/communities of the finetuned GPT-2 baseline and our method on two ranking tasks. The 10 targets/communities are the ones with the largest coefficient change between the two methods, either positively or negatively. From left to right, the targets/communities are sorted by their performance changes. All results are statistically significant at $p < 0.05$ level.

with this movement might be diluted or confused when aggregating messaging across communities, thereby diminishing the model's performance.

Ablation Study on Random Message Passing. A plausible counter-argument could be that the enhancement observed through our message passing approach merely results from an enlargement of each community's finetuning data pool. According to this perspective, one could just as easily enrich each corpus by drawing randomly from other community corpora, negating the need for a reference to the *community retweet network*. In light of this, we conduct an ablation study, creating an alternative community retweet network with edge weights between communities assigned randomly. In this network the message passing does not follow the communities retweeting activities. Comparisons between this random message passing method and our approach are illustrated in Table 3. Observations indicate that models finetuned with random message passing tend to underperform, providing a robust argument that our proposed method of finetuning via message passing, informed by the *community retweet network*, cannot be reduced to a simplistic random data augmentation for each community's corpus. This further validates the crucial role played by the *community retweet network* in directing the information flow and helping each community language model learn more relevant information.

	Finetuned GPT-2	**Finetuned GPT-2**
	+Random MP	**+ MP**
P1	0.20±0.0	0.25±0.04
P2	0.18±0.03	0.23±0.01
P3	0.14±0.02	0.20±0.01
P4	0.22±0.02	0.24±0.01

Table 3: Spearman rank correlation of our method and an ablated method where each community exchanges information following a community retweet network whose edge weights are randomly assigned. All results are statistically significant at $p < 0.05$ level.

6 Conclusion

We explore the complex ideologies of ad-hoc online communities towards different political figures and social groups. Our approach probes these ideological stances by finetuning language models on community-authored tweets and exchanging community information through message passing. Our method aligns with real-world survey data and outperforms existing baselines. Our work underscores the potential of leveraging social media data to monitor and understand societal dynamics in the digital age.

Our method offers a promising pathway for future research. Potential avenues include expanding the study to other social media platforms, analyzing how ideological stances of online communities evolve over time, and finetuning a single language model for different communities. Our approach also holds the promise of providing an in-depth exploration of intricate ideological postures of the communities, facilitating a broader array of applications, including the examination of community emotional reaction to wedge issues (Guo et al., 2023) and affective polarization (Iyengar et al., 2019).

Limitations

Our study, while valuable, does have several limitations that must be acknowledged. First, our research primarily focuses on Twitter, a single social media platform. This may limit the generalizability of our findings, as user behavior and community dynamics can vary significantly across different platforms. Secondly, we concentrate primarily on U.S. based English-speaking communities. This focus restricts the applicability of our findings, as language nuances, cultural factors, and political landscapes can greatly affect the expression and perception of ideologies in online communities. Additionally, our method relies heavily on the quality of community retweet networks for information exchange. If the underlying network is not well-constructed or does not accurately reflect community interactions, it may compromise the effectiveness of our approach. Moreover, our model also assumes that communities are static and does not account for potential temporal changes in community formation, sentiments, and interactions. In reality, these elements can dynamically evolve over time. These limitations highlight valuable areas for future research and should be taken into account when interpreting the findings of our study.

Ethics Statement

Our study investigates online communities on Twitter, focusing on their political orientations and the propagation of different ideological stances. While this understanding is essential for addressing societal challenges such as misinformation and polarization, we are aware that our work could potentially be misused. For instance, our methods could be exploited to manipulate public opinion or target specific communities for propaganda or harassment. We condemn such misuse and advocate for the responsible application of our research findings.

Regarding data privacy, we employ publicly available Twitter data, respecting the platform's guidelines. No personal identifying information is used in our analysis, maintaining user anonymity. We acknowledge the potential risks of re-identification and take precautions to minimize this risk.

We also recognize that our work might unintentionally perpetuate biases present in the data, given that the language models are trained on real-world data, which might reflect societal biases. As such, the models' ideology probing could potentially reinforce and amplify these biases. Efforts were made to mitigate this risk by ensuring the diversity of the communities studied and clearly acknowledging this limitation in our research.

Overall, we believe that the potential benefits of our research, such as enabling better understanding of online communities and fostering healthier online discourse, outweigh these risks. However, we emphasize the need for continued ethical consideration and caution as the research progresses and its findings are put to use.

References

Emily Allaway and Kathleen Mckeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*.

Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 258–265. IEEE.

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

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Emily Chen, Ashok Deb, and Emilio Ferrara. 2021. # election2020: the first public twitter dataset on the 2020 us presidential election. *Journal of Computational Social Science*, pages 1–18.

Emily Chen, Kristina Lerman, Emilio Ferrara, et al. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR public health and surveillance*, 6(2):e19273.

Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language models trained on media diets can predict public opinion. *arXiv preprint arXiv:2303.16779*.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. *Berkman Klein Center Research Publication*, 6.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Siyi Guo, Zihao He, Ashwin Rao, Eugene Jang, Yuanfeixue Nan, Fred Morstatter, Jeffrey Brantingham, and Kristina Lerman. 2023. Measuring online emotional reactions to offline events. *arXiv preprint arXiv:2307.10245*.

Zihao He, Negar Mokhberian, António Câmara, Andres Abeliuk, and Kristina Lerman. 2021. Detecting polarized topics using partisanship-aware contextualized topic embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2102–2118.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77.

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.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. Communitylm: Probing partisan worldviews from language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6818–6826.

Julie Jiang, Emily Chen, Shen Yan, Kristina Lerman, and Emilio Ferrara. 2020. Political polarization drives online conversations about covid-19 in the united states. *Human Behavior and Emerging Technologies*, 2(3):200–211.

Julie Jiang, Xiang Ren, and Emilio Ferrara. 2023. Retweet-bert: political leaning detection using language features and information diffusion on social networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 459–469.

Ashiqur R KhudaBukhsh, Rupak Sarkar, Mark S Kamlet, and Tom Mitchell. 2021. We don't speak the same language: Interpreting polarization through machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14893–14901.

Jon Kingzette, James N Druckman, Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan. 2021. How affective polarization undermines support for democratic norms. *Public Opinion Quarterly*, 85(2):663–677.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

George Lakoff. 2014. *The all new don't think of an elephant!: Know your values and frame the debate*. Chelsea Green Publishing.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260.

Jeremiah Milbauer, Adarsh Mathew, and James Evans. 2021. Aligning multidimensional worldviews and discovering ideological differences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*, pages 206–219. Springer.

Mohsen Mosleh and David G Rand. 2022. Measuring exposure to misinformation from political elites on twitter. *nature communications*, 13(1):7144.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ashwin Rao, Fred Morstatter, Minda Hu, Emily Chen, Keith Burghardt, Emilio Ferrara, and Kristina Lerman. 2021. Political partisanship and antiscience attitudes in online discussions about covid-19: Twitter content analysis. *Journal of medical Internet research*, 23(6):e26692.

Ashwin Rao, Fred Morstatter, and Kristina Lerman. 2022. Partisan asymmetries in exposure to misinformation. *Scientific Reports*, 12(1):15671.

Kate Starbird. 2017. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 230–239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sze-Yuh Nina Wang and Yoel Inbar. 2021. Moral-language use by us political elites. *Psychological Science*, 32(1):14–26.

Albert Webson, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick. 2020. Are “undocumented workers” the same as “illegal aliens”? disentangling denotation and connotation in vector spaces. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4090–4105.

Sam Whitt, Alixandra B Yanus, Brian McDonald, John Graeber, Mark Setzler, Gordon Ballingrud, and Martin Kifer. 2021. Tribalism in america: behavioral experiments on affective polarization in the trump era. *Journal of Experimental Political Science*, 8(3):247–259.

A ANES Survey

30 targets studied in the ANES survey: (1) *people*: Donald Trump, Barack Obama, Joe Biden, Elizabeth Warren, Bernie Sanders, Pete Buttigieg, Kamala Harris, Amy Klobuchar, Mike Pence, Andrew Yang, Nancy Pelosi, Marco Rubio, Alexandria Ocasio-Cortez, Nikki Haley, Clarence Thomas, Dr. Anthony Fauci, and (2) *groups*: blacks, whites, Hispanics, Asians, illegal immigrants, feminists, the #MeToo movement, transgender people, socialists, capitalists, big business, labor unions, the Republican Party, the Democratic Party.

B Community Stance Aggregation

The reason on using sententiment analysis as a proxy of stance detection. Admittedly, the stance towards a target expressed in a sentence might be different from the overall sentiment of the sentence, and the most ideal case would be using a pretrained stance detection (He et al., 2022; Allaway and Mckeown, 2020) model on the target to detect the stance of the generated response towards it. However, not all stance detection models pretrained on the 30 targets are publicly accessible. Nevertheless, by manually inspecting the generated responses, we find that all the generated responses are simple sentences with no convoluted semantics[4] where sentiment analysis and stance detection would produce the same result. We further validate this observation by comparing the results from the sentiment analysis model with two pretrained stance detection models on Trump and Biden for generated responses on them, which show trivial differences.

C Experimental Setup

Model Finetuning. We finetune the GPT-2 model on a Tesla A100 with 40GB memory. We use a batch size of 160 and learning rate of $5e-5$. We leave 2% of data for validation. The model is finetuned for a total of 10 epochs. When finetuning with our proposed method, message passing is conducted once after the 5th epoch, and thus every community exchanges information only with its direct neighbors.[5] The model checkpoint with best performance (loss) on the validation set is saved for further evaluation.

[4] For example, "Joe Biden is a joke. He is by no means presidential material."

[5] We experimented on more frequent message passing during training, where each community could obtain information from k-hop (k$\geq$ 1) neighbors, but we did not see non-trivial performance improvement.