Not What it Used to Be: Characterizing Content and User-base Evolution in Newly Created Online Communities

Alex Atcheson*
Vinay Koshy*
samuela7@illinois.edu
vkoshy2@illinois.edu
University of Illinois, Urbana-Champaign
Urbana, Illinois, USA

Karrie Karahalios
kkarahal@illinois.edu
University of Illinois, Urbana-Champaign
Urbana, Illinois, USA

ABSTRACT

Attracting new members is vital to the health of many online communities. Yet, prior qualitative work suggests that newcomers to online communities can be disruptive – either due to a lack of awareness around existing community norms or to differing expectations around how the community should operate. Consequently, communities may have to navigate a trade-off between growth and development of community identity. We evaluate the presence of this trade-off through a longitudinal analysis of two years of commenting data for each of 1,620 Reddit communities. We find that, on average, communities become less linguistically distinctive as they grow. These changes appear to be driven almost equally by newcomers and returning users. Surprisingly, neither heavily moderated communities nor communities undergoing major user-base diversification are any more or less likely to maintaining distinctiveness. Taken together, our results complicate the assumption that growth is inherently beneficial for online communities.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**; • **Applied computing** → **Sociology**.

KEYWORDS

Online communities, Growth, Content Moderation, Computational Social Science

ACM Reference Format:
Alex Atcheson, Vinay Koshy, and Karrie Karahalios. 2024. Not What it Used to Be: Characterizing Content and User-base Evolution in Newly Created Online Communities. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3613904.3642769

*Both authors contributed equally to this research.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642769

1 INTRODUCTION

Online communities possess a unique capacity for rapid, large-scale growth [18, 19]. This stems from both the low barrier to entry for participation and from the ability for members to participate anywhere at any time. Often, this is a strength. Communities devoted to niche topics can quickly establish themselves and attract enough members to maintain a vibrant culture [19]. Still, this capacity for growth comes with a cost. Online communities can easily expand beyond their initial audiences [16, 32]. In such cases, they might deviate from their original purposes or lose topical focus. Qualitative work suggests that community moderators and members are sensitive to this phenomenon [16, 18, 29, 32]. In interviews with moderators, Seering et al. [32] found that the perception that a community's population was undergoing a "sudden diversification," often prompted moderators to introduce new community rules to keep the community focused. In such cases, the question of *whose* interests a community serves becomes particularly salient. Communities face a choice: do they accommodate the interests of new members, or do they take measures to preserve the existing community identity?

Despite these qualitative accounts, the extent to which new communities deal with disruptive growth is unclear. A few factors work to prevent such disruptions. First, newcomers choose which communities to participate in[19]. Given the range of open communities available online, potentially disruptive newcomers may end up being steered towards communities that are a better fit. Further, prior work indicates that users can effectively learn certain community norms prior to participating for the first time [27]. Thus, diverse newcomers may be able to pick up on existing community norms and blend in seamlessly.

While prior work has looked at how exogenous population shocks can impact online communities [18, 22], to our knowledge, no work has studied organic growth in a generalizable sample of newborn communities. Filling this gap in the literature is an important step towards building a holistic understanding of online community trajectory. Further, since prior studies focus on single communities [10, 18] or small groups of communities [22], a more generalizable sample enables community designers to reason about the extent to which discovered trends may apply to their groups.

As such, we conduct a longitudinal analysis of the first two years of growth in 1,620 Reddit communities. We track two particular attributes of community identity over this period – the distinctiveness of a community's language use [39] and the diversity of a community's user base. The former measure is inspired by a rich history of research tying language use to community identity

[3, 6, 10, 12, 34, 39], while the latter is motivated by qualitative accounts of the factors that lead to changes in community norms [32]. Crucially, we believe these measures capture the underlying tension communities may face between accommodating users with a wide range of backgrounds and interests and maintaining a unique, shared identity.

We leverage a dataset of approximately 300 million comments and 12 million unique authors to answer the following research questions:

> **RQ1**: To what extent does the linguistic behavior of a community become less distinctive as the community grows? How much does this vary between communities?
> **RQ2**: To what extent do moderated communities see a stronger or weaker association between growth and distinctiveness?
> **RQ3**: To what extent does diversifying communities see a stronger or weaker association between growth and distinctiveness?

Answering these questions is beneficial to both community moderators and users. Fundamentally, answering RQ1 can clarify the trade-offs associated with growth, an attribute which is often treated as a measure of success for online communities [2, 5, 9, 17]. RQ2 and RQ3 help tease apart why some communities may be more strongly impacted by growth than others. Specifically, answering RQ2 provides preliminary evidence for whether or not existing moderation interventions mitigate loss of community identity, a perceived harm of growth [16, 32]. Meanwhile, answering RQ3 can help clarify the kinds of growth that most strongly impact communities.

Using a hierarchical modeling approach, we find that, on average, subreddits experience a small drop in linguistic distinctiveness during periods of growth, providing some evidence for a trade-off between distinctiveness and growth. One potential explanation for this phenomenon is that during growth periods, newcomers introduce more generic language into the community. We find some evidence for this – on average, newcomers use slightly less community-specific language. However, this difference is likely too small to fully explain the observed distinctiveness drops. Surprisingly, we find that neither moderation nor community diversification are significantly associated with changes to distinctiveness during growth periods. Taken together, our results both support and complicate hypotheses derived from qualitative work [16, 18, 22, 32]. Although there already exists a large body of qualitative work demonstrating how moderators can steer an online community's culture, we believe our findings highlight the need for more nuanced quantitative work to complement the existing literature.

2 RELATED WORK

In this paper, we study the relationships between several attributes of online communities: growth, moderation, diversity of contributors, and distinctiveness of linguistic behavior. Prior work has provided preliminary evidence that these variables are related, though no studies have evaluated their relationship at the scale present in this paper.

2.1 Norms and Growth in Online Communities

The adoption of distinct norms has long been a feature of online communities. Burnett and Bonnici [4] surface this in an early qualitative study of Usenet newsgroups. They distinguish between explicit norms (e.g., rules and community FAQs) and implicit norms (uncodified community habits). More recent studies raise the possibility that newcomers may actually shift norms in the communities they join rather than merely adapting to the existing norms [7, 10, 11, 16, 22, 31].Danescu-Niculescu-Mizil et al. [10] provide evidence for this, finding that newcomers within an online community tend to use trending vocabulary. Similarly, Dev et al. [11] find substantive differences between newcomers and long-time users in StackExchange Q&A groups. Despite preliminary evidence that newcomers act differently, evidence that they drive changes in norms is limited. Lin et al. [22] examine a small set of Reddit communities that received a massive influx of newcomers, finding no evidence for changes in language use. On the other hand, Chan et al. [7] find that influxes of newcomers in subreddits lead to more comment removals by moderators, suggesting that newcomers may be more likely to violate explicit community norms.

2.2 Moderation and Norms

On Reddit, volunteer moderators are often responsible for setting the explicit norms within their communities [31]. A substantial amount of research has explored the kinds of norms that moderators choose to enforce [8, 13, 30, 32]. Within a platform, it is common to see a wide range of explicit norms across communities [8, 14] .Grimmelmann [14] argues that this variation in norms helps to promote diverse content.

Mechanisms for enforcing explicit norms vary. Penalties, such as bans and content removals, are most common [21], though example-setting [31] and mediation [30] are also utilized. The latter two strategies may also help to enforce both implicit and explicit community norms.

2.3 Moderation and Growth

Moderators play an important role in managing activity as communities grow. Seering et al. [32] finds that moderators may deal with a greater frequency of norm violations as communities outgrow an initially homogeneous user group. Qualitative work highlights a few attributes of communities that contribute to effective newcomer management [18], highlighting the technological and leadership capabilities of moderator teams, as well as a shared sense of responsibility amongst the community [18]. In an empirical analysis of several subreddits that experienced massive, exogenous growth, Lin et al. [22] find that moderator responsiveness was associated with positive perceptions of content quality amongst community members. However, prior work has shown that moderation can lead to churn in community membership, suggesting that over-moderation may negatively affect growth [15, 33].

While we motivate our work through qualitative accounts of user experiences in online communities, we emphasize that we do not make explicit value judgments about any of our study measures. Rather, our measures were chosen based on qualitative accounts of community evolution.. We contribute to existing literature by

analyzing the relationship between all four of our measures in the context of a large-scale, generalizable dataset.

3 DATA COLLECTION

In this section, we discuss how we selected the 1,620 subreddits we studied and the data we collected to characterize their evolution over time.[1]

3.1 Subreddits

Following Mensah et al.'s [26] study of growing subreddits, we chose to collect data from all subreddits created between March 1, 2018 and December 31, 2019 that grew to 10,000 subscribers within two years of creation. This ensures that subreddits in our pool experienced a non-trivial amount of growth and allows us to track subreddits from their inception.

We used a three-step process to identify these subreddits. First, we used the public Pushshift comment archives to identify all subreddits containing at least one comment in 2018 or 2019. Then, we filtered out subreddits created before 2018 using the Reddit API. Finally, we used Pushshift to identify the largest recorded subscriber count for each subreddit every month. We filtered out any subreddits that never reached 10,000 subscribers.

3.2 Comments

For each subreddit in our sample, we use Pushshift to collect all comments from each subreddit's first two years of existence. For each comment, we extracted the source subreddit, author name, body text, creation time, and removal status. Because comments can be removed by moderators or deleted by their authors after being indexed by Pushshift, we use the Reddit API to update the comments' final statuses. Following prior work [35], we apply a few simple filters to drop suspected bot accounts from our dataset. These filters, and an evaluation of their efficacy, are provided in Appendix B.

3.3 Additional Filtering

We refine our dataset to ensure that each month contains enough data to support reliable content- and user-level analyses. For each subreddit, we only analyze months where at least 50 distinct users and comments are present. Because we use an auto-regressive model in some of our analysis, we drop subreddits that do not contain at least two consecutive months that meet these criteria. We filtered out "Not safe for work" (e.g. pornography) subreddits and non-English language subreddits. Our approach for identifying non-English subreddits is included in the supplement. As a result, our final dataset consists of 1,620 subreddits, 12 million unique authors, 291 million active comments, 10 million removed comments, and 17 million deleted comments.

4 DATA ANALYSIS

4.1 Subreddit Measures

For each subreddit i in each month t, we compute four community-level measures across a two-year observation period. To respect users' privacy, we exclude comments deleted by their authors from all analysis.

- *Subscribers* ($s_{i,t}$): The number of users subscribed to subreddit i in month t.
- *Removal rate* ($r_{i,t}$): The proportion of comments in month t of subreddit i that were removed by subreddit i's moderators.
- *Distinctiveness* ($d_{i,t}$): The linguistic uniqueness of commenting behavior in month t of subreddit i compared to a random sample of comments from elsewhere across Reddit.
- *Diversity* ($v_{i,t}$): The extent to which community members of subreddit i in month t vary in their participation across subreddits.

While the subscribers and removal rate measures are relatively straightforward to compute, our distinctiveness and diversity measures are more complex, relying on neural embeddings trained to capture similarity between Reddit comments and users, respectively [25, 36, 38].

4.2 Distinctiveness

To compute the distinctiveness score for a subreddit i in month t, we first generate an embedding for each comment posted to i in month t. If a month contains more than 1,000 comments, we randomly sample 1,000 comments to reduce computational cost. We compute a month-specific embedding for the subreddit using the average of these comment embeddings. To provide a baseline for Reddit-wide linguistic behavior in month t, we compute the average embedding of 10,000 comments [2] randomly sampled from across all active non-NSFW, English-speaking subreddits in that month (potentially including subreddits in our sample of 1,620). Intuitively, subreddits are considered distinctive if their linguistic behavior differs substantially from this Reddit-wide baseline. Thus, the distinctiveness score of subreddit i in month t is one minus the cosine similarity between the subreddit-specific vector and the Reddit-wide baseline vector.

We use a fine-tuned S-BERT model [28] to compute the comment embeddings. To fine-tune our model, we use the generalized end2end (GE2E) loss function [38]. Intuitively, this loss function takes in an input set of comments, grouped by subreddit, and rewards comments for: (i) being close to their subreddit's centroid and (ii) being far from other subreddits' centroids. As such, subreddits whose centroids are closer together are subreddits that are harder to distinguish from one another in terms of linguistic behavior. This makes the loss function a good fit for our purposes, since subreddits with low distinctiveness scores will be harder to distinguish from the baseline Reddit-wide sample. All fine-tuning was conducted on a separate dataset of held-out subreddits (refer to appendix D).

Formally, GE2E loss takes a batch of $N \times M$ comments, where N is the number of subreddits, and M is the number of comments per subreddit. It then computes a centroid $\vec{v}_s$ for each subreddit, as well as a similarity matrix S. This matrix contains the cosine similarities between comment embeddings $\vec{y}_{ji}$ and the computed centroid. Following the suggestion in [28], we used the modified centroid $v_s^{(-i)} = \frac{1}{M-1}\left(\sum_{m \neq i}^{M} \vec{y}_{jm}\right)$ when computing the similarity

[1] Data collection and analysis code, as well as most data, can be found in the paper's GitHub repository.

[2] In appendix C we provide a brief justification for our choice of sample size here

Figure 1: Two-dimensional projection of the embedding space of subreddits in December, 2021, with the visualization centered on r/PokemonLetsGo. The five most similar subreddits to r/PokemonLetsGo are: r/PokemonSwordAndShield (.956), r/PokemonSwordShield (.938), r/ProfessorOak (.899), r/PokemonHome (.893), and r/pokemongobrag (.879).

between v_s and y_{ji} when $j = s$. We scaled all similarities by learned parameters w and b, resulting in a matrix:

$$S_{ji,s} = \begin{cases} w \cos\left(y_{ji}, v_s^{(-i)}\right) + b & \text{if } j = s, \\ w \cos\left(y_{ji}, v_s\right) + b & \text{otherwise.} \end{cases} \quad (1)$$

The final loss function is

$$S_{ji,s} = L(y_{ji}) = \log\left(\sum_{k=1}^{N} e^{S_{ji,k}}\right) - S_{ji,j}. \quad (2)$$

For qualitative evaluation, we present the most and least distinctive subreddits in December, 2021 in table 1a and table 1c. The most distinctive subreddits tended to be goal-oriented. They focus on technical topics like legal advice, technology, or health and wellness advice. In contrast, comments in the least distinctive subreddits tended to be simple responses to humorous content. Notably, these subreddits tended to lack clear community-wide goals around commenting behavior.

4.3 Diversity

We quantify the diversity, $v_{i,t}$, of subreddit i in month t based on a set of embeddings representing users. To generate user embeddings, we use a modified, context-based word2vec procedure [20] first proposed by Waller and Anderson [36]. Intuitively, this procedure creates embeddings for *subreddits* so that subreddits (i.e., "contexts") sharing many participants (i.e., "words") will have similar embeddings. We use a skip-gram model and negative sampling to train a set of user embeddings for each month i. The training set consists of all user-subreddit co-occurrences across Reddit in month i, not just those in our sample of 1,620 subreddits. This improves the robustness of the learned subreddit representations.

Subreddit	Distinctiveness	Topic
PokemonGoFriends	.87	Gaming
dailywash	.84	Wellness
whitecoatinvestor	.82	Financial Advice
WeightLossNews	.79	Wellness
ScientificNutrition	.78	Wellness

(a) Most distinctive subreddits

Subreddit	Diversity	Topic
NuclearRevenge	.45	Storytelling
DiabloImmortal	.44	Video game
StoicMemes	.42	Humor
karens	.42	Storytelling
TheCircleTV	.42	Television

(b) Most diverse subreddits

Subreddit	Distinctiveness	Topic
MXRplays	.06	YouTuber
Helluvaboss	.07	Television
NuxTaku	.07	Twitch Streamer
NuxTakuSubmissions	.07	Twitch Streamer
SrGrafo	.07	Web Comic

(c) Least distinctive subreddits

Subreddit	Diversity	Topic
Market76	.02	Gaming
BruceDropEmOff	.03	Streamer
PokemonHome	.04	Gaming
SaintRampalJi	.04	Spirituality
Etorr	.04	YouTuber

(d) Least diverse subreddits

Table 1: Top five most and least distinctive and diverse subreddits in our sample. Scores were pulled from the last available month of data for each subreddit. A full list of subreddits sorted by distinctiveness and diversity scores are included in the supplement.

In each month, we represent a user as a weighted average of the embeddings of all subreddits that appear in their commenting history. These embeddings are weighted based on the number of contributions made by a user to a given subreddit. For each subreddit in our pool, we compute the centroid of embeddings of participating users in a given month. The diversity score of a subreddit is one minus the weighted average of the cosine similarities between each participant and the computed centroid. If there are more than 1,000 active users in a month, we randomly select 1,000 users with which to compute the score. Intuitively, if a subreddit consists of users who mainly participate in the same subreddits, the average cosine similarity between user embeddings and the subreddit centroid will be large, making the diversity score small. To train the embeddings in each month, we use the same hyperparameters as Waller and Anderson [36].

Table 1b and table 1d contain the most and least diverse subreddits in our sample in December, 2021. The least diverse subreddits featured insular commenters who rarely commented in other subreddits. In contrast, the most diverse subreddits were associated with more mainstream and general-interest topics. A full list of subreddits sorted by distinctiveness and diversity scores in December 2021 are included in the supplementary materials.

4.4 Bayesian Linear Regression Analysis

We used a Bayesian auto-regressive linear regression model to analyze the month-to-month changes in distinctiveness of each subreddit in our sample. We model each month-to-month change as a function of the growth in subscribers that a subreddit experienced from month $t-1$ to month t. We measure growth by computing the difference in log-scaled subscriber counts between consecutive months (i.e., $\log(s_{i,t}) - \log(s_{i,t-1})$), allowing us to estimate the effect of growth on changes to community distinctiveness, answering RQ1.

RQ2 and RQ3 focus on understanding the extent to which the growth-distinctiveness association is moderated by other variables. Thus, we include an interaction term, γ, between growth and removal rate (RQ2) and an interaction term, η, between growth and diversification (RQ3). Although they do not directly answer to our research questions, we also include terms ψ and ρ for modeling the main effects of removal rate and diversification, respectively.

To assess the variation in trends across subreddits, we model the growth coefficients (α_i) hierarchically. In other words, each subreddit has its own coefficient governing the association between growth and change in distinctiveness. We model growth coefficients as being drawn from a Normal distribution with mean μ_α and standard deviation σ_α. This allows us to reason about the "typical" association between growth and change in distinctiveness and how this association varies between communities. We also include varying subreddit-specific intercepts β_i to account for the first month of data, where no distinctiveness change can be observed. Month-specific varying intercepts (θ_T) were used to account for real-world events, such as the start of the COVID-19 pandemic, that may have shifted Reddit-wide linguistic behavior. We use $T(i,t)$ to refer to the absolute month corresponding to the t-th month of data for subreddit i.

Because the range of distinctiveness scores is between 0 and 1, we use a Beta likelihood function with logit-link to ensure that the model predictions $d_{i,t}$ also range between 0 and 1. This yields the following final model:

$$
\begin{aligned}
d_{i,t} &\sim \text{Beta}(\delta_{i,t}, \phi) \\
\text{logit}(\delta_{i,t}) &= \text{logit}(\delta_{i,t-1}) + \alpha_i(\log(s_{i,t}) - \log(s_{i,t-1})) + \psi r_{i,t} \\
&\quad + \rho(v_{i,t} - v_{i,t-1}) + \theta_{T(i,t)} \\
\text{logit}(\delta_{i,0}) &= \beta_i + \theta_{T_i,0} \\
\alpha_i &= \bar{\alpha_i} + \gamma r_{i,t} + \eta(v_{i,t} - v_{i,t-1}) \\
\beta_i, \bar{\alpha_i} &\sim MVNormal([\mu_\beta, \mu_\alpha], \Sigma) \\
\theta_T &\sim Normal(\mu_\theta, \sigma_\theta)
\end{aligned}
\tag{3}
$$

We use a multivariate-normal prior for the subreddit-specific slope and intercept terms following McElreath [24]. Prior distributions for all model parameters are included in Appendix E.

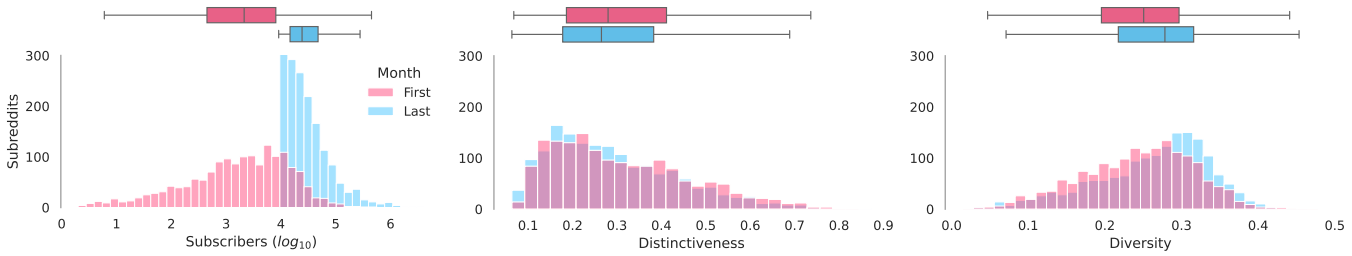

Figure 2: Histogram and box plots depicting distributions of community size, distinctiveness, and diversity of subreddits both early in their existence and approximately two years later.

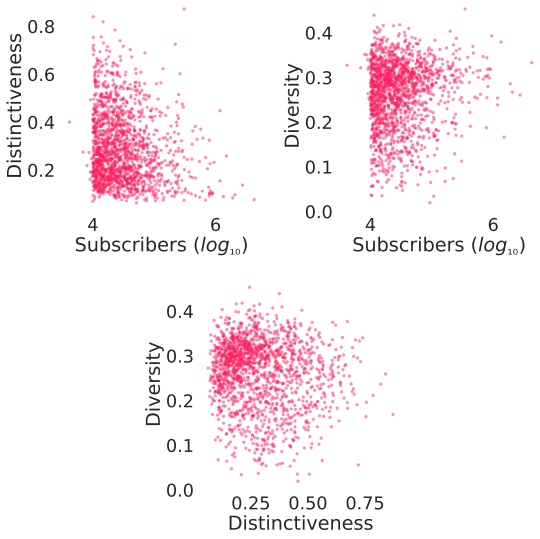

Figure 3: Scatter plots depicting relationships between community size, distinctiveness, and diversity.

5 RESULTS

Because our distinctiveness and diversity measures are novel, we use section 5.1 and section 5.2 to present summary statistics that help build intuition. We then answer our primary RQs using the hierarchical linear model in section 5.3. Finally, in section 5.4 we explore mechanistic explanations for the regression result by comparing newcomer and returning user commenting behaviors.

5.1 Patterns in Community Growth, Distinctiveness, and Diversity

We first characterize changes in diversity and distinctiveness over the study year range. We compare the distributions of subscriber counts, diversity scores, and distinctiveness scores in each subreddit's first and last available month of data. On average, the first available month was between months four and five ($M = 4.66$, $SD = 5.57$). This is because many subreddits had little-to-no activity in the first few months of data. The last available month was between months 22 and 23 on average ($M = 22.1$, $SD = 4.58$).

fig. 2 illustrates the distribution comparisons. 60.5% of subreddits saw a decrease in distinctiveness ($M_\delta = -.019$, $SD_\delta = .088$), while 66.9% of subreddits saw an increase in diversity ($M_\delta = .019$, $SD_\delta = .061$). 97.1% of subreddits saw an increase in subscribers ($M = 54,040$, $SD = 176,517$), providing an extra level of validation that our inclusion criterion successfully identified growing subreddits.

5.2 Correlations between Measures

We now look at the covariation of measures at a fixed point in time. We calculate Pearson's r between each pairing of community size, distinctiveness, and diversity in the last month of data for each subreddit. We find a slight positive correlation between size and diversity ($r = .0764$, $p < .05$, $CI_{95\%}$ $[-.0285, .121]$), a slight negative correlation between size and distinctiveness ($r = -.168$, $p < .05$, $CI_{95\%}$ $[-.212, -.118]$), and a slight negative correlation between diversity and distinctiveness ($r = -.118$, $p < .05$, $CI_{95\%}$ $[-.161, -.0747]$).

Qualitatively, we find many subreddits that exemplify these observed trends. For instance, high-distinctiveness/low-diversity subreddits tend to be knowledge sharing groups for niche hobbies like r/IndieMusicFeedback and r/Pathfinder2e (a table-top roleplaying game). Meanwhile, low-distinctiveness/high-diversity communities tend to center on sharing humorous content and have few specific rules for commenting behavior. These include subreddits like r/CoupleMemes or r/SailorMood, two meme-sharing communities.

Still, fig. 3 demonstrates that there are many communities that run counter to the overall trend. For example, many subreddits related to physical and mental wellness (e.g., r/waterbros, r/veggieshake, and r/LifeAfterSchool) scored high on both diversity and distinctiveness. Low-diversity/low-distinctiveness communities tended to center on memes and casual discussion for niche user groups, like r/teenagersnew and r/Jesser (a subreddit for fans of a particular YouTuber).

5.3 Linear Modeling

We now present the results from our linear model. This model allows us to assess the association between monthly distinctiveness changes and our other study measures, answering our primary research questions. We emphasize that results from our regression should be interpreted as *comparisons* rather than causal statements. That is, our regression model allows us to estimate the average difference in monthly distinctiveness change between two subreddits that differ by one of the predictor variables.

Table 2 contains 95% credible intervals for key model parameters. Recall that our model is hierarchical in nature, meaning that each subreddit is given an individualized $\bar{\alpha}_i$ parameter. Thus, the main effect of growth on distinctiveness is subreddit-specific. μ_α corresponds to the average of these main effects across all subreddits. As expected, on average, growth is significantly associated with a decrease in distinctiveness, answering RQ1. However, the estimate for the standard deviation of these subreddit-specific main effects is large, indicating substantial variation around this mean.

Surprisingly, we do not see significant interactions between removal rate and growth (γ, RQ2) or diversification and growth (η, RQ3), conflicting somewhat with results from prior qualitative work [18, 32]. We provide several possible explanations for this in section 6.Note that although the main effect of removal rate on distinctiveness change is significant, the size of the association is extremely small.

To make the strength of associations in our model more intuitive, we visualize effect sizes in fig. 4. We plot the predicted distinctiveness change associated with different monthly growth factors. We do this for all combinations of two hypothetical comment removal rates (0% and 5%) and three hypothetical changes in diversity score (-0.1, 0, and 0.1). Consider a set of subreddits with a 0% removal rate that experiences no diversification. Our model predicts that, after doubling in size, these communities' distinctiveness scores would drop by .014 (95% CI [.012, .016]), on average. In communities that increase ten times in size, we would expect an average change in distinctiveness of .051 (95% CI [.044, .057]). To put these numbers into perspective, a change of .014 is roughly equivalent to the difference in distinctiveness between r/pothos (.513) and r/monstera (.494) – two subreddits about houseplants. A change of .051 approximates the difference between r/ratemydessert (.420) and r/ratemyplate (.370), subreddits for rating user-submitted pictures of desserts and more general food, respectively. Figure 4 also demonstrates that the predicted average distinctiveness changes are similar across different removal rate and diversification numbers, even though the main effect of removal rate is marginally significant. Although the predicted changes to distinctiveness are small, they correspond to modeled averages for changes over the span of a single month. Thus, our results are still consistent with larger changes experienced over broader time spans.

5.4 Comparing Newcomers and Returning Users

Although growth is associated with a decrease in distinctiveness, the mechanism underlying this phenomenon is unclear. One explanation is that newcomers, unfamiliar with community norms, make comments containing linguistic patterns that are different from that of the community. Thus, distinctiveness would decrease during growth periods when there are many newcomers. To test this hypothesis, we compare the distinctiveness of comments created by subreddit newcomers and returning users. For each subreddit-month pair, we define a newcomer to be a user who commented in the subreddit for the first time that month and a returner to be a user who commented in the subreddit that month and also in some previous month. We compute two distinctiveness scores for each subreddit-month pairing: a newcomer distinctiveness score, using *only* the first comment from each newcomer that month, and a returner distinctiveness score using the first comment from each returner that month. The returner-newcomer gap is the difference between these two scores. We use only the first comment for each user since newcomers may learn community norms during subsequent contributions. We compute these scores in months where there are at least 50 newcomers and at least 50 returning users. We compute these scores using a maximum of 100 newcomers or returners.

We use a hierarchical mixed effects model to analyze the data. We model the newcomer and returner distinctiveness scores as a linear function of the subreddit growth that month, plus a set of subreddit-specific random intercepts (refer to appendix F). In fig. 5 we plot our model's posterior predictions for the returner-newcomer gap. We do this for 80 randomly selected subreddit-month pairs, half of which had high overall distinctiveness (above the 90^{th} percentile) and half of which had low overall distinctiveness (below the 10^{th} percentile). Although we do find that the returner distinctiveness scores are larger than those of newcomers on average, these differences are extremely small. Further, the returner-newcomer gaps were not significantly larger during periods of growth. Thus, it is unlikely that the returner-newcomer gap entirely explains the changes to distinctiveness observed during growth periods.

6 DISCUSSION

In this work, we study changes in online community distinctiveness during growth periods. This represents one of the first attempts to validate qualitatively derived theories about community evolution and moderation [16, 18, 32]. In this section we discuss implications of our work and contextualize our findings within the literature.

6.1 A Trade-off between Growth and Distinctiveness

Our regression analysis indicates that, on average, subreddits become less distinctive during growth periods. Size and activity level are often used as measures of success in foundational online communities research [2, 5, 9, 17]. However, our findings provide quantitative support to a line of qualitative work that suggests communities may have a more complicated relationship with growth [16, 18, 32]. Hwang and Foote [16], for example, found that users felt smaller communities allowed for topical focus and niche discussion, which our linguistic analysis supports.

6.2 Mechanistic Explanations for the Growth-Distinctiveness Trade-off

To make the association between growth and distinctiveness actionable for community moderators, it is imperative to understand the mechanisms underlying this relationship. We explore one possible explanation: that newcomers' tend to use more indistinct language, reducing the community's overall distinctiveness. Although we find support for this theory, the difference between the language use of newcomers and returners is extremely small and is unlikely to fully explain the observed losses in community distinctiveness. This finding is not entirely surprising – prior work suggests that users may learn community norms before they begin actively participating [27].

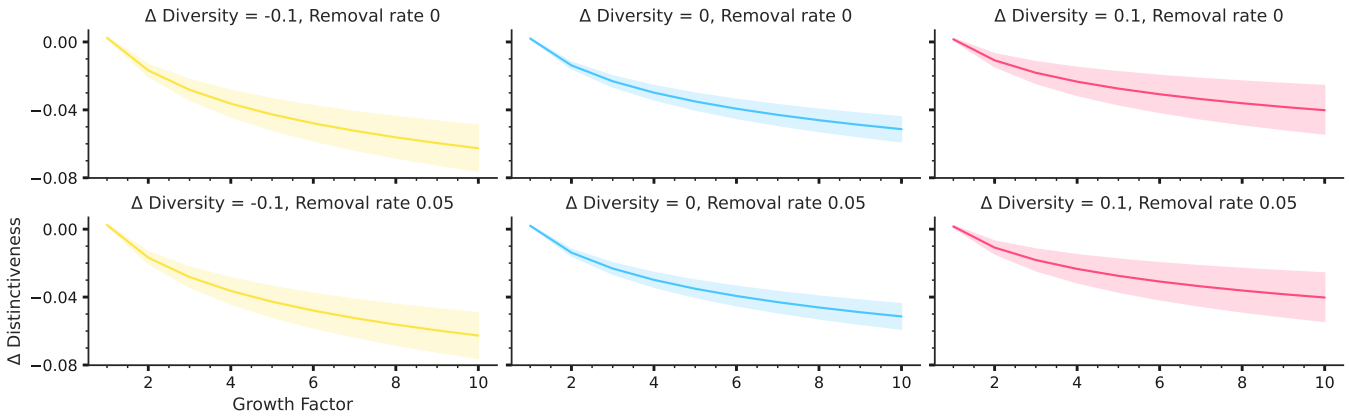

Figure 4: Model predictions of average monthly distinctiveness changes for communities that experience varying amounts of growth, diversification, and comment removal rates. In general, we see that growth is associated with larger losses of distinctiveness. These losses are comparable in communities that experience greater divresification and communities that moderate more heavily. Error bands indicate 95% credible intervals.

Parameter	Mean	95% CI	Interpretation
μ_α	-0.020	(-0.032, -0.025)	Growth coefficient mean
σ_α	0.05	(0.047, 0.053)	Growth coefficient variation
η	6.2e-4	(-0.001, 0.002)	Growth-diversification interaction
γ	-1.7-4	(-0.003, 0.002)	Growth-removal rate interaction
ρ	6.8e-4	(-0.002, 0.003)	Diversification main effect
ψ	6.2e-4	(3.8e-5, 1.16e-3)	Removal rate main effect

Table 2: μ_α and σ_α correspond to the mean and standard deviation of the subreddit-specific association between growth and distinctiveness. η corresponds to the interaction effect between growth and diversification on distinctiveness, and γ corresponds to the interaction effect between growth and removal rate on distinctiveness. ρ and ψ refer to the main effects of diversification and removal rate on distinctiveness change respectively

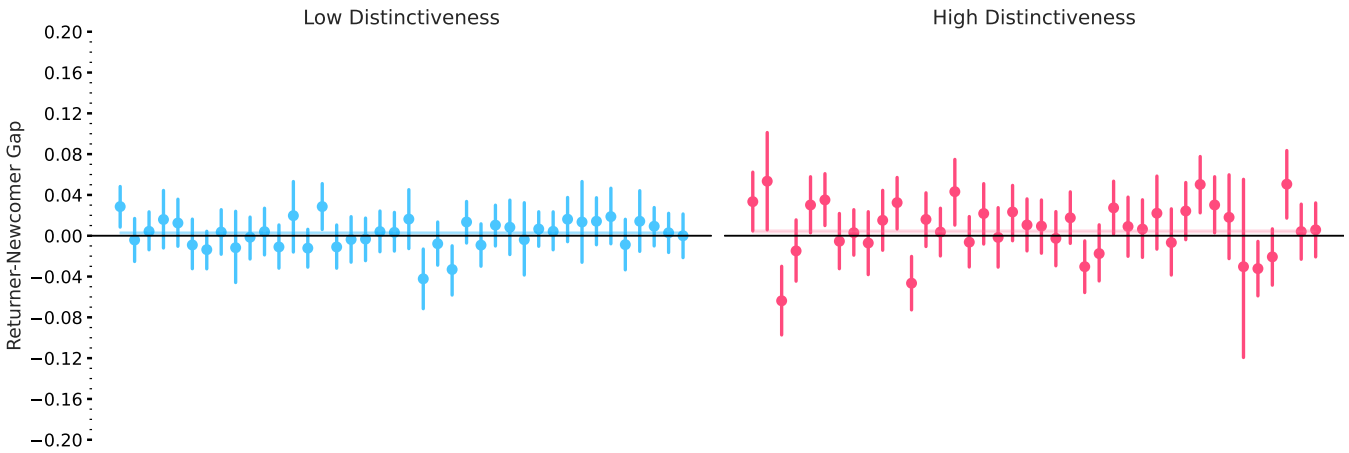

Figure 5: Model predictions of the difference between newcomer and returner distinctiveness scores for 80 randomly selected subreddit-month pairs. We find newcomers' comments are slightly less distinctive, on average, across the pairs, with some variation around this trend. Horizontal bands indicate 95% credible intervals for the means of each group.

Given the linguistic similarity between newcomer and returner contributions, we suggest two alternative explanations for the association between growth and distinctiveness, each with different implications for moderators. First, influxes of newcomers may cause decreases in community distinctiveness that affect both newcomers and returning users alike. For example, newcomers might post off-topic content which returning users end up engaging with, decreasing both groups' distinctiveness scores. If future work finds support for this mechanism, moderators should think carefully before intervening to try to preserve distinctiveness – the newer, less distinctive content may actually be engaging for old and new users alike. Second, reverse causality may be at play, whereby newcomers are more likely to join a subreddit during periods when the community's distinctiveness is lower (e.g., when the community shifts discussion towards more general-interest topics). If this is the case, interventions that attempt to lower community distinctiveness could be an effective tool for attracting new users. In a community on a technical topic, for example, moderators could cultivate a more welcoming environment for newcomers by encouraging returning users to use less jargon.

6.3 Explaining Additional Variation

Although we find an aggregate negative association between growth and distinctiveness, our hierarchical modeling approach reveals substantial variation around this trend – in some communities, growth was actually associated with an *increase* in distinctiveness. The presence of this variation is nearly as important as the observed aggregate trend, as it indicates that growing communities are not destined to lose distinctiveness. Inspired by prior work, we explore two potential sources of this variation: degree of moderation [18] and community diversification [32]. Surprisingly, we find neither degree of moderation nor diversification to be significantly associated with responsiveness to growth. Below, we provide several potential explanations for this finding.

6.3.1 Moderation. We highlight two factors could explain the discrepancy between our findings around moderation and qualitative accounts of community growth [18]. First, and perhaps most important, is that communities may decide to moderate more aggressively whenever they anticipate larger-than-normal changes to community distinctiveness. Because our regression analysis does not estimate the outcome in the counterfactual case (i.e., what would have happened to communities had they not moderated?), we could be underestimating the impact of comment removals. Second, other moderation tools might be more effective at preserving distinctiveness [30]. Moderators have a wide range of non-punitive strategies for setting, monitoring, and maintaining community norms. These include stepping in to promote desirable content, increasing the visibility of community guidelines [23], and providing alternative channels for off-topic content, like separate threads devoted to casual conversation. Interestingly, qualitative findings suggest that Reddit moderators tend to use these tools less often than moderators on other platforms, like Twitch [32]. Currently, there is little quantitative work on non-punitive moderation interventions, suggesting the need for future work on this topic.

6.3.2 Diversification. Following Seering et al. [32], we expected that communities undergoing a combination of growth and diversification would be especially likely to become less distinctive. However, we found no significant interaction between the two measures. A possible explanation is that, regardless of background, newcomers are equally able to pick up on existing community norms before participating [27]. Selection bias may play a role, as well, since users may decide to join a community only if they are already supportive of the existing norms [16]. Still, our results around diversification do not rule out the possibility that the background of incoming users effects the magnitude of the associated distinctiveness change. For example, its possible that users from specific kinds of communities, like those with higher tolerances for trolling, might be especially disruptive. A large influx of such users could actually make a community appear less diverse, even if it comes with a drop in distinctiveness.

Ultimately, communities are interested in understanding what will happen to *them*, not what will happen to the hypothetical "average" community. Knowing what kinds of communities lose more or less distinctiveness during growth periods would be invaluable to community designers, who may desire both size and distinctiveness in a community. While we made an attempt at disentangling the sources of variation underpinning the trade-off between growth and distinctiveness, we believe that this remains a ripe area for future work.

7 LIMITATIONS

Although we focus our analysis on linguistic distinctiveness, we recognize that it represents a single dimension of online community culture. A community could undergo significant changes to discussion content while remaining at the same level of distinctiveness, as measured by our embedding-based approach. Future work could conduct a similar analysis on other computational measures of community culture.

Further, while we believe that both our user and comment embeddings produce meaningful measures of distinctiveness and diversity, we acknowledge two limitations in our current approach. First, our embedding-based methodology does not accurately capture the underlying uncertainty in the data [1]. This is important given that subreddits tend to have few comments during the first several months, and because of the relative diversity of comments across Reddit. Though the overall size of our dataset helps to mitigate this concern, our analysis may be overconfident for specific subsections of the data. Second, we acknowledge that the evidence for the construct validity of both measures is largely qualitative. Although both approaches are inspired by prior work [25, 36, 37], further validation of these measures would strengthen our current analysis.

8 CONCLUSION

Drawing upon prior qualitative work [16, 18, 31], we conduct a large-scale evaluation of community evolution in growing communities on Reddit. We find that communities experience a decrease in distinctiveness and an increase in diversity over the study period. Further regression analysis indicates that growth is associated with a decrease in distinctiveness. For users interested in interacting

with like-minded others around niche topics, we demonstrate that growth may not be an inherent good. Interestingly, we find that moderation does not have a significant effect on a community's capacity to maintain distinctiveness. We highlight the need for future research into developing tools and interventions to help moderators shape group identity over the course of a community's lifespan.

REFERENCES

[1] Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics* 6 (2018), 107–119.

[2] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 44–54.

[3] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18, 2 (2014), 135–160.

[4] Gary Burnett and Laurie Bonnici. 2003. Beyond the FAQ: Explicit and implicit norms in Usenet newsgroups. *Library & Information Science Research* 25, 3 (2003), 333–351. https://doi.org/10.1016/S0740-8188(03)00033-1

[5] Brian S Butler. 2001. Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information systems research* 12, 4 (2001), 346–362.

[6] Justine Cassell and Dona Tversky. 2005. The language of online intercultural community formation. *Journal of Computer-Mediated Communication* 10, 2 (2005), JCMC1027.

[7] Jackie Chan, Aditi Atreyasa, Stevie Chancellor, and Eshwar Chandrasekharan. 2022. Community Resilience: Quantifying the Disruptive Effects of Sudden Spikes in Activity within Online Communities. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–8.

[8] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

[9] Tiago Cunha, David Jurgens, Chenhao Tan, and Daniel Romero. 2019. Are all successful communities alike? Characterizing and predicting the success of online communities. In *The world wide web conference*. 318–328.

[10] 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*. 307–318.

[11] Himel Dev, Chase Geigle, Qingtao Hu, Jiahui Zheng, and Hari Sundaram. 2018. The size conundrum: Why online knowledge markets can fail at scale. In *Proceedings of the 2018 world wide web conference*. 65–75.

[12] Jacob Eisenstein. 2015. Written dialect variation in online social media. *Charles Boberg, John Nerbonne, and Dom Watt, editors, Handbook of Dialectology. Wiley* (2015).

[13] Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.

[14] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.

[15] Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*. 163–172.

[16] Sohyeon Hwang and Jeremy D Foote. 2021. Why do people participate in small online communities? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[17] Sanjay Ram Kairam, Dan J Wang, and Jure Leskovec. 2012. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining*. 673–682.

[18] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. Surviving an" eternal september" how an online community managed a surge of newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1152–1156.

[19] Robert E Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.

[20] Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 302–308.

[21] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. Measuring the Monetary Value of Online Volunteer Work. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 596–606.

[22] Zhiyuan Lin, Niloufar Salehi, Bowen Yao, Yiqi Chen, and Michael S Bernstein. 2017. Better when it was smaller? community content and behavior after massive growth. In *Eleventh International AAAI Conference on Web and Social Media*.

[23] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.

[24] Richard McElreath. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan, 2nd Edition* (2 ed.). CRC Press. http://xcelab.net/rm/statistical-rethinking/

[25] Reid McIlroy-Young, Yu Wang, Siddhartha Sen, Jon Kleinberg, and Ashton Anderson. 2021. Detecting Individual Decision-Making Style: Exploring Behavioral Stylometry in Chess. *Advances in Neural Information Processing Systems* 34 (2021), 24482–24497.

[26] Humphrey Mensah, Lu Xiao, and Sucheta Soundarajan. 2020. Characterizing the Evolution of Communities on Reddit. In *International Conference on Social Media and Society*. 58–64.

[27] 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*, Vol. 14. 557–568.

[28] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[29] Joseph Seering. 2020. Reconsidering community self-moderation: the role of research in supporting community-based models for online content moderation. *Proc. ACM Hum.-Comput. Interact* 3 (2020).

[30] Joseph Seering, Geoff Kaufman, and Stevie Chancellor. 2022. Metaphors in moderation. *New Media & Society* 24, 3 (2022), 621–640. https://doi.org/10.1177/1461444820964968 arXiv:https://doi.org/10.1177/1461444820964968

[31] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 111–125.

[32] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443.

[33] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[34] Trang Tran and Mari Ostendorf. 2016. Characterizing the Language of Online Communities and its Relation to Community Reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 1030–1035. https://doi.org/10.18653/v1/D16-1108

[35] Veniamin Veselovsky and Ashton Anderson. 2023. Reddit in the Time of COVID. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 878–889.

[36] Isaac Waller and Ashton Anderson. 2019. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *The World Wide Web Conference*. 1954–1964.

[37] Isaac Waller and Ashton Anderson. 2020. Community embeddings reveal large-scale cultural organization of online platforms. *arXiv preprint arXiv:2010.00590* (2020).

[38] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4879–4883.

[39] Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the international AAAI conference on web and social media*, Vol. 11. 377–386.

A IDENTIFYING NON-ENGLISH SUBREDDITS

We apply simple Naive Bayes classifiers (implemented in the Python *langtext* module) to a random sample of 500 comments from each subreddit to identify non-English-speaking subreddits. If more than 50% of comments are predicted to be in English, we consider a subreddit English-speaking.

B BOT DETECTION APPROACH

After curating our comments dataset, we manually inspected known bot accounts to identify behavioral traces indicative of bot activity.

We observed that bot accounts tended to signal their bot status through either their account name or messages included in their comments. This led us to include two sets of filters for flagging bot accounts.

First we begin by examining account names. If an account name contains 'transcriber', 'automoderator', or 'savevideo', or if the name ends with some form of '-bot', '_bot', or 'bot', we flag the account. For any account flagged by this name filter, we collect the account's entire comment history. Among accounts flagged by the name filter, we mark an account as a bot account if it has either: (i) commented 10,000 or more times or (ii) commented between 10 and 10,000 times, with 80% of the comments in a single subreddit. This was designed to catch bots that create comments across all of Reddit (e.g. haikubot, which turns random reddit comments into haikus), as well as bots that assist with specific subreddit functionalities (e.g. DeltaBot, which manages a leaderboard for r/ChangeMyView).

Still, many bot accounts evade this first filter (e.g., RugScreen). For this reason, we also examine the text of each accounts comments. We flag any accounts that contain one of a handful of keyword phrases in their body text:

- ^^I ^^am ^^a ^^bot **OR** ^I ^am ^a ^bot **OR** ^^I'm ^^a ^^bot **OR** ^I'm ^a ^bot
- ^^this ^^comment ^^was ^^written ^^by ^^a ^^bot **OR** ^this ^comment ^was ^written ^by ^a ^bot
- [Info] **OR** [*Info*] **OR** [**Info**] **OR** [(Info)] **OR** [^^Info]
- This bot wants to find the best and worst Reddit bots
- I detect haikus. And sometimes, successfully
- this comment was written by a bot
- I am a bot **OR** I'm a bot
- [Source](htts://github
- Was I a good bot?
- Summon me with
- spambotdetector
- This is a bot

If an account has at least 20 comments flagged and at least 50% of its history is flagged after the first flag, we mark the account as a suspected bot. A threshold of 20 comments was chosen because manual review revealed that non-bot accounts may occasionally use the above phrases, but usually not repeatedly. An account only needs to be flagged by *either* the name filter or the comment text filter.

We evaluate our bot detection approach by manually reviewing a random sample of suspected bot and non-bot accounts. We curate a corpus of suspected bot and non-bot accounts and comments by randomly sampling 300 suspected bot accounts and 700 suspected non-bot accounts from Pushshift's aggregated comment files from March, 2018 to December, 2021. For each account, we collect 10% of the comment history. Two of the paper authors manually and individually reviewed a random assortment of 500 accounts. For each account, the authors judged if the account was a bot or not. If the author's were unsure based on the comment history, they reviewed the account's name. If still unsure, the authors reviewed the account's page on Reddit to view a greater proportion of the account's comment history, and check the account description. When comparing our automated approach to human review, we find that our approach achieves a recall of .96 and a precision of .8.

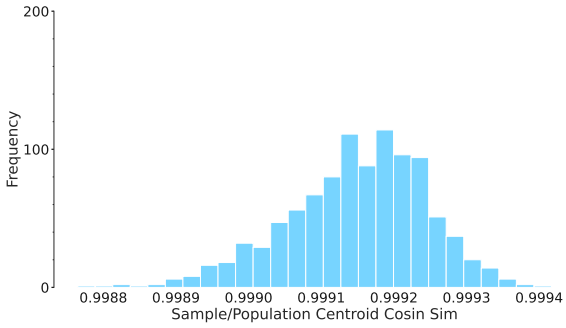

Figure 6: Histogram showing the distribution of cosine similarities between subsample centroids and the sample centroid. All 1,000 subsample similarities fall between 0.99 and 0.9995

C REDDIT-WIDE CENTROID SAMPLE-SIZE ANALYSIS

Given the diverse nature of content on Reddit, one concern is that 10,000 comments may not provide a stable estimate of the centroid of embeddings of comments across all of Reddit in a given month. To validate our choice of sample size, we conduct a bootstrapping-inspired analysis to check the generalizability of generated centroids.

Using the month of August, 2021 as a test case, we first randomly sample 100,000 comments across non-NSFW, English-speaking subreddits. We compute the centroid of the embeddings of these comments. We then randomly subsample with replacement groups of 10000 comments out of this initial sample, and compare the centroid of each subsample to the centroid of larger sample of 100,000. We repeat this subsampling 1000 times. Intuitively, if the subsample centroids are very similar to the sample centroid, we know that our subsamples generalize well. Figure 6 contains a histogram of computed cosine similarity scores. All scores fall between 0.990 and 0.9995, indicating that subsamples of 10,000 comments generalize extremely well to the larger sample.

D COMMENT-EMBEDDING MODEL FINE-TUNING DETAILS

We conduct all model fine-tuning and hyperparameter tuning on a separate set of 1,639,400 comments from 16,394 subreddits sampled during the same window of time as our main dataset. Importantly, this fine-tuning dataset contains only comments from subreddits that *do not* appear in our final sample of 1,620 subreddits. We included 100 randomly selected comments from any non-NSFW, English-speaking subreddit that had at least 100 comments in 2018 or 2019. As such, the model is tuned to produce embeddings that are generally effective at separating subreddits in the embedding space, rather than specifically learning to separate the 1620 subreddits we focus on in our study. We use similar NSFW and English language-based filters on this additional set. We use a 70%-15%-15% training-validation-test split across subreddits and conducted all

model training on a single GPU provided by Google Colab. We use the first 128 tokens of each comment to produce an embedding.

After conducting a modest grid search to select hyperparameters, we fine-tuned our embedding model with a learning rate of 5e-6 on batches consisting of 10 subreddits and 10 comments per subreddit. We fine-tuned for a single epoch. We report two measures to assess the quality of our embeddings before and after fine-tuning: the average loss over the test set, and performance on a few-shot classification task described by McIlroy-Young et al. [25]. In this task, each subreddit in the test set is given a set of 50 known "reference" comments. We then match sets of 50 unlabelled "query" comments to the subreddits in the test set. This is done by computing centroids for the reference and query sets, and then matching each query centroid to the closest reference centroid. Overall, our model achieved an accuracy of 87.6% prior to fine-tuning and 88.1% after fine-tuning, suggesting a modest performance gain. We see a larger improvement to average test set loss, with an average loss of .958 prior to fine-tuning and .867 after.

E DISTINCTIVENESS MODEL PRIORS

$$
\begin{aligned}
\Sigma &= (\Lambda\Lambda^T) * \begin{pmatrix} \sigma_\alpha^2 & \sigma_\alpha\sigma_\beta \\ \sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \\
\Lambda &\sim LKJCholesky(2) \\
\eta, \rho, \psi, \gamma &\sim Normal(0, 0.5) \\
\mu_\alpha &\sim Normal(0, 1) \\
\mu_\beta &\sim Normal(0, 1) \\
\phi &\sim Exponential(1)) \\
\sigma_\alpha, \sigma_\beta, \sigma_\theta &\sim Exponential(1))
\end{aligned}
\tag{4}
$$

* is used to denote element wise multiplication

F RETURNER-NEWCOMER GAP MODEL

Let $n_{i,t}$ refer to the newcomer distinctiveness score for subreddit i in month t and $r_{i,t}$ refer to the returner distinctiveness score for i in month t. To evaluate whether returner distinctiveness scores are consistently higher than newcomer distinctiveness scores, we fit a simple hierarchical beta regression model to the data. We model each score as a linear function of the growth experienced by the subreddit that month ($\log(s_{i,t}) - \log(s_{i,t-1})$) plus a set of subreddit-specific varying intercepts ($\beta_{n,i}$, $\beta_{r,i}$). A separate set of varying intercepts was learned for the newcomer and returner scores. These intercepts were modeled hierarchically, allowing us to compare the learned distribution of newcomer and returner intercepts across subreddits. This results in the following model:

$$
\begin{aligned}
n_{i,t} &\sim \text{Beta}(\delta_{n,i,t}, \phi_n) \\
r_{i,t} &\sim \text{Beta}(\delta_{r,i,t}, \phi_r) \\
\text{logit}(\delta_{n,i,t}) &= \alpha_n(\log(s_{i,t}) - \log(s_{i,t-1})) + \beta_{n,i} \\
\text{logit}(\delta_{n,i,t}) &= \alpha_r(\log(s_{i,t}) - \log(s_{i,t-1})) + \beta_{r,i} \\
\beta_{n,i} &\sim Normal(\mu_n, \sigma_n) \\
\beta_{r,i} &\sim Normal(\mu_r, \sigma_r)
\end{aligned}
\tag{5}
$$

We use the following prior distributions for model parameters:

$$
\begin{aligned}
\phi_n &\sim Exponential(1)) \\
\phi_r &\sim Exponential(1)) \\
\alpha_n &\sim Normal(0, 1) \\
\alpha_r &\sim Normal(0, 1) \\
\mu_n &\sim Normal(0, 1) \\
\mu_r &\sim Normal(0, 1) \\
\sigma_n &\sim Exponential(1) \\
\sigma_r &\sim Exponential(1)
\end{aligned}
\tag{6}
$$