Distinguishing In-Groups and Onlookers by Language Use

Joshua R. Minot[*]
Vermont Complex Systems Center
University of Vermont
joshua.minot@uvm.edu

Milo Z. Trujillo[*]
Vermont Complex Systems Center
University of Vermont
milo.trujillo@uvm.edu

Samuel F. Rosenblatt
Department of Computer Science
Vermont Complex Systems Center
University of Vermont
samuel.f.rosenblatt@uvm.edu

Guillermo de Anda Jáuregui
National Institute of Genomic Medicine
(INMEGEN)
Universidad Nacional Autónoma de México
gdeanda@inmegen.edu.mx

Emily Moog
University of Illinois at Urbana-Champaign
Sandia National Laboratories
ermoog@sandia.gov

Briane Paul V. Samson
Center for Complexity
and Emerging Technologies
De La Salle University
briane.samson@dlsu.edu.ph

Laurent Hébert-Dufresne
University of Vermont
laurent.hebert-dufresne@uvm.edu

Allison M. Roth
University of Florida
amr2264@columbia.edu

Abstract

Inferring group membership of social media users is of high interest in many domains. Group membership is typically inferred via network interactions with other members, or by the usage of in-group language. However, network information is incomplete when users or groups move between platforms, and in-group keywords lose significance as public discussion *about* a group increases. Similarly, using keywords to filter content and users can fail to distinguish between the various groups that discuss a topic—perhaps confounding research on public opinion and narrative trends. We present a classifier intended to distinguish members of groups from users discussing a group based on contextual usage of keywords. We demonstrate the classifier on a sample of community pairs from Reddit and focus on results related to the COVID-19 pandemic.

1 Introduction

Online communities today have unprecedented power to impact the course of disease spread (Prandi and Primiero, 2020; Armitage, 2021), sway elections (Bovet and Makse, 2019; Persily, 2017), and manipulate global markets (Anand and Pathak, 2022). However, studies of online communities are often limited to single platforms due, in part, to the fact that the overlap in users across platforms is never explicitly known or because user networks and user behavior may differ across platforms (Hall et al., 2018; Trujillo et al., 2021; Grange, 2018). Nevertheless, there are some exceptions (*inter alia* (Yarchi et al., 2021; Alatawi et al., 2021; Horawalavithana et al., 2019)) and account mapping is an area of active research (*inter alia* (Chen et al., 2020)).

A powerful alternative to account mapping is to track language rather than users, which only requires data on the content of the platform and not necessarily their user base. There remain important caveats to this approach, however: 1) shifts in language can be hard to differentiate from shifts in user demographics and 2) language *about* a group of interest can look very similar to the language *of* the group itself. This is especially true if in-group vocabulary is used by outsiders when discussing the group, or if the in-group's vocabulary percolates into the general lexicon. An example of such language spread involves the word "incel", which was popularized in a specific online community before becoming more widely known.

Here, we address the second problem of distinguishing in-group members from onlookers engaged in discussion about the in-group, based on language alone. We introduce a group-classifier,

[*] These authors contributed equally to this work

Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, pages 157 - 171
May 26, 2022 ©2022 Association for Computational Linguistics

which labels users as being in a group or discussing a group. We train our classifier on Reddit, an online forum broken into explicit sub-communities (i.e., “subreddits”). We identify pairs of subreddits, where one subreddit focuses on a particular topic (e.g., COVID conspiracies), and a second subreddit of “onlookers" discusses the first community or topic. Consistent user participation in a subreddit implies group membership, providing training labels; we filter outlier users who participate in or “troll” their chosen subreddit’s counterpart. Our classifier attempts to distinguish users from each community based on their usage of topic words.

Our contributions in this piece are focused on two main points:

1. We propose a framing for in-group and onlooker discussion communities and discuss the value of differentiating between them in downstream analyses. This point is especially important for future work on cross-platform community activity.

2. We collect a novel data set of in-group and onlooker subreddit pairs and present a baseline classification pipeline to demonstrate the feasibility of separating groups of users accounts based on the content of their posts. We go on to present preliminary results on how this automatic labelling of user accounts may affect downstream analyses relative to the ground truth data.

The rest of this manuscript is organized as follows: in Section 2 we provide an overview of prior work, mainly in the complimentary spaces of stance detection and counter speech. In Section 3 we outline our methods, including the collection of a novel dataset of subreddit pairs. In Section 4 we present the results from our in-group and onlooker classifier along with the impact of automatic labelling on resulting language distributions. We discuss the implications of our work in Section 5 and concluding remarks in Section 6. Finally, in Section 7 we suggest areas for future work which could build upon our in-group and onlooker framing, improve our classification pipeline, and address broader research questions.

2 Previous work

We classify authors as being “in a group”, or “discussing a group”, not necessarily in an adversarial way. This closely resembles stance detection (Küçük and Can, 2020; Alkhalifa and Zubiaga, 2021). Research involving stance detection may be divided into two main categories (Alkhalifa and Zubiaga, 2021):

1. Predicting the likelihood of a rumor being true (i.e., rumor detection) by examining whether the stance of posts is supporting, refuting, commenting on, or questioning the rumor (Zubiaga et al., 2016, 2018; Hardalov et al., 2021).

2. Assessing whether the stance of a post is “pro”, “against”, or “neither” with respect to any given subject (Anand et al., 2011; Augenstein et al., 2016; Joshi et al., 2016; Abercrombie and Batista-Navarro, 2018; Alkhalifa and Zubiaga, 2021).

In some cases, manually labelled datasets are used to evaluate the quality of stance detection pipelines (Joseph et al., 2021) or train stance classifiers using supervised learning (Mønsted and Lehmann, 2022).

Similar to the latter category of stance detection, topic-dependent argument classification in argument mining also parallels our classification scheme, as it may work to evaluate whether a sentence argues for a topic, argues against a topic, or is not an argument (Mayer et al., 2018; Reimers et al., 2019; Lawrence and Reed, 2020).

“Perspective identification” works to assess an author’s point of view, e.g., classifying individuals as “democrats” or “republicans” based the content of their post (Lin et al., 2006; Wong et al., 2016; Sobhani, 2017; Bhatia and Deepak, 2018). Our work also relates to the automated identification of “counter-speech”, in which hateful or uncivil speech is countered in order to establish more civil discourse (Wright et al., 2017; He et al., 2021).

Our work is similar to the form of stance detection that evaluates “pro”, “anti”, or “neither” attitudes, but the problems of stance detection tend to assume that any discussion about a group are adversarial. However, the problem of distinguishing the language *about* a group from language *of* the group is much more general, as people discussing an emerging subculture do not necessarily oppose it. For example, onlookers may talk about non-political groups formed around new music scenes, small social movements or communities surrounding specific activities without holding opposing views to these groups. Political or not, identifying these onlookers can be of critical importance when studying a specific subculture.

3 Methods

3.1 Data Selection

Reddit partitions content into “subreddits”: forums dedicated to a particular topic, with individual community guidelines and moderation policies. We identified seven (7) pairs of subreddits where one subreddit was focused on a highly-specific topic and another subreddit was dedicated to discussion about the first community. We selected clearly distinguishable communities that formed pairs of in-group and onlooking group subreddits. For example, r/NoNewNormal is a COVID-conspiracy and anti-vaccination group, while r/CovIdiots is dedicated to discussing anti-vaccination and COVID conspiracy theories (see Fig. 1 for an overview of 2-gram distributions for these subreddits). We selected this pair as our main case study because of the timeliness of the COVID-19 topic and the volume of conversation in each community. Partially owing to the contentious nature of the communities we were interested in, many of the subreddits we examined had previously been banned. Since data from banned subreddits remains available (Baumgartner et al., 2020), this did not inhibit our study or reproducibility.

Relationships between the primary community and the onlooking community were typically antagonistic. However, this does not mean that the results from standard sentiment analysis would have been able to correctly classify utterances from each group. For example, the r/NoNewNormal community may express negative opinions about vaccines or masking mandates, while r/CovIdiots may express positive sentiment about both topics, but negative sentiment about the opinions held by members of r/NoNewNormal.

For some of our subreddit pairs, the onlooker subreddit was created specifically to discuss the in-group subreddit. For example, r/TheBluePill was created in response to r/TheRedPill. For other pairs, both subreddits discussed the same topic from different viewpoints but were not directly connected. For example, r/ProtectAndServe is a subreddit populated by current and former law enforcement officers, while r/Bad_Cop_No_Donut is a subreddit dedicated to the criticism of law enforcement, but it is not specifically a criticism of r/ProtectAndServe itself. Including both types of subreddit pairs allowed us to measure the effectiveness of our classifier on communities with varying degrees of similarity.

3.2 Subreddits Chosen

The following are qualitative descriptions of each subreddit pair we examined. The size of each subreddit corpus, in terms of users and comments, as well as the mean comment score on each subreddit, can be found in the appendix (Table 4).

r/NoNewNormal and r/CovIdiots

r/NoNewNormal self-described as discussing “concerns regarding changes in society related to the coronavirus (COVID-19) pandemic, described by some as a ‘new normal’, and opposition to [those societal changes]." Most posts focused on perceived government overreach and fear-mongering. Reddit banned the subreddit on September 1st, 2021.

r/CovIdiots is dedicated to “social shaming" of covid conspiracy theorists, “anti-maskers," and “anti-vaxxers."

r/TheRedPill and r/TheBluePill

r/TheRedPill is a “male dating strategy" subreddit, commonly associated with extreme misogyny and a broader collection of “Manosphere" online communities including incels, men’s rights activists, and pick up artists.

r/TheBluePill is a satirical subreddit targeting content from r/TheRedPill.

r/BigMouth and r/BanBigMouth

r/BigMouth is an online fan community that discusses the Netflix television series, “Big Mouth." The show often features coming of age topics, including puberty and teen sexuality.

r/BanBigMouth was a community focused on associating the TV show with pedophilia and child grooming, and petitioning for the show to be discontinued and removed. Reddit banned the subreddit in June, 2021 for promoting hate.

r/SuperStraight and r/SuperStraightPhobic

r/SuperStraight was an anti-trans subreddit that defined “Super Straight" as heterosexual individuals who were not attracted to trans people. Reddit banned the subreddit for promoting hate towards marginalized groups in March, 2021.

r/SuperStraightPhobic was an antagonistic subreddit critiquing the users, posts, and intentions of the r/SuperStraight subreddit. It was banned shortly after r/SuperStraight.

r/ProtectAndServe and r/Bad_Cop_No_Donut

r/ProtectAndServe is self-described as "a place where the law enforcement professionals of Reddit can communicate with each other and the general public." Users who submit documents proving their active law enforcement status have identifying labels next to their usernames.

Bad_Cop_No_Donut is a subreddit for documenting law enforcement abuse of power and misconduct. Most posts are links to news articles, while comments discuss article content and general police behavior.

r/LatterDaySaints and r/ExMormon

r/LatterDaySaints is an unofficial subreddit for members of the Church of Latter-Day Saints. While non-members of the church are permitted to ask questions and engage in conversation, criticizing church doctrine, policy, or leadership is forbidden, and the subreddit is heavily moderated.

r/ExMormon is a subreddit for former members of the Mormon church to discuss their experiences. Posts are typically highly critical of the church.

r/vegan and r/antivegan

r/vegan is a broad vegan community, with topics ranging from cooking tips, to animal cruelty, environmental impacts of meat consumption, and social challenges with veganism.

r/antivegan is ideologically opposed to veganism. Much of the subreddit's content is satirical, or critical discussion about the actions of perceived vegan activists.

3.3 Data Collection

For each pair of subreddits, we first chose an "ending date" for data collection: If either subreddit was banned prior to the start of our study, we used the earliest ban-date as our ending date. Otherwise, we used the date of our data download. We then downloaded all comments made in the subreddit for one year prior to the ending date, using pushshift.io, an archive of all public Reddit posts and comments which is frequently used by researchers (Baumgartner et al., 2020). We then filtered out comments made by bot users, using a bot list provided by (Trujillo et al., 2021).

We anecdotally observed users from some of our selected subreddits "raiding" other selected subreddits. For example, users from subreddits opposed to the r/NoNewNormal COVID-conspiracy group sometimes harassed users in r/NoNewNormal, and vice-versa. We did not want these harassment-comments to bias our text-analysis, so we filtered out all users who had an average comment-score less than unity for their comments in the subreddit. In other words, we only kept comments from users that the community did not strongly disagree with. This did not filter out coordinated attacks, where many members of one community raided another, upvoted their raiding comments, and downvoted the in-community comments. However, this type of attack (often referred to as "brigading") is a bannable offense on Reddit, and we did not observe it in our dataset.

3.4 Determining In-Group Vocabulary

To compare the n-gram distributions of pairs of subreddits we used rank-turbulence divergence (RTD) (Dodds et al., 2020). We used RTD to both summarize overall divergence and highlight specific n-grams that contributed most to this divergence value. We found RTD to be an effective choice when making more nuanced comparisons between the disjoint distributions of subreddit pairs. It avoids construction of the mixed-distribution found in other divergence measures—such as Jensen-Shannon divergence (JSD)—which may be less effective at highlighting salient terms with the subreddit-scale distributions.

The rank-turbulence divergence between two sets, Ω_1 and Ω_2, is calculated as follows,

$$\begin{aligned} D^R_\alpha(\Omega_1||\Omega_2) &= \sum \delta D^R_{\alpha,\tau} \\ &= \frac{\alpha+1}{\alpha} \sum_\tau \left| \frac{1}{r^\alpha_{\tau,1}} - \frac{1}{r^\alpha_{\tau,2}} \right|^{1/(\alpha+1)}, \end{aligned}$$

where $r_{\tau,s}$ is the rank of element τ (n-grams in our case) in system s and α is a tunable parameter that affects the impact of starting and ending ranks.

We used a divergence-of-divergence metric (RTD^2) to identify n-grams that contributed to disagreement between base-divergence results derived from n-gram distributions. More specifically, we ranked the RTD values calculated from the ranks of the RTD contributions to divergence results for ground truth and predicted distributions (using our classifiers). Said another way, in cases where n-grams had high RTD^2 values, those n-grams would either be over- or under-emphasized in the data re-

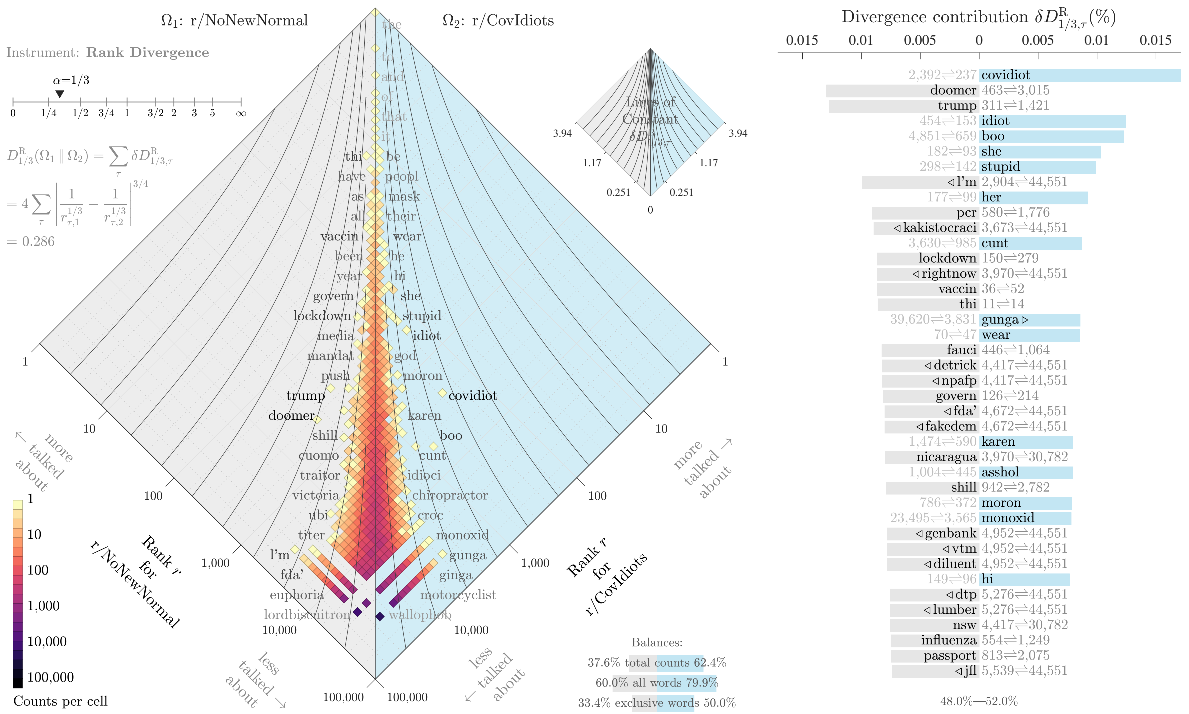

Figure 1: **An allotaxonograph (Dodds et al., 2020) showing the 1-gram rank distributions of `r/NoNewNormal` and `r/CovIdiots` along with rank-turbulence divergence results.** The central diamond shaped plot shows a rank-rank histogram for 1-grams appearing in each subreddit. The horizontal bar chart on the right shows the individual contribution of each 1-gram to the overall rank-turbulence divergence value ($D^R_{1/3}$). The 3 bars under "Balances" represent the total volume of 1-gram occurring in each subreddit, the percentage of all unique words we saw in each subreddit, and the percentage of words that we saw in a subreddit that were unique to that subreddit.

sulting from our classification pipeline when compared with the ground truth.

3.5 In-group and out-group prediction

We inferred membership of individual users in in-group or onlooker subreddits using two binary classification models. These models were applied to the entire concatenated comment history of users for a given subreddit. In addition to the data filtering described in Section 3.3, we removed users whose concatenated comment histories contained fewer than 10 1-grams. In order to investigate the effect of comment length on classification performance, we created a second training and evaluation data set—referred to as the "threshold" data set—with users whose comment histories contained at least 100 1-grams and who made at least 10 comments on their assigned subreddit. Due to the large class imbalance in most subreddit pairings, we under-sampled the majority class to rebalance the training and testing data sets.

To establish a baseline, we trained a logistic regression model on term frequency-inverse document frequency (TF-IDF) features. For the logistic regression model, we generated TF-IDF features by selecting 1-grams that appeared in at least 10 documents and at most 95% of total documents. We also removed English stopwords before feeding these features to a logistic regression model.

We compared the performance of the logistic regression model with a Longformer-based classifier (Beltagy et al., 2020). The Longformer model uses a sparse attention mechanism to address the quadratic memory scaling of the standard transformers (Vaswani et al., 2017)—in our cases allowing for the consideration of longer documents (comment histories). For the Longformer model, we used the default Transformers library (Wolf et al., 2020) implementation of a sequence classifier with a maximum sequence length of 2,048.

4 Results

4.1 Language classifier

For all subreddit pairs, we found that both language classifiers performed better than random, with some variation along subreddit size and community characteristics, as in Figs. 4 and 5. The Longformer model performed better in all cases (as indicated by the Matthews correlation coefficient (MCC) in Table 1). However, with sufficient data volume, the logistic regression classifier was able to achieve comparable results, especially notable given the reduced model complexity.

For the Longformer model trained and evaluated on r/NoNewNormal and r/CovIdiots, we achieved precision and recall values of approximately 0.75 for both classes Table 5. For the other subreddits, precision and recall values ranged between approximately 0.65 and 0.9 with near parity between the classes. See Fig. 2 for receiver operator characteristic (ROC) curves for the Longformer model.

The logistic regression classifier offered lower performance but relatively similar results with the added benefit of interpretable feature importance scores. In the case of r/NoNewNormal and r/CovIdiots, we report feature importance for the logistic regression model in Table 3. The feature importance results provide some insights on how bag-of-words models are capturing community-specific language. For instance, "media", "doomer", and "trump" are language features highly predictive of the r/NoNewNormal subreddit accounts. On the other hand, "idiots", "crocs", and "5g" are language features highly predictive of the r/CovIdiots accounts.

4.2 Divergence results

4.2.1 Initial observations

We found that RTD identified salient terms when comparing the 1-gram distributions of r/NoNewNormal and r/CovIdiots. As seen in Fig. 1, we found that terms relating to specific people and institutions such as "trump", "fda", and "fauci" drove RTD contributions from the r/NoNewNormal distribution. For the same subreddit, we found 1-grams related to vaccines—"vaccine[s]", "dtp" (Diphtheria-Tetanus-Pertussis), and "npafp" (Non-polio Acute Flaccid Paralysis)—which ranked higher than the opposing subreddit. Finally, some 1-grams related to non-pharmaceutical interventions ranked relatively higher in the r/NoNewNormal distribution, including "lockdown" and "passport". From the r/CovIdiots 1-gram distribution, we saw the eponymous term "covidiot" contributing the greatest to RTD followed by insults such as "stupid" and "karen"—illustrating the insulting critiques that many of the r/CovIdiots posts level at r/NoNewNormal.

The RTD results suggest a few characteristics of each subreddit. Both r/NoNewNormal and r/CovIdiots discussed prominent topics related to the pandemic—as seen by terms such as "mask", "vaccine", and "lockdown" ranking in the top 300 1-grams for each subreddit. The subreddits' focuses constrast each other with r/NoNewNormal appearing more focused on discussion that is critical of pandemic interventions and r/CovIdiots criticizing r/NoNewNormal (as evidenced by a higher degree of insulting language).

4.2.2 Effect of classifier on divergence results

Overall RTD values were similar for both the ground truth and predicted distributions ($D^{\mathrm{R}}_{1/3} = 0.286$ and 0.274, respectively). In Table 2 we present the top 20 1-grams as highlighted by RTD^2 . We saw fluctuations for terms related to internet memes (e.g., "gunga", "ginga", and "boo"). In other cases, function words like "he" and "be" are ranked as contributing notably to the RTD^2 results—this may be owing to nuanced differences in speech patterns between the two communities that are amplified by the classification and RTD^2 results. For some highly topical 1-grams, such "trump", "covidiot", and "influenza", we found shifts in rank limited to an order of magnitude—in these cases the salient 1-grams contributed more to RTD in the classifier-derived data set, likely owing to the bias of the model.

4.3 Accuracy versus user attributes

We expected our classifier to perform better on active users who received praise from a community (as indicated by the voting score on their comments). To confirm this hypothesis, we plotted the likelihood of correctly labeling users that post in r/NoNewNormal compared to their number of comments in the subreddit, total comment-score, and mean comment-score, shown in Fig. 3.

Our classifier performed most reliably on users with ten to three hundred comments in the subreddit, and ten to five hundred total karma. Performance decayed for users with over 400 comments, but there were only 520 users in this category out of about 58,000 r/NoNewNormal users. Anecdotally, this small subset of users engaged in longer

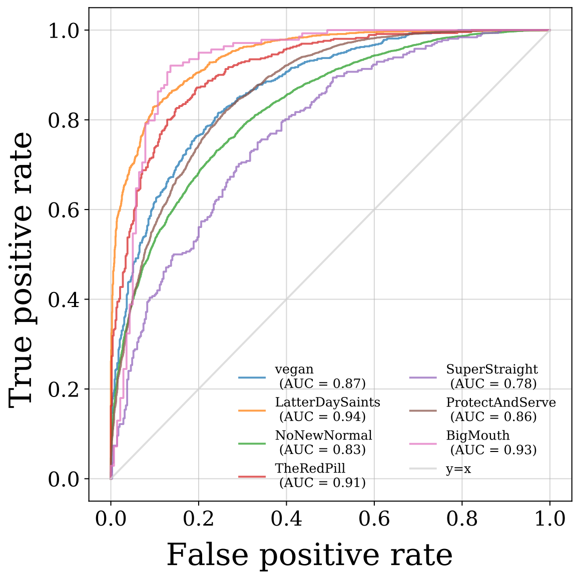

Figure 2: **Receiver operator characteristic curves for classification models evaluated on the subreddit pairs.** For each subreddit pair we trained a binary classifier based on the Longformer language model. The classifier trained on `r/BigMouth` and `r/BanBigMouth` showed the best performance (AUC = 0.93) while our primary case study—`r/NoNewNormal` and `r/CovIdiots`—had an AUC value of 0.83. It is worth noting the variation in sample sizes and as described in Table 1.

and more general discussions, and as a result, used language that is more common and more difficult to classify compared to their less active peers.

To filter out low-activity users, we re-ran our classifier after pruning accounts with less than under 100 one-grams in their comment history or less than 10 total comment in their associated subreddit. This filtering is discussed in Section 3.5 and labeled "Threshold" in Table 1 where we present the classification results. The threshold data generally improved the performance of both the logistic regression and Longformer models.

5 Discussion

The work outlined here is motivated by the challenge of accurately classifying communities that discuss the same topics but are distinct in their exact views. Further, we are motivated by the task of identifying these communities in the absence of interaction data that may allow for the construction of a social graph.

Our methodology addresses the challenge of analyzing online conversation around contentious topics where there may be polarized communities that share similar linguistic features. For instance, when studying online discourse around a specific topic one approach to collecting relevant content is anchor wording (selecting posts based on the presence of key words defined by a researcher). In the case of `r/NoNewNormal` and `r/CovIdiots`, "vaccine", "mask", and "covid" share similar rank values in the 1-gram distributions for each subreddit (55, 37; 24, 28; 51, 58; respectively). A naive anchor-word selection would capture much of the conversation in each of these communities. However, anchor word selection would fail to disambiguate the dramatically differing views held by the majority of users in each community. This has impacts on down stream analysis such as sentiment analysis, tracking narrative diffusion, and topic modelling.

Considering our main motivation was a problem description and initial demonstration of a classification pipeline, we did not extensively explore model architectures or hyperparameters. We included n-gram order in the initial hyperparameter sweep when developing the logistic-regression pipeline, and results suggested that 1-grams were most effective. However, including higher order n-grams is still worth exploring more in-depth, and may have benefits for model interpretabilility and down stream results (e.g., feature importance). Further, we selected the word-embedding model (the Longformer) based mainly on considerations related to maximum sequence length and preliminary performance observations. Additional word-embedding models could be considered—choosing models trained on more recent and/or domain specific data may be especially helpful.

As in stance detection (Alkhalifa and Zubiaga, 2021), there are several limitations to the methodology we present. First, our data set covers a limited time frame, and past work has demonstrated that models which are trained on old data sets may perform relatively poorly when fed new data (Alkhalifa et al., 2021; Alkhalifa and Zubiaga, 2021). Additionally, our methodology does not account for the fact that users may change opinions throughout time. For example, a user may initially be a member of a group, but a shift in opinion may cause the user to leave the group but still engage in discussion about said group. Lastly, our classifier is only trained on English posts, and we cannot guarantee the same level of performance across languages.

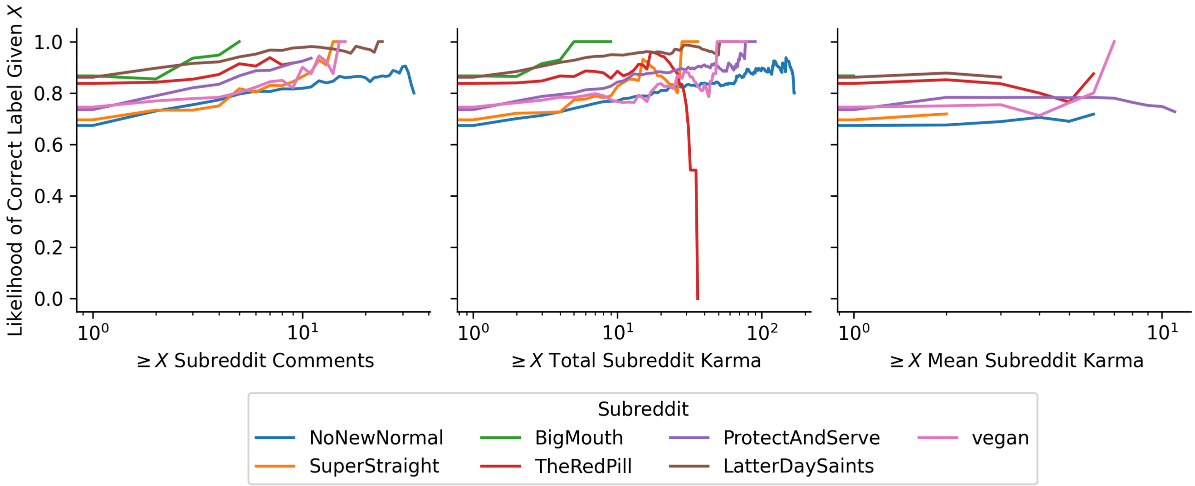

Figure 3: **Likelihood of correctly labeling users in in-group subreddits by user attributes.** From left to right, correct labeling versus user comments in the subreddit, correct labeling versus total karma in the subreddit, and correct labeling versus mean karma in the subreddit. In all cases, the classifier performed poorly with low-activity users, better with moderate activity. We have pruned the 10% of users with the highest attributes from this plot, to improve legibility. An unabridged version of the plot is in the appendix, with a more detailed explanation. Plots include only users that commented in the primary "of" subreddit. Results from base-LR classifier.

Table 1: **Data set size and classification performance for logistic regression (LR) and Longformer (LF) models.** Subreddit pairs, primary "of" community first, "onlooking" subreddit second. Matthews correlation coefficient (MCC) refers to performance on the test set. The threshold results refer models trained on a thresholded data set where user comment histories must contain at least 100 1-grams and at least 10 comments. Results excluded due to small sample size are represented with an "*".

Subreddits	**MCC**				**Data set size**	
	Base		**Threshold**		**Base**	**Threshold**
	LR	**LF**	**LR**	**LF**		
r/NoNewNormal v. r/Covidiots	0.41	0.48	0.57	0.60	44185	6778
r/TheRedPill v. r/TheBluePill	0.55	0.65	*	*	4680	402
r/BigMouth v. r/BanBigMouth	0.64	0.80	*	*	1394	140
r/SuperStraight v. r/SuperStraightPhobic	0.35	0.43	*	*	3310	584
r/ProtectAndServe v. r/BadCopNoDonut	0.50	0.55	0.65	0.76	41158	6930
r/LatterDaySaints v. r/ExMormon	0.65	0.72	0.80	0.83	15062	4122
r/vegan v. r/antivegan	0.49	0.56	0.65	0.72	6896	1692

6 Conclusion

In the present study, we frame the research challenge of classifying in-groups and onlookers based on the linguistic features of social media posts. The classification task is made difficult by the significant intersection of terms shared between the two communities, which may confound classification attempts. We collect a data set of seven (7) subreddit pairs that match the in-group and onlooker-group criteria, focusing our efforts on a case study of pro- and anti-COVID mitigation communities. These subreddits provide an appealing proving ground for group identification tasks, because subreddit participation acts as a noisy label in lieu of ground truth for group identity. We identify salient 1-grams that differentiate each communities' language distributions. Using the full collection of subreddit pairs, we train two classifiers to assign users to communities based on their posts. We demonstrate the feasibility of the classi-

1-gram	RTD^2 Rank	RTD rank (pred.)	RTD rank (actual)
he	1	11.0	446.0
be	2	4285.0	19.0
vaccin	3	7.0	104.0
thi	4	143.0	8.0
nyt	5	15.0	459.0
they	6	27.0	3414.5
diffrent	7	42.5	17076.0
ginga	8	73.5	9.0
gunga	9	24.0	5.0
shill	10	103.0	13.0
titer	11	11026.0	59.5
boo	12	2.0	1.0
covidiot	12	1.0	2.0
sham	14	52.0	4253.0
voluntari	15	53.0	4420.5
influenza	16	14.0	103.0
purg	17	1694.5	44.0
postul	18	16.0	123.0
trump	19	8.0	3.0
dui	20	51.0	1956.0

Table 2: **Rank-turbulence divergence (RTD) of divergence results from actual and predicted 1-gram distributions.** As a divergence-of-divergences measurement, RTD^2, shows disagreement between the divergence results derived from 1-gram distributions of generated with ground truth labels and the distribution generated with our classification pipeline. Highly ranked RTD^2 values highlight the 1-grams that have the greatest difference in rank of contribution to the divergence results for each pairing. For instance, "trump" is the 1-gram with the 3rd highest contribution in ground-truth data, whereas the 1-gram is ranked 8th in the classifier-generated data. We stemmed the 1-grams prior to calculation of divergence results.

fication scheme with these results. In most cases, our classifier recovers 70% or more of a community's users. From these results, we show how our initial language distribution divergence results may be affected by using data labelled by our classifier. In the case of the COVID subreddits, the true and classifier-generated distributions are qualitatively similar, identifying notable 1-grams in each case. We hope the research questions and combined set of results is motivating for future work that leverages training generalizable classifiers on labelled community data that can then be used in a variety of settings.

7 Future Work

We present a first attempt at in-group classification based on contextual language use, in a challenging environment where both the in-group and onlookers discuss many of the same topics. We believe that classifiers in this domain have important applications for cross-platform group detection, where more reliable labels like consistent usernames and network interactions are unavailable. More powerful classifiers may account for additional text features, including user sentiment, shared topics, stance towards those topics, and language style. Longer time-span studies should be wary of semantic drift over time (Schlechtweg et al., 2019), as well as more specific changes in group language and stance on topics. Models of community language style (Tran and Ostendorf, 2016) could also help identify communities across platforms, as long as platform-specific language style features are identified and controlled for.

References

Gavin Abercrombie and Riza Theresa Batista-Navarro. 2018. Identifying opinion-topics and polarity of parliamentary debate motions. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*. Association for Computational Linguistics.

Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. 2021. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374.

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2021. Opinions are made to be changed: Temporally adaptive stance classification. In *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*, pages 27–32.

Rabab Alkhalifa and Arkaitz Zubiaga. 2021. Capturing stance dynamics in social media: Open challenges and research directions. *arXiv preprint arXiv:2109.00475*.

Abhinav Anand and Jalaj Pathak. 2022. The role of Reddit in the GameStop short squeeze. *Economics Letters*, 211:110249.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 1–9.

R. Armitage. 2021. Online 'anti-vax' campaigns and COVID-19: censorship is not the solution. *Public Health*, 190:e29–e30.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Sumit Bhatia and P Deepak. 2018. Topic-specific sentiment analysis can help identify political ideology. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 79–84.

Alexandre Bovet and Hernán A. Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10(1):7.

Hongxu Chen, Hongzhi Yin, Xiangguo Sun, Tong Chen, Bogdan Gabrys, and Katarzyna Musial. 2020. Multi-level graph convolutional networks for cross-platform anchor link prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 1503–1511. ACM.

Peter Sheridan Dodds, Joshua R Minot, Michael V Arnold, Thayer Alshaabi, Jane Lydia Adams, David Rushing Dewhurst, Tyler J Gray, Morgan R Frank, Andrew J Reagan, and Christopher M Danforth. 2020. Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems. *arXiv preprint arXiv:2002.09770*.

Camille Grange. 2018. The generativity of social media: Opportunities, challenges, and guidelines for conducting experimental research. *International Journal of Human–Computer Interaction*, 34(10):943–959.

Margeret Hall, Athanasios Mazarakis, Martin Chorley, and Simon Caton. 2018. Editorial of the special issue on following user pathways: Key contributions and future directions in cross-platform social media research. *International Journal of Human–Computer Interaction*, 34(10):895–912.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. A survey on stance detection for mis-and disinformation identification. *arXiv preprint arXiv:2103.00242*.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, page 90–94, New York, NY, USA. Association for Computing Machinery.

Sameera Horawalavithana, Abhishek Bhattacharjee, Renhao Liu, Nazim Choudhury, Lawrence O. Hall, and Adriana Iamnitchi. 2019. Mentions of security vulnerabilities on reddit, twitter and github. In *IEEE/WIC/ACM International Conference on Web Intelligence*, page 200–207. ACM.

Kenneth Joseph, Sarah Shugars, Ryan Gallagher, Jon Green, Alexi Quintana Mathé, Zijian An, and David Lazer. 2021. (Mis) alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 312–324.

Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. 2016. Political issue extraction model: A novel hierarchical topic model that uses tweets by political and non-political authors. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 82–90.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander G Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 109–116.

Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. 2018. Argument mining on clinical trials. In *COMMA*, pages 137–148.

Bjarke Mønsted and Sune Lehmann. 2022. Characterizing polarization in online vaccine discourse—a large-scale study. *PloS one*, 17(2):e0263746.

Nathaniel Persily. 2017. The 2016 US Election: Can democracy survive the internet? *Journal of democracy*, 28(2):63–76.

Lorenzo Prandi and Giuseppe Primiero. 2020. Effects of misinformation diffusion during a pandemic. *Applied Network Science*, 5(1):82.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746.

Parinaz Sobhani. 2017. *Stance detection and analysis in social media*. Ph.D. thesis, Universite d'Ottawa/University of Ottawa.

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*, pages 1030–1035, Austin, Texas. Association for Computational Linguistics.

Milo Trujillo, Sam Rosenblatt, Guillermo de Anda Jáuregui, Emily Moog, Briane Paul V Samson, Laurent Hébert-Dufresne, and Allison M Roth. 2021. When the echo chamber shatters: Examining the use of community-specific language post-subreddit ban. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 164–178.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. 2016. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE transactions on knowledge and data engineering*, 28(8):2158–2172.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62, Vancouver, BC, Canada. Association for Computational Linguistics.

Moran Yarchi, Christian Baden, and Neta Kligler-Vilenchik. 2021. Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 38(1–2):98–139.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.

Appendix

Subreddit Corpus Sizes

Table 4 indicates the size of each subreddit, in terms of user count and comment count, after pruning bots and low-karma users as specified in our methodology. It also includes the mean karma (comment score) for remaining comments in each subreddit corpus.

Comparison of Subreddit Activity

If subreddits in a pair have dramatically different activity levels, such as much longer comments in one subreddit than another, these differences in writing style may correlate with classification difficulty. Figs. 4 and 5 show cumulative distributions of comment length and comment count per user, respectively, to illustrate which subreddits are closer in behavior than others.

Uniquely Identifying Words

Table 3 shows the words that most strongly correlate with membership in r/NoNewNormal and r/CovIdiots.

Labeled Language versus Predicted Language

Fig. 1 shows word use divergence between r/NoNewNormaland r/CovIdiotsusing all comments from users in each subreddit. For comparison, Fig. 7 shows the same word use divergence based only on users our classifier predicted as members of each subreddit.

Classifier performance metrics

Table 5 shows F1 scores and precision values for the logistic regression and longformer model.

Classifier Accuracy versus User Attributes

Our classifier performs best on accounts with above 10 comments and a minimum comment-karma threshold. However, the classifier cannot reliably label every user in the tail of the distribution. This leads to a misleading visualization, conflating the low-density of users that have high comment counts or karma scores with classifier performance. Therefore, we did not include the tail of each performance graph in Fig. 3. For posterity, we have included an unabridged version of the graph that includes these misleading tails, in Fig. 6.

r/NowNewNormal	r/CovIdiots
media	covidiots
emails	covidiot
questioning	retard
lockdown	cunt
jab	nnn
power	report
restrictions	idiot
narrative	deniers
woke	idiots
yall	idiocy
guys	crocs
passport	ugh
msm	5g
subreddit	selection
dystopian	wedding
sheep	frustrating
doomer	fox
doomers	hoax
sub	beard
trump	department

Table 3: **Feature importance for logistic regression classifier trained on r/NowNewNormal and r/CovIdiots**. The two columns correspond to the text features that are most strongly predictive of each subreddit.

Subreddit	**Users**	**Comments**	**Mean Karma**
r/NoNewNormal	57966	1245398	4.743
r/CovIdiots	28427	174056	4.119
r/TheRedPill	10149	59388	3.608
r/TheBluePill	2744	9616	4.716
r/BigMouth	6252	19904	1.895
r/BanBigMouth	981	3226	1.359
r/SuperStraight	5914	46491	2.686
r/SuperStraightPhobic	1897	11498	1.449
r/ProtectAndServe	25096	241328	7.484
r/Bad_Cop_No_Donut	77288	314933	5.898
r/LatterDaySaints	9130	131055	2.498
r/ExMormon	35672	852607	3.440
r/vegan	62544	622069	4.908
r/antivegan	4492	47738	3.878

Table 4: Users and comments in each subreddit, after filtering out bots and low-karma users

Subreddits	**F1**				**Precision**				**Data set size**	
	Base		**Threshold**		**Base**		**Threshold**		**Base**	**Threshold**
	LR	**LF**	**LR**	**LF**	**LR**	**LF**	**LR**	**LF**		
r/NoNewNormal v. r/Covidiots	0.71	0.74	0.83	0.80	0.71	0.74	0.83	0.80	44185	6778
r/TheRedPill v. r/TheBluePill	0.79	0.84	*	*	0.84		*	*	4680	402
r/BigMouth v. r/BanBigMouth	0.80	0.88	*	*	0.80	0.88	*	*	1394	140
r/SuperStraight v. r/SuperStraightPhobic	0.67	0.69	*	*	0.67	0.69	*	*	3310	584
r/ProtectAndServe v. r/BadCopNoDonut	0.75	0.78	0.90	0.88	0.75	0.78	0.90	0.88	41158	6930
r/LatterDaySaints v. r/ExMormon	0.83	0.86	0.95	0.91	0.83	0.86	0.95	0.91	15062	4122
r/vegan v. r/antivegan	0.75	0.78	0.88	0.86	0.75	0.78	0.88	0.86	6896	1692

Table 5: **Data set size and classification performance for logistic regression (LR) and Longformer (LF) models.** Subreddit pairs, primary "of" community first, "onlooking" subreddit second. F1 scores and precision values are calculated using weighted average for the balanced data sets. F1, precision, and recall (not shown) values were all approximately equal for specific models and subreddit pairs in our experiments—partially owing to the balanced datasets. The threshold results refer models trained on a thresholded data set where user comment histories must contain at least 100 1-grams and at least 10 comments. Results excluded due to small sample size are represented with an "*".

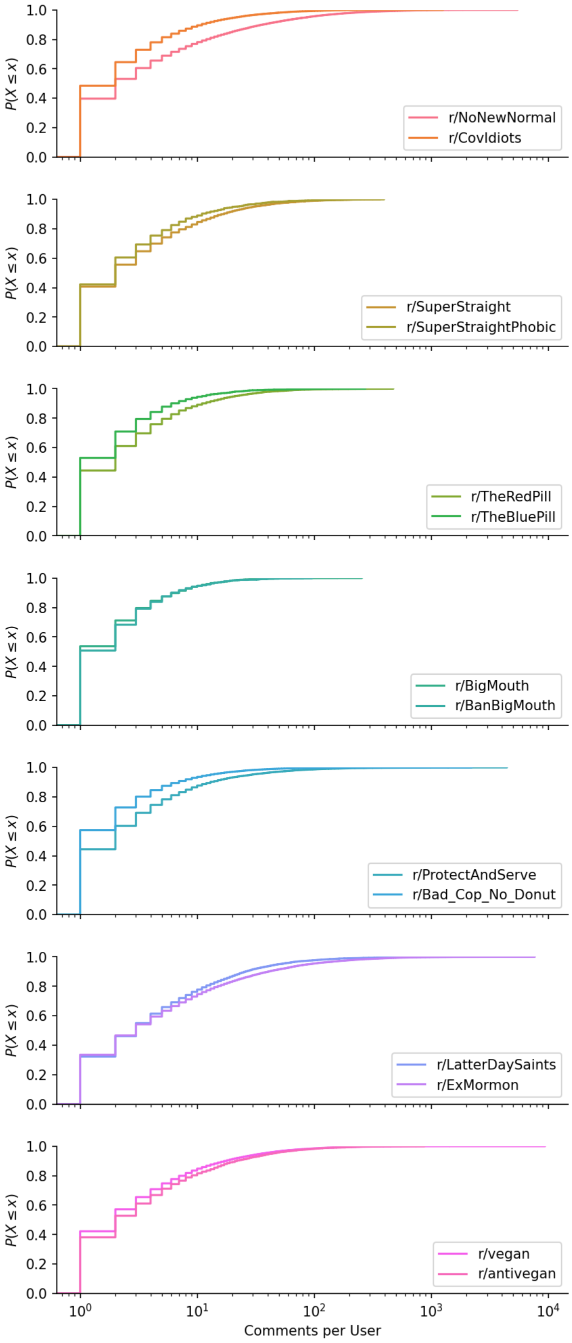

Figure 4: **Cumulative distribution of comments made by each user in each examined subreddit pair.** Distribution taken after filtering.

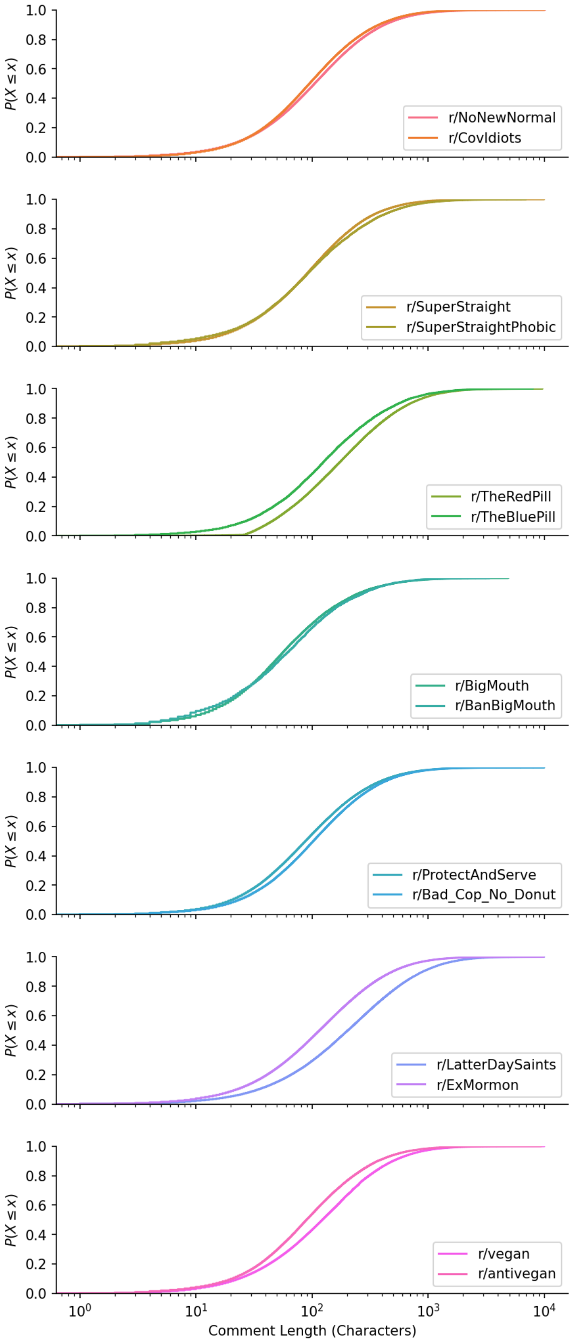

Figure 5: **Cumulative distribution of comment length in each examined subreddit pair.** Distribution taken after filtering.

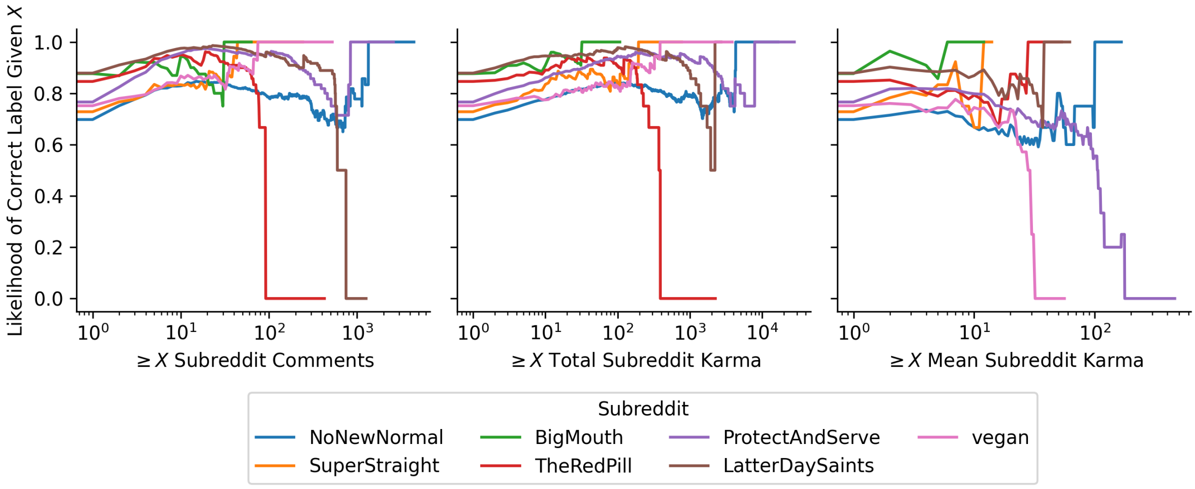

Figure 6: **Likelihood of correctly labeling users in in-group subreddits by user attributes.** This is the unabridged version of Fig. 3, including unstable long-tail behavior when classifying the small minority of high-activity accounts.

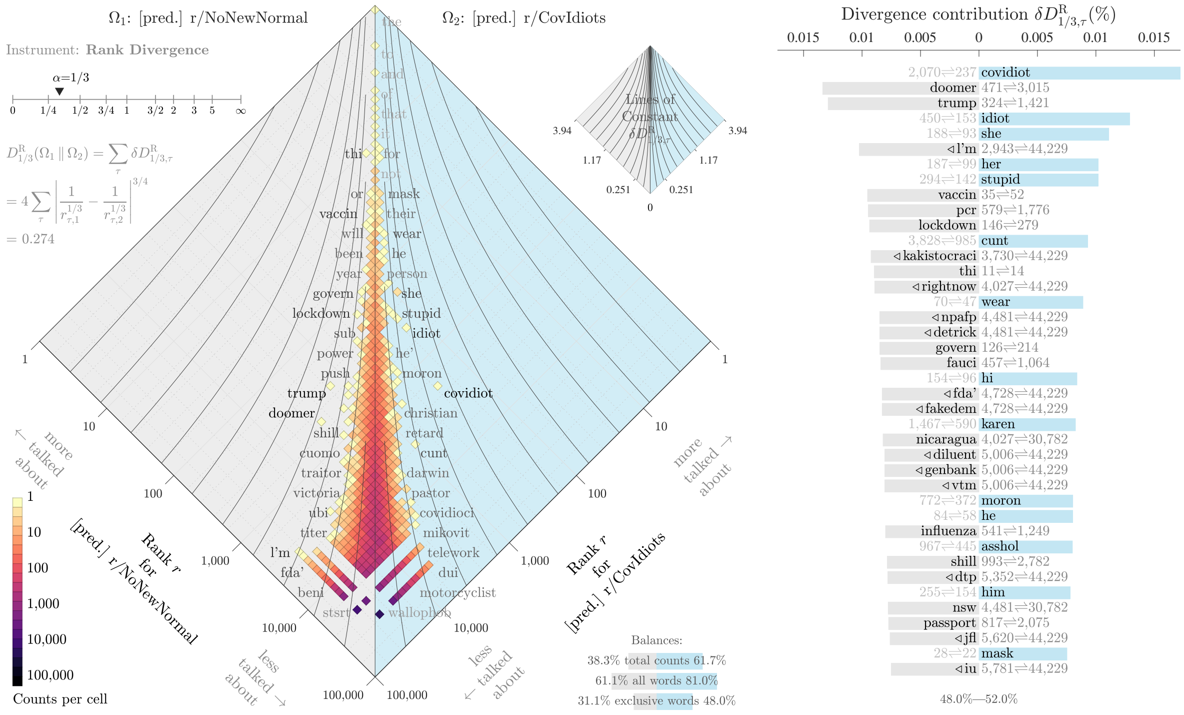

Figure 7: **An allotaxonograph (Dodds et al., 2020) showing the 1-gram rank distributions of predicted users of `r/NoNewNormal` and `r/CovIdiots` using our classifier to assign membership.** See Fig. 1 for allotaxonograph of actual users. The central diamond shaped plot shows a rank-rank histogram for 1-grams appearing in each subreddit. The horizontal bar chart on the right show the individual contribution of each 1-gram to the overall rank-turbulence divergence value ($D^R_{1/3}$). The 3 bars under "Balances" represent the total volume of 1-gram occurring in each subreddit, the percentage of all unique words we see in each subreddit, and the percentage of words that we see in a subreddit that are unique to that subreddit.