# Friend or Foe? Detecting Adversaries in Hateful Subreddits

Bui, Vy Chen, Chiu-Chun Helms, Derek Hickey, Dan
Oregon State University
{buivy, chenchiu, helmsd, hickeyda}@oregonstate.edu

## Abstract

*Recent work that seeks to make inferences about how individuals become radicalized in online environments samples users from hateful subreddits. However, these studies make the assumption that all users in these subreddits are aligned with the community's greater interests. In reality, many users enter hateful subreddits and employ "counter-speech," which is language that refutes the hateful ideologies of the subreddits' core members. We take inspiration from the counter-speech detection and domain adaptation literature to build a model to detect such users in hateful subreddits. We then use this model to demonstrate how the quality of research on hateful subreddits can be improved by filtering adversaries of the subreddits out of samples.*

## 1. Introduction

Online hate communities present a significant threat to the world, both online and offline. Numerous mass killings have been linked to perpetrators' engagement in hateful online communities, and hate can make social media platforms less welcoming, especially for minority identity groups. Social media platforms, such as Reddit, have a unique role to play in this issue, as they make it easier than ever for hateful individuals to form communities and communicate with each other [3]. It is therefore essential to document the actions and dynamics of online hate groups to understand further how to moderate them and mitigate radicalization. Previous work in this domain has revealed important characteristics of hateful subreddits, such as the effect becoming active in a hateful subreddit has on users' language outside of the subreddits [28], or that members of hateful subreddits frequently attack newcomers, reducing their engagement in the hateful communities [13]. However, work of this nature that attempts to make inferences about hateful online users by sampling from hateful subreddits operates under the assumption that each member of each subreddit is genuinely hateful or prone to radicalization. While it is likely this is true for most users that post in hateful subreddits, this can nonetheless have an impact on studies like this. In some studies that sample from hateful subreddits, the quality of the samples can be improved by filtering out users who are not genuinely hateful. In other cases, knowing which users are not hateful can greatly affect the conclusions of a study. For example, in the study done by Hickey et al [1] [13], users of hateful subreddits were found to be hostile toward newcomers, and this hostility was associated with newcomers leaving the subreddits. The authors of this paper came to the conclusion that members of hateful subreddits were ineffective at recruiting new members through direct social interaction. However, if members of hateful subreddits are only hostile toward users who enter the community to criticize it, and those users are the ones who are being driven away, then that provides evidence that members of hateful subreddits are effective at driving away users they do not want to engage with their community. Furthermore, there is evidence subreddits benefit from being hostile toward outsiders who attack their community [18].

In this report, we address and investigate this assumption by detecting "adversaries" in hateful subreddits – users who express views that oppose ideologies of hate or the general views of core members of hateful subreddits. We pair datasets and methods from the counter-speech detection literature with the domain adaptation technique of self-training [26] to detect such users. We test our model on a manually annotated set of 500 comments from hateful subreddits, then estimate the proportion of adversaries in hateful subreddits, as well as perform an analysis from Hickey et al. [13] with samples adjusted for adversaries to demonstrate the importance in adjusting for them in analyses. Overall, we find that while results can be replicated from a prior counter-speech detection study, performance significantly drops for this new task and domain. Some strategies to adapt the task to our new domain result in dramatic improvements above the initial drop from the direct replication, though they do not get as high as the performance of the original study. Using our models, we demonstrate that adversaries are indeed prevalent in these subreddits, but not to the degree that they change the results of the analysis

[1] This paper is not published yet, so the citation is fake. Just email Dan if you are curious about the details.

done by Hickey et al [13]. However, the performance of our method for detecting adversaries as well as the quality of our annotations are not high, so we suggest the results of this study be taken lightly and a more rigorous follow-up with an improved training set and annotation process, or more extensive domain adaptation experiments, be conducted.

## 2. Related Work

### 2.1. Online Hate Groups and Community Conflict

Many previous works have sought to understand online hate and radicalization, which can inform moderation efforts of social media platforms. Causal modeling is commonly used in such studies as controlled experiments are difficult in social media settings [8]. Reddit users dramatically increase their usage of hateful language outside of hate groups when they become active in hate groups [28]. Similarly, some studies have explored the impact of moderation of hateful communities, such as quarantines [4] or bans [14] on the members of those communities, finding that hateful community membership is reduced with moderation efforts, but the users who stay may become more extreme. Other research measures the negative impact of hate speech on users' mental health [27]. While the harms of hateful communities and language on social media are well-documented, there is still much work to be done on understanding why users become radicalized in the first place. Hickey et al. [13] attempt to help answer this question by documenting what factors of interaction in hateful communities cause users to keep posting in them.

While most subreddits are dedicated to discussions surrounding specific topics, including hate, it is not true that every user who posts in a given subreddit is a proponent of the topic of the subreddit. "Brigading" is a well-documented phenomenon, where an online community will attack another community in an organized manner, by either directing abusive speech toward the targeted community or exploiting popularity metrics (e.g. upvotes/downvotes) to reduce the visibility of views important to the targeted community [11, 16]. One study that explored inter-community conflict on Reddit observed a "colonizing" effect of attacks on subreddits, where users from a community that attack a certain subreddit become more active in the attacked subreddit following the attack, while core users of the attacked community become less engaged. The authors observed that hostile responses toward such attacks often mitigated the colonizing effect of the attack [18]. Furthermore, another study of inter-subreddit conflict found subreddits similar to the ones studied in this report to be frequently involved in such conflicts [6].

### 2.2. Counter-speech Detection

Counter-speech is an approach to combating hate speech alternative to content moderation, wherein users of social media platforms reply to hateful language with arguments that refute or denounce such behavior [22]. Counter-speech can take many forms, such as presenting facts to object to hate speech, responding to hate speech with humor, or even responding to hate speech with hostility. Mechanisms to promote counter-speech on social media are commonly recommended to platforms, as moderation efforts receive pushback from individuals who are concerned about excessive censorship [21]. Counter-speech detection classifiers have been built using data from several domains, including YouTube [22], Twitter [1, 12], and Reddit [29, 30]. Some of these studies have also demonstrated how incorporating context into counter-speech datasets can improve performance [1,30]. He et al. [12] use their counter-speech classifier to find exposure to counter-speech decreases users' likelihood of becoming hateful on Twitter. To our knowledge, while counter-speech classifiers have been built using Reddit data, no extensive analyses of counter-speech in Reddit communities have been performed.

### 2.3. Domain Adaptation in Natural Language Processing

Most machine learning algorithms operate under the assumption that the test set is drawn from the same distribution as the training set [25]. Often, a model's performance will drop when applied to a different distribution, but in-distribution performance is also strongly correlated with out-of-distribution performance [23]. Common methods of domain adaptation involve changing aspects of the source model, such as the loss function or model structure. Other methods involve applying semi-supervised learning to unlabeled data [25]. Semi-supervised learning techniques include self-training, in which a model is trained on labeled data and used to predict classes on unlabeled data [26]. The predictions above a certain confidence threshold are automatically labeled with their predicted class and removed from the set of unlabeled data. The high-confidence predictions are then used along with the original labeled data to train the model again. This process is repeated until there are no more high-confidence predictions. Another type of semi-supervised learning called tri-training uses three classifiers trained on bootstrapped labeled data. Unlabeled data points are added to the training set when all three models agree on the prediction. The process is repeated until no more data points can be added to the training set [26]. We opt to employ self-training in this paper to improve performance under domain shift.

Label shift is another domain adaptation that can be considered for our problem. Black Box Shift Correction (BBSC) is a method to detect and quantify shifts between

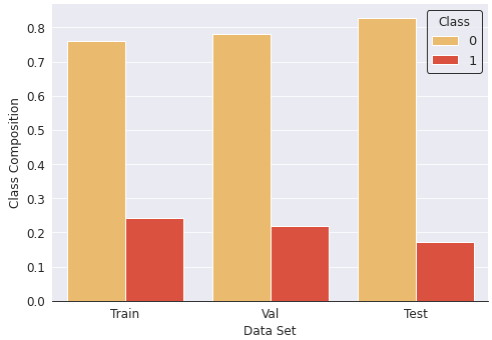


Figure 1. Class composition per data split, with train and validation sets containing silver + gold from Yu et al. and test set containing observations from newly collected Subreddit data.

training and test set distributions, and apply corrections to the posterior probabilities in order to reflect original model performance onto a new class label distribution [19]. With fine-tuning being performed on the hate and counter hate speech data set originally proposed by Yu et al. [30] and evaluation being performed on our new data collected from 25 different hateful subreddits, class labels have seen a shift as our domain has become hate-content specific. However, as the class distribution of our test set is similar to the class distribution of the data we trained our model on, we opted not to use this method.

# 3. Methodology

## 3.1. Replication of Yu et al.

As a starting point for developing our own counter-speech detection model, we replicate a model built by Yu et al. [30]. The authors used a dataset of 6,846 observations collected from Reddit by retrieving comments containing one or more hate words, which were then marked as the target variable, and their parent comments were marked as the context variable for that target [30]. This creates the context/target pairs, which correlate to a "post and reply" general format to represent a conversation between two users. Annotation of these context/target pairs was completed through anonymizing and crowdsourcing the text to 674 different annotators, and was then split into two separate subsets: "gold" and "silver" data sets. The gold set consists of 4,751 pairs that had high agreeance among annotators (Krippendorf's $\alpha \geq 0.6$) while the silver set contains the 2,095 remaining pairs that had a lower agreement (Krippendorf's $\alpha < 0.6$). The Multi-Annotator Competence Estimation (MACE) score [15] was used to determine posts with high agreement [30], as annotations from individuals with low MACE scores were dropped to increase the agreement scores. The test set was derived only from examples in the gold dataset, while the silver dataset was used for training and validation in their highest-performing methods.

As a benchmark for replication, we refer to the highest F1 score reported by Yu et al. [30] for counter-speech detection, which was 0.53. To achieve this score, the authors implemented pre-training a RoBERTa transformer [20] on a stance detection dataset [24] and fine-tuned the model with a combination of both gold and silver data sets, utilizing the combination of both context and target variables. We report results following this exact method, though we test other models that stray from this method, described in more detail in Section 3.2.

While the authors built a multi-class classifier consisting of classes of "hate," "neutral," and "counter-speech," our problem focuses mainly on the ability to replicate these methods only for the counter-speech class. This is because our domain is hateful subreddits rather than subreddits in general, and we want to detect users who explicitly state opposition to hateful ideologies. Furthermore, we make the assumption that there is a strong chance that users who use neutral language in hateful subreddits are hateful. Therefore, for all evaluations, we frame this as a binary classification problem.

## 3.2. Evaluating Counter-hate Detection Models on Hateful Subreddit Data

To Expand on the work originally proposed by Yu et al. [30], we used a dataset of users' first posts in a set of 25 different hateful subreddits, spanning multiple targeted identity groups, collected by Hickey et al [13]. As these data are unlabeled, we manually labeled a small subset of 500 observations, sampled equally from each hateful subreddit. The posts were labeled to mirror the three classes from the Yu et al. data: hate (0), neutral (1), and counter-hate speech (2). Labelling was implemented by having each team member classify 50 of these observations in order to determine overall class agreement, and a single team member who is an expert in hateful subreddits moved forward with labeling the remaining observations. The observed Fleiss Kappa score among all annotators was 0.55, indicating moderate agreement. The distribution of labels between the original data and newly collected Subreddit data are nearly identical, with Figure 1 representing the per class composition for each split. Utilizing this data to expand on the previous work, we evaluated multiple counter-hate detection models on the hateful subreddit data and reported values for both the F1 and ROC-AUC scores, with scores being evaluated based on the performance of the counter-hate speech class. In combination with this, the models were also evaluated on the original test data created by Yu et al. [30] in order to compare model performance with the shift from general to hateful subreddit data for both metrics.

We evaluated multiple hypotheses of what could be done

| Method | Yu et al. F1 | Yu et al. ROC-AUC | New Domain F1 | New Domain ROC-AUC |
|---|---|---|---|---|
| Original Yu et al. Results [30] | **0.53** | N/A | N/A | N/A |
| Random Baseline | 0.32 | 0.5 | 0.26 | 0.5 |
| Yu et al. Replication | 0.49 | 0.65 | 0.32 | 0.65 |
| 0% <NULL> Context Injection | 0.40 | 0.67 | 0.39 | 0.73 |
| 25% <NULL> Context Injection | 0.41 | 0.68 | **0.46** | **0.75** |
| 50% <NULL> Context Injection | 0.43 | 0.69 | 0.39 | 0.73 |
| Target Only | 0.44 | **0.73** | 0.43 | 0.75 |
| Target Only + Yu et al. Test Data | N/A | N/A | 0.39 | 0.74 |
| Self-Training Method (Target Only) | 0.43 | 0.70 | 0.29 | 0.73 |

Table 1. Experimental results comparing Yu et al. test data vs new data performance. The <NULL> context injection models and target only models were trained without the "silver" dataset from Yu et al.

to maximize performance on our annotated test set. Firstly, we slightly alter the configuration of the model from what Yu et al. report achieved the best performance [30]. As we noticed performance on the validation set was highest when only considering the "gold" dataset, we did not use the "silver" dataset for any of our further experiments. Then, we randomly replace the context training examples with <NULL> tokens select thresholds to measure the difference in performance among models with missing context. The reasoning for this is that the PushShift API is missing some data [9], which means context examples will be unobtainable in certain cases. Indeed, 75% of the examples in our annotated test set are missing context, with the average proportion per subreddit varying greatly (min. 16%, max. 71%, we double-checked the discrepancy between these values and the proportion in the test set and conclude we must have gotten very unlucky in sampling the test set). The second method involved training on target variables only, with no marker for missing context. This method was also re-implemented with the addition of the gold test set from Yu et al. [30] during training, and evaluated only on the new subreddit data to determine the change in performance from increased training data sizes. Finally, self-training [26] was employed using a set of 25,000 unlabeled Reddit posts, with 1,000 coming from each Subreddit. The confidence threshold for adding unlabeled predictions to the training set was 0.8.

In addition to the results for each model, we also report random baseline performance, wherein predicted labels were generated with an equal probability and the performance F1 score was calculated from those labels. The random label generation process was repeated 1,000 times and the mean F1 score was used as the final random baseline. Other random baselines were generated using a lower probability of labeling an example as counter-hate, but they were not as high as the equal probability case.

### 3.3. Analysis of Adversaries in Hateful Subreddits

To demonstrate the utility of our models, we apply our self-training model to the full unlabeled dataset collected by Hickey et al [13]. To provide greater context for this analysis, we will first summarize the previous work: First, all posts from 25 different hateful subreddits were collected using the Pushshift API [2]. The first post of each user was taken from each subreddit, as well as the first reply from each of those posts if they received replies. The authors also recorded whether or not each user continued to post within other threads on the subreddit they initially posted in. Users who received replies to their first posts were matched with users who did not receive replies using causal inference methods, and the groups were compared. The authors found that users who received replies to their first posts in hateful subreddits were less likely to keep posting in them than users who did not receive replies. Furthermore, the authors matched hateful subreddits to non-hateful subreddits and found replies have the opposite effect in non-hateful subreddits. To investigate this discrepancy, the authors used the "attack on commenter" model from the Perspective API [10] and found that personal attacks are much more common in hateful subreddits compared to non-hateful ones, and attacks are strongly negatively associated with reduced engagement across all subreddits.

There are two relevant problems that can be addressed using these data – firstly, it is useful to know the overall prevalence of adversaries in hateful subreddits and which subreddits they are more common in. To do this, we measure the proportion of positive predictions from each users' first post in each hateful subreddit. It should be noted this treats all users equally and does not adjust for activity levels, so adversaries may be measured as more prevalent than they would be perceived by a user on the Reddit platform.

The next question that can be addressed is whether removing adversaries from the samples of users will change the results from Hickey et al. in any way. For the purpose

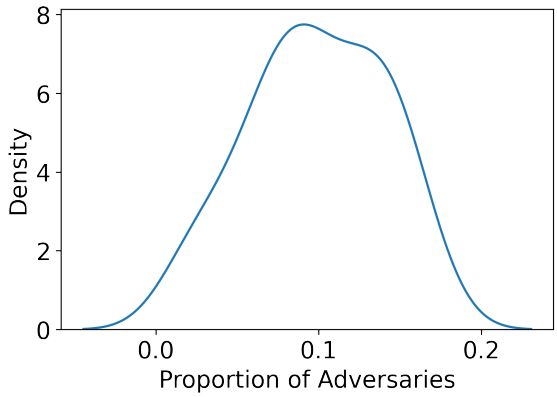


Figure 2. Distribution of proportions of adversaries in each hateful subreddit.

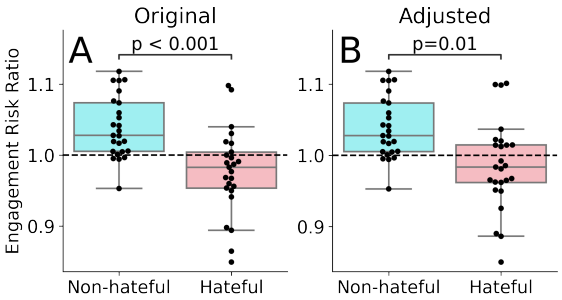


Figure 3. Results comparing the effects of replies on engagement in hateful subreddits using (A) the original sample from Hickey et al. and (B) the sample after removing matched pairs of users containing adversaries from hateful subreddits. An Engagement Risk Ratio greater than one indicates replies have a positive effect on engagement, while a ratio lower than one indicates replies have a negative effect.

of this report, we focus on the comparison between hateful and non-hateful subreddits in terms of the effect replies have on further engagement in the subreddits. As the samples consist of users who received replies matched directly with users who did not receive replies, any matched pair that contains an adversary is removed from the sample. In an ideal scenario, the matching process would be repeated after removing all adversaries from the samples, but for the sake of time, we simply remove matched pairs.

## 4. Results

**Results from Yu et al. could be replicated.** Table 1 displays the results of all models we experimented with. We achieve an F1 score of 0.49 when replicating the method Yu et al. obtained their best results with. While this is not as high as the benchmark of 0.53, the result is still expected to happen by chance, as Yu et al. report a mean F1 of 0.5 and a standard deviation of 0.03 when training their model using a series of different random seeds [30]. This indicates competence in our ability to train models, as we achieve performance comparable to the standard for published research using this training dataset. Additionally, it reinforces the validity of the research conducted by Yu et al.

**Domain adaptation techniques show mixed results.** When testing different levels of missing context in the training dataset, we found having context removed from some training examples helped improve performance on our labeled test set (Table 1). The best performance was acheived by randomly removing context from 25% of the training examples (F1=0.46). However, training without any context was very close in performance (F1=0.43). Surprisingly, self-training results in the lowest performance on the labeled test set (F1=0.29), aside from the random baseline.

**Adversaries are prevalent in hateful subreddits.** The distribution of proportions of adversaries in hateful subreddits is shown in Figure 2. The mean is 0.1 with a standard deviation of 0.04.

**Results from Hickey et al. stay consistent when accounting for adversaries.** Figure 3 displays the results of the analysis of overall effects of replies in hateful subreddits before and after adjusting for adversaries. We find that the results are still statistically significant in favor of the original result from Hickey et al. [13]. However, the overall negative effect of replies is smaller after adjusting for adversaries.

## 5. Conclusion

In this study, we trained multiple counter-speech classification models to detect users in hateful subreddits who do not align with the general views of the community. We also demonstrated how the results of prior research can change when considering such users.

We attempted several strategies for improving performance under domain shift with mixed results overall. Some strategies, such as incorporating more training data, either in the form of the test set from the original training distribution, or weakly labeled data from the self-training method, performed worse than we expected. Furthermore, incorporating training data with low inter-annotator agreement also resulted in lower performance. While it is counter-intuitive that less data results in better predictions, it is possible that our model was overfitting to the distribution in the original domain when given more data from that distribution, so it generalized better to the new domain with less data. For the self-training method, it is possible that biases from the original distribution affected the labeling of data in the new distribution, and errors early on in the self-training process then propagated throughout the rest of the training set. However, more extensive validation of the parameters of the self-training process, such as training dataset size and con-

fidence threshold, should be done before drawing any concrete conclusions.

Lastly, while our classifiers are not extremely accurate, we nonetheless demonstrate how it is important for practitioners to consider how their samples of users from different hateful subreddits can include users that are not of interest to their research question. While we do not demonstrate that this changes the results of another research paper, it is still evident to us that this is a category of users that should be considered. We urge researchers studying hate groups or hate speech online to consider this possibility. Furthermore, we feel more confident in the conclusions made by Hickey et al. as we could not disprove them through this analysis.

## Limitations and Future Directions

**Annotation Quality.** Our whole test set was only annotated by one person, and for the portion that was annotated by multiple people, agreement among raters was moderate. Furthermore, capturing the nuances of the language used in specific hateful online communities requires extensive background knowledge and training. For these reasons, we caution that the final results of this report be taken lightly. An in-depth, expert-guided annotation process with disagreements discussed by annotators, as implemented by Vidgen et al. [29] would provide more clarity regarding the true performance of our model.

**Quality of training data.** While we chose to focus on data from Yu et al. [30] to train our counter-hate model due to its similarity to our domain, we acknowledge the observed upper limit for overall performance on this dataset is low (F1=0.53). Notably, other researchers have achieved much higher F1 scores of 0.72 [22] and 0.85 [12] on data from YouTube and Twitter, respectively. It is possible that the data used to train these classifiers are of higher quality than the dataset we used, and the benefit of the increase in quality could overpower the detriment of a greater domain shift. Furthermore, both prior studies use data annotated by a small set of experts rather than annotations crowdsourced from Amazon Mechanical Turk (as was done by Yu et al [30]). While we acknowledge the importance of reducing bias in annotations by sourcing them from a wide variety of people, there are also issues with annotations sourced from Mechanical Turk [7]. Namely, we suspect the task of annotating counter-speech to be a difficult one that crowd workers have not been sufficiently trained for.

**More rigorous testing of domain adaptation strategies.** In this study, we evaluated the strategy of self-training to improve the performance of counter-speech classification under domain shift. Surprisingly, performance decreased when using this method. However, due to time constraints, this was the only method we tested, and there are many other methods that could be applied [25]. Even within the self-training method, there are hyperparameters we could have tuned given more time, such as the confidence threshold or max number of iterations. Furthermore, strategies such as tri-training have been demonstrated to outperform self-training [26]. Performance would likely be improved by considering other domain adaptation methods.

**Impact of adversaries on hateful subreddits.** In this study, we observed how hateful individuals respond to their adversaries in hateful subreddits and how that can influence the staying power of the adversaries. However, we did not investigate the influence adversaries may have on the success of the hateful subreddits or the behavior of their members. This would be an important problem to investigate, as previous studies warn of the potential dangers of hateful individuals migrating to platforms where their views are challenged less frequently [5, 14, 17].

## References


[1] Abdullah Albanyan and Eduardo Blanco. Pinpointing fine-grained relationships between hateful tweets and replies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10418–10426, 2022. 2

[2] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset, 2020. 4

[3] Manuela Caiani and Patricia Kröll. The transnationalization of the extreme right and the use of the internet. *Int. J. Comp. Appl.*, 39(4):331–351, 2015. 1

[4] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Quarantined! examining the effects of a community-wide moderation intervention on reddit. *ACM Trans. Comput.-Hum. Interact.*, 29(4), mar 2022. 2

[5] Simon Copland. Reddit quarantined: Can changing platform affordances reduce hateful material online? *Internet Policy Review*, 9(4):1–26, 2020. 6

[6] Srayan Datta and Eytan Adar. Extracting inter-community conflicts in reddit. In *Proceedings of the international AAAI conference on Web and Social Media*, volume 13, pages 146–157, 2019. 2

[7] Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, pages 413–420, 2011. 6

[8] Antigoni-Maria Founta and Lucia Specia. A survey of online hate speech through the causal lens. *arXiv:2109.08120*, 2021. 2

[9] Devin Gaffney and J Nathan Matias. Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *PloS one*, 13(7):e0200162, 2018. 4

[10] Google Jigsaw. Perspective api, 2017. 4

[11] Timothy Graham and Aleesha Rodriguez. The sociomateriality of rating and ranking devices on social media: A case study of reddit's voting practices. *Social Media+ Society*, 7(3):20563051211047667, 2021. 2

[12] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 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*, pages 90–94, 2021. 2, 6

[13] Daniel Hickey, Matheus Schmitz, Paul Smaldino, Daniel Fressler, Goran Muric, and Keith Burghardt. No love among haters: Negative interactions reduce online hate community engagement. *arXiv*, 2023. 1, 2, 3, 4, 5

[14] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *CSCW*, 5(CSCW2):1–24, 2021. 2, 6

[15] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia, June 2013. Association for Computational Linguistics. 3

[16] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *TOCHI*, 25(2):1–33, 2018. 2

[17] Neil F Johnson, Rhys Leahy, N Johnson Restrepo, Nicholas Velásquez, Minzhang Zheng, Pedro Manrique, Prajwal Devkota, and Stefan Wuchty. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773):261–265, 2019. 6

[18] Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *TheWebConf*, pages 933–943, 2018. 1, 2

[19] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3122–3130. PMLR, 10–15 Jul 2018. 3

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[21] Binny Mathew, Anurag Illendula, Punyajoy Saha, Soumya Sarkar, Pawan Goyal, and Animesh Mukherjee. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24, 2020. 2

[22] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380, 2019. 2, 6

[23] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021. 2

[24] John Pougué-Biyong, Valentina Semenova, Alexandre Matton, Rachel Han, Aerin Kim, Renaud Lambiotte, and Doyne Farmer. Debagreement: A comment-reply dataset for (dis) agreement detection in online debates. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 3

[25] Alan Ramponi and Barbara Plank. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, 2020. 2, 6

[26] Sebastian Ruder and Barbara Plank. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1044–1054, 2018. 1, 2, 4, 6

[27] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. Prevalence and psychological effects of hateful speech in online college communities. In *WebSci*, pages 255–264, 2019. 2

[28] Matheus Schmitz, Keith Burghardt, and Goran Muric. Quantifying how hateful communities radicalize online users, 2022. 1, 2

[29] Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online, June 2021. Association for Computational Linguistics. 2, 6

[30] Xinchen Yu, Eduardo Blanco, and Lingzi Hong. Hate speech and counter speech detection: Conversational context does matter. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5918–5930, Seattle, United States, July 2022. Association for Computational Linguistics. 2, 3, 4, 5, 6