Enter At Your Own Risk: The Impacts of Joining a Hateful Subreddit

Kaitlyn Ko
University of Chicago
Chicago, IL 60637
Email: katiehko@uchicago.edu

Keith Burghardt
University of Southern California
Information Sciences Institute
Marina Del Rey, CA 90292
Email: keithab@isi.edu

Goran Muric
University of Southern California
Information Sciences Institute
Marina Del Rey, CA 90292
Email: gmuric@isi.edu

***Abstract*—The Internet has become one of the most valuable assets for extremists to disseminate information to a broader audience. We are interested in whether racist online communities are simply rendezvous of individuals who already subscribe to hateful ideology, or if the existence of these online communities has the capacity to radicalize more neutral individuals. Our preliminary findings show that joining a racist subreddit increases hate word usage for several months in hateful and non-hateful subreddits. This has broad implications on the role of the Internet in not only supporting, but growing racist behavior.**

I. Introduction

The greatest resource to the modern world has also become its greatest threat: the Internet. As it works to connect a previously unfathomable number of people, it ultimately polarizes communities, instead of unifying them [1]. Through this, racism has found it especially easy to flourish. Online echo chambers of racism not only to make these ideas more accessible to a wider audience [2], but have been implicated to have a direct impact in coordinating hate crimes as well [3].

The RECRO model has proposed that previously neutral individuals can be radicalized by entering extremist online communities [4]. This model was empirically substantiated with conspiracy communities on Facebook and Reddit [5]. However, the overall empirical impact of joining *hateful* alt-right subreddits on individual users has yet to be studied extensively.

A large proportion of racist online activity happens in more backstreet social media communities, especially those with fewer hate speech regulations such as 4chan, Parler, and Reddit. Of particular interest is Reddit, which is helpful for studying these hateful communities due to its partitioning of each post by "subreddits" [6].

This leads us to the question, How does engaging with hateful online communities affect the individual user? More specifically, does joining a hateful subreddit make an average Reddit user use more hate speech upon joining? We found that, among users who have posted in r/GreatApes, joining this subreddit increased their usage of hate words for several months in outside subreddits. We are currently working to determine if this trend does not occur for users who did join r/GreatApes.

II. Methodology

For this study, we focused on r/GreatApes and r/uncensorednews as "hateful subreddits". While several subreddits may qualify as hateful, r/GreatApes is one of few subreddits that was expressly created to be anti-black. r/uncensorednews is also marked with racism and xenophobia as it focuses primarily on the "bad things perpetrated by members of minority groups" [7]. Larger subreddits which are more notoriously known to be hateful such as r/The_Donald were not studied due to the short time scope of the REU in which this study was conducted.

In order to determine the correlation between joining a hateful subreddit and hate speech, we calculated the percentage of hate words per total words, averaged across all users before and after users joined the subeddit. Hate speech was defined as instances of hate words from a lexicon created for another anti-black subreddit, r/CoonTown[1] [8]. We will aggregate this data over both aforementioned subreddits.

Additionally, we will divide the users into *treatment* and *control* groups using Mahalanobis Distance Matching. Users will be matched on the basis of several variables, such as the time they joined Reddit, the types of subreddits they are interested in, frequency of posts and comments, and popularity. Then, we will compare "hate speech" between matched users via causal modeling, to bolster the observed relationship between joining a hateful subreddit and hate speech (Figure 1).

III. Discussion

Preliminary data from r/GreatApes were analyzed with correlation testing. As shown in Figure 3, users who joined r/GreatApes, on average, showed a spike of hate word usage outside of the subreddit for approximately 10

[1] Complete manually-filtered lexicon can be found at https://tinyurl.com/hatewords (contains offensive content).

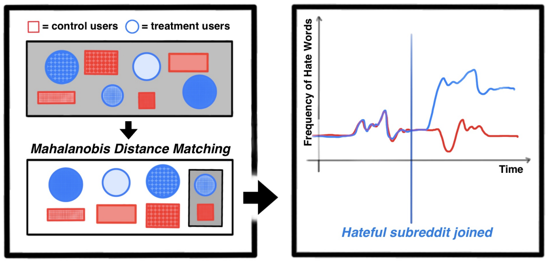
□ = control users ○ = treatment users
Mahalanobis Distance Matching
Frequency of Hate Words
Time
Hateful subreddit joined

Fig. 1. A schematic of the methodology. In the left panel, red squares represent users from the control group while blue circles represent users from the treatment group (i.e., users who joined r/GreatApes). Shapes and their sizes and patterns represent different user feature values. Matched users will be compared in the right panel, where we compare their usage of hate words upon joining the hateful subreddit.

months upon joining. The average hate words per total words increased from 0.06% to 0.07% from before and after joining ($\Delta = 0.01\%$).

We will repeat this analysis with r/uncensorednews, then perform causal modeling. To further extrapolate upon the findings, we may expand the definition of "hate speech" to include more qualitative means of hate speech in addition to hate words [9]. Moreover, we will try to identify heterogeneous effects on our results by investigating user characteristics that may have contributed to the observed trends. Finally, we will explore other types of subreddits, such as anti-vaccine, to determine if the observed trend is only true with racist hate words or other types of vernacular as well.

While we will have to wait to see our full results, we believe our preliminary results show promise that joining a hateful subreddit like r/GreatApes is associated with an increase in hate speech usage outside of the subreddit, at least in the short-term. This may show that the Internet is instrumental in growing racist ideology, as well as elucidate measures to stifle "echo chambers" of hate speech [10] and consequent hate crimes.

Acknowledgments

The author would like to thank the NSF REU program at the Information Sciences Institute (Grant No. 2051101) for sponsoring this research.

References

[1] A. Rao, F. Morstatter, M. Hu, E. Chen, K. Burghardt, E. Ferrara, and K. Lerman, "Political partisanship and anti-science attitudes in online discussions about covid-19," *arXiv preprint arXiv:2011.08498*, 2020.

[2] M. Caiani and P. Kröll, "The transnationalization of the extreme right and the use of the internet," *International Journal of Comparative and Applied Criminal Justice*, vol. 39, no. 4, pp. 331–351, 2015.

[3] M. L. Williams, P. Burnap, A. Javed, H. Liu, and S. Ozalp, "Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime," *The British Journal of Criminology*, vol. 60, no. 1, pp. 93–117, 2020.

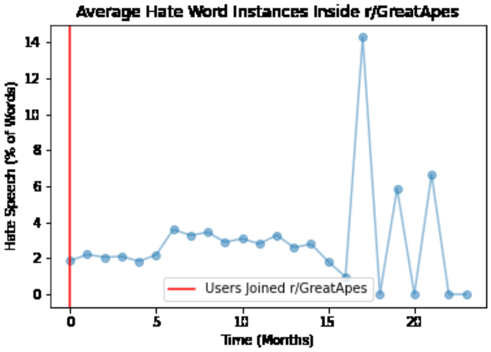

Fig. 2. Hate words as a percentage of total words in posts inside r/GreatApes vs. months after joining r/GreatApes. Note that the proportion of hate words per total words was considerably greater in posts inside r/GreatApes than in those outside (Figure 3). The large variance in data after 16 months was likely due to the fact that a smaller sample continued to post in r/GreatApes for extended periods after joining.

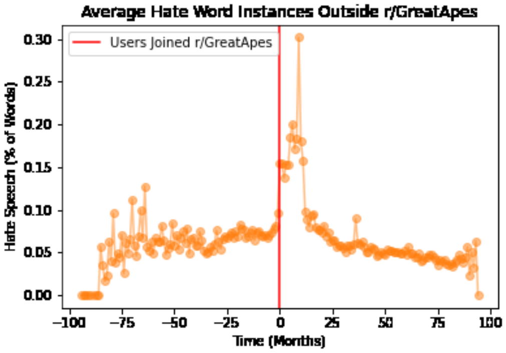

Fig. 3. Hate speech over time in posts outside r/GreatApes. Hate speech percentage spiked immediately after users joined r/GreatApes but plateaued back to previous levels approx. 10 months after joining.

[4] L. S. Neo, "An internet-mediated pathway for online radicalisation: Recro," in *Violent Extremism: Breakthroughs in Research and Practice*, pp. 62–89, IGI Global, 2019.

[5] N. Van Raemdonck, "The echo chamber of anti-vaccination conspiracies: mechanisms of radicalization on facebook and reddit," *Institute for Policy, Advocacy and Governance (IPAG) Knowledge Series, Forthcoming*, 2019.

[6] M. Z. Trujillo, S. F. Rosenblatt, G. d. A. Jáuregui, E. Moog, B. P. V. Samson, L. Hébert-Dufresne, and A. M. Roth, "When the echo chamber shatters: Examining the use of community-specific language post-subreddit ban," *arXiv preprint arXiv:2106.16207*, 2021.

[7] T. Squirrell, "Linguistic data analysis of 3 billion reddit comments shows the alt-right is getting stronger," Aug 2017.

[8] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–22, 2017.

[9] T. Colley and M. Moore, "The challenges of studying 4chan and the alt-right:'come on in the water's fine'," *New Media & Society*, p. 1461444820948803, 2020.

[10] T. Gaudette, R. Scrivens, G. Davies, and R. Frank, "Upvoting extremism: Collective identity formation and the extreme right on reddit," *New Media & Society*, p. 1461444820958123, 2020.