Understanding the Effect of Deplatforming on Social Networks

Shiza Ali
Boston University, USA
shiza@bu.edu

Mohammad Hammas Saeed
Boston University, USA
hammas@bu.edu

Esraa Aldreabi
Binghamton University, USA
ealdrea1@binghamton.edu

Jeremy Blackburn
Binghamton University, USA
jblackbu@binghamton.edu

Emiliano De Cristofaro
University College London, UK
e.decristofaro@ucl.ac.uk

Savvas Zannettou
Max Planck Institute for Informatics, Germany
szannett@mpi-inf.mpg.de

Gianluca Stringhini
Boston University, USA
gian@bu.edu

ABSTRACT

Aiming to enhance the safety of their users, social media platforms enforce terms of service by performing active moderation, including removing content or suspending users. Nevertheless, we do not have a clear understanding of how effective it is, ultimately, to suspend users who engage in toxic behavior, as that might actually draw users to alternative platforms where moderation is laxer. Moreover, this *deplatforming* efforts might end up nudging abusive users towards more extreme ideologies and potential radicalization risks. In this paper, we set to understand what happens when users get suspended on a social platform and move to an alternative one. We focus on accounts active on Gab that were suspended from Twitter and Reddit. We develop a method to identify accounts belonging to the same person on these platforms, and observe whether there was a measurable difference in the activity and toxicity of these accounts after suspension. We find that users who get banned on Twitter/Reddit exhibit an increased level of activity and toxicity on Gab, although the audience they potentially reach decreases. Overall, we argue that moderation efforts should go beyond ensuring the safety of users on a single platform, taking into account the potential adverse effects of banning users on major platforms.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**; • **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

Online social networks, deplatforming, moderation

WebSci '21, June 21–25, 2021, Virtual Event, United Kingdom

ACM ISBN 978-1-4503-8330-1/21/06…$15.00
https://doi.org/10.1145/3447535.3462637

ACM Reference Format:
Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the Effect of Deplatforming on Social Networks. In *13th ACM Web Science Conference 2021 (WebSci '21), June 21–25, 2021, Virtual Event, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447535.3462637

1 INTRODUCTION

Over the past years, toxic activity on social media like hate speech, cyberbullying, and harassment, has become an increasingly important problem [1, 36]. To curb abuse, social media platforms have adopted different mitigation strategies, including providing users with tools to flag abusive behavior [29]. They also suspend or ban accounts which are deemed unfit for the community. Suspended accounts are usually guilty of violating the platform's terms of service, which forbid illegal behavior (e.g., sending spam) as well as partaking in antisocial behavior. Depending on the policy of each social network, offending accounts are blocked for some time or banned permanently. Suspension actions targeted at individuals who took part in hateful and harassing conduct are commonly referred to as *deplatforming* [49].

Most research in this space has looked at deplatforming in a siloed fashion, evaluating the effect that these actions have on the platforms the accounts where banned from [11, 23, 57]. However, users are obviously not bound to a single platform, but can migrate to other online services where moderation is possibly more lax. In fact, anecdotal evidence shows that once hateful users get banned from Twitter, they often move to Gab, an alternative social network with an open lack of moderation marketed as protection of "free speech" [2, 43, 50]. The effect of these migrations is not well understood. In particular, it is not clear if users tend to become more toxic after they move to a more extreme community. What is clear is that these communities have been used as outlets for violent actors, e.g., Robert Bowers' anti-Semitic posting on Gab just prior to murdering eleven and wounding six people at the Tree of Life synagogue in Pittsburgh [46].

In this paper, we aim to measure the effect that deplatforming events have on users, looking at how their activity changes when they get suspended on a social network and move to an alternative one. To do so, we focus on users who were suspended on Twitter and

Reddit and moved to Gab, a social network with laxer moderation. Overall, we aim to answer the following research questions:

- **RQ1 – *Account Creation:*** Do suspended users create an account on an alternative platform after being suspended on a mainstream one, or did they already have one?
- **RQ2 – *Toxicity:*** Do suspended users become more toxic after migrating? Platforms might be banning users to enforce codes of conduct, perhaps in the hopes that users will reform (especially with respect to temporary suspensions). However, it is possible that users will actually become more toxic when migrating to other platforms, especially when the new platforms have an emphasis on more lax moderation.
- **RQ3 – *Activity:*** Do suspended users become more active after they migrate to other platforms? We want to understand whether moving to a less moderated platform can contribute to a user's activity, resulting in them posting more content.
- **RQ4 – *Audience:*** Do suspended users get bigger or smaller audiences after their migration to other platforms? We want to understand whether after moving to an alternative social network users are able to retain their following. If a user's audience is reduced, they may reach fewer people and the effects of their toxic activity will be contained. Intuitively, we expect that the audience reachable by a user will shrink after they are banned from a major social network and move to a smaller one.

To answer these questions, we first need to identify users on Gab who were suspended from Twitter and Reddit. To do so, we start from a large corpus of 29M posts gathered from Gab, and cross-reference profile names with those that used to be active on Twitter and Reddit but were suspended. Overall, there are several reasons why a user would reuse the same profile name on a different platform [32], e.g., for continuity and recognizability with their followers [30]. However, it is also possible that, especially for "common" profile names, there are accounts on multiple social networks that are not controlled by the same person. To identify these accounts, we manually labeled a sample of the data and developed a classifier which achieves an accuracy of 94.5%. We consider 3,074 out of 4,790 (64%) suspended Twitter accounts with a corresponding profile name on Gab and 5,216 out of 6,308 (82.7%) on Reddit as controlled by the same person.

We find that 73.68% of the accounts on Gab were created *after* being suspended from Twitter or Reddit, highlighting a non-negligible user migration as a result of deplatforming. We also find that users tend to become more active and their posts more toxic after moving to Gab. However, since they lose followers, their audience tends to decrease.

Overall, our study paints a comprehensive picture of the efficacy of deplatforming on the Web community that suspended the users, the one users migrated to, as well as the effects on user behavior. We argue that deplatforming seems to assist in safeguarding users from the platform that took the moderation action, while substantially reducing the audience of problematic users. Nevertheless, suspended users are becoming more active and more toxic, which is possibly an indication of online radicalization that might have an overall negative impact both on the online and offline world (i.e., users perpetrating real-world violence due to online radicalization [24, 28]). These broad implications should be taken into careful consideration by policymakers, social network operators, and by the research community studying these problems.

2 RELATED WORK

In this section, we review relevant related work.

Hate Speech on Social Media. Silva et al. [52] study the main targets of hate speech in online social media by analyzing content shared on Twitter and Whisper, while Mondal et al. [36] analyze common hate expressions, the effect of anonymity on hate speech, and the most hated groups across regions. [33] study the spread of hate speech in Gab, showing that hateful content has greater outreach and spreads faster. As hate speech on social media become increasingly popular, previous work has also worked on automated detection using machine learning [6, 9, 14, 18, 48, 61].

Malicious Accounts. A wealth of research has studied malicious accounts on social networks, from those involved in sending spam [22, 54, 57] to performing fraud [15, 26, 63], to taking part in online harassment [12, 19] Alorainy et al. [5] analyze suspended accounts and argue that suspended accounts are a reliable source for hate speech prediction. They analyze three sources of data sets: suspended, active, and neutral ones. Their emotional analysis indicate that tweets from suspended accounts show more disgust, negative, fear, and sadness emotions than the ones from active accounts. In another study, Volkova et al. [59] predict suspicious, i.e., deleted or suspended accounts in social media. They analyze multiple datasets of thousands of active, deleted, and suspended Twitter accounts and produce predictive behaviors that lead to the removal or shutdown of an account. They observe that the presence of certain terms in tweets increases the likelihood for that account to be deleted or suspended.

Moderation on Social Media. Jhaver et al. [27] analyze users' reactions to Reddit's moderation process. They find that 18% of the participants believed that their posts were removed for the right reasons, 37% did not know why their post was removed, and 29% expressed resentment towards the removal of their posts. Furthermore, Habib et al. [23] undertake a comprehensive study on whether it is feasible to proactively moderate Reddit communities. They build a machine learning model to study the characteristics of subreddits and predict the future behavior of that subreddit. They also analyze the impact of different events on user behavior and find that banning and quarantining subreddits does not have an impact on the overall civility of the users involved in that community. Unlike our work, their focus is on banning entire subcommunities (i.e., subreddits) and not single users. They suggest that there is a need for more active and nuanced intervention strategies to effectively moderate malicious accounts. However, [11] study the 2015 ban of two hate communities on Reddit, /r/fatpeoplehate, and /r/CoonTown; they analyze the effects of the ban on users and the communities and conclude that quarantining these communities was successful for Reddit since users left the site or reduced their hate speech.

Newell et al. [38] explore how a period of community unrest on Reddit affected users migration off platform. Similarly to our study, they use an algorithm (with a lower bound of 0.6 precision) which

makes use of user names to match users across a variety of other platforms (e.g., HackerNews and Voat). While their findings are relevant to the current work, as they expose some of the motivations behind users migrating off of Reddit, there is a fundamental difference: our study focuses on users that were *forcibly* removed from Twitter or Reddit.

Ribeiro et al. [47] analyze data from two communities r/The_Donald and r/Incels that were banned from Reddit and migrated to their own websites. Compared to this research, our focus is on the migration of single users after suspension and not on the reaction of users to entire communities being banned.

Remarks. To the best of our knowledge, this paper is the first to study how the activity of online users changes once they get suspended on one platform and migrate to a different one.

3 METHODOLOGY

Our main goal is to identify and study accounts that were suspended on one or more social platforms – namely, Twitter and Reddit – and later moved to alternative platforms such as Gab. We choose Gab because, according to previous work [65], it has laxer moderation policies and it attracts accounts suspended from major social networks.

To identify pairs of accounts that were controlled by the same person (the suspended account on Twitter or Reddit and the one on Gab) we proceed backwards. We first identify Gab users from a large dataset of publicly available Gab posts. Next, we identify accounts on Twitter and Reddit that used the same profile name as the ones on Gab, and that have since been suspended. To avoid false positives, we build a classifier to determine whether or not two accounts on different platforms were likely controlled by the same people. In this section, we describe these steps in detail.

3.1 Gab Data Collection

We use the Gab dataset made available by PushShift [7]; this contains 29 million posts made over a 1.5 year period (between 2016 and 2018) by 322,397 unique Gab users. We then extract the profile names of these accounts and check whether accounts with the same profile names were also present on Twitter and Reddit but have been suspended.

3.2 Identifying Pairs of Accounts with the Same Profile Name

We start by assuming that users who get suspended on Twitter and Reddit and move to Gab will create an account with the same profile name. Next, for each account in the Gab dataset, we look up their profile name on Twitter and Reddit, and check whether an account with that profile name existed and was suspended. We began collecting our data for Twitter and Reddit in January 2020.

Twitter. For Twitter, we use the Twython Twitter API [41] user searching functionality [35] and check whether the user is active, suspended (these get a HTTP 403 error), or not found (HTTP 404 error), indicating that the account either never existed or was deleted by its owner. From the 322K Gab profile names, we find 200,303 users on Twitter with the same profile name. Of these, 20,967 Twitter accounts were suspended at the time of collection

Reddit. To check whether accounts with a certain profile name used to exist on Reddit, we look up *reddit.com/u/<profilename>.json*. If the object that is returned contains the message *Forbidden*, with HTTP error code 403, we consider that user to be suspended. We find 145,835 users on Reddit with the same profile name on Gab and out of these 6,308 were suspended.

3.3 Collecting Data of Suspended Users on Twitter and Reddit

Next, we collect data about suspended Twitter/Reddit users with a matching Gab profile name.

Twitter. Twitter accounts can be suspended or have their activity limited for security purposes, or they may have violated the Twitter Rules, or have some features limited due to suspicious activity [58]. Note that Twitter's Terms of Service prevent researchers from keeping records of accounts or tweets once they are deleted. Therefore, we cannot use data collected from, e.g., the 1% Streaming API; rather, we use publicly available archives from the Wayback Machine [62] and the dataset of verified Twitter accounts from Pushshift [45]. Out of the 20,967 suspended users, we are able to find data for 580 users on Pushshift and 4,210 users on the Wayback Machine, ultimately gathering about 1M tweets.

Reddit. As part of their content policy, Reddit uses content removal (e.g., if it incites violence), including banning users, as well as subreddit quarantines. Since these accounts were suspended, we are not able to collect all their posts by using the regular Reddit API. Instead, we retrieve their posts by querying the search tool provided by the Pushshift API. We are able to find data of 5,216 out of the 6,308 suspended users using the Pushshift API [44] and extract all the posts made by the user.

Ethics. Our study only uses data that is publicly available and, since we do not interact with users in any way, it is not considered as human subjects research by the IRB at our institution. Nonetheless, we acknowledge that linking user accounts across different social networks may have some ethical implications. We limit our study to accounts that used the same profile name on two platforms; arguably, this excludes from our analysis users who did not want to be found after being suspended on Twitter or Reddit, for example by changing their profile name. Moreover, we only report aggregated statistical information and do not perform any analysis at a single account granularity, except the one needed for determining ground truth. Finally, we do not use the collected information to further de-anonymize the users.

3.4 Classifying Accounts as Belonging to the Same Person

Even though two accounts with the same profile name exist on two different platforms, this does not necessarily mean that the accounts belong to the same person. This is particularly true for common profile names. Therefore, to reliably identify accounts controlled by the same person, we look at features of the online accounts on the different platforms and build a machine learning classifier that confirms whether or not the two accounts on the different platforms belong to the same user.

Figure 1: An example of Twitter account metadata we extract: (1) Profile name, (2) Display name, (3) Description , and (4) Location.

Preprocessing. We first process the metadata associated to user accounts by extracting the display name, description, and location of the user. An example of the Twitter metadata is showed in Figure 1. We also extract any URLs in the profile description.

Features. For Twitter, we use the following features to help us characterize whether two accounts with the same profile name on two platforms belong to the same person:

- Jaro Similarity [60] of the display names of the accounts on both platforms. We use this metric to match different display names because it assigns a positive weight for close strings and normalizes it according to the length of the strings, hence outperforming other string compare methods [42].
- Jaccard similarity [39] of the profile descriptions. We use Jaccard Similarity because it performs well for topic modeling and comparing keywords [53, 56]
- Mention of the user's original Twitter handle in the Gab description or viceversa. (We use this feature as manual inspection indicates that several users had the same description mentioned on both platforms, possibly to signal the followers that they are the same person.)
- Number of matching hashtags in the profile descriptions.
- Matching location mentioned on Twitter with the one mentioned in the profile description on Gab.

For Reddit, since the information provided in user profiles is very limited, we only use Jaro similarity between the account's display names.

Labeling. Identifying if two accounts are controlled by the same person cannot be easily automated. To establish ground truth, we had three authors of this paper manually annotate the same subset of 400 randomly selected accounts (200 Twitter+Gab pairs and 200 Reddit+Gab pairs). Each data point was labeled by the three annotators and the label was chosen by majority vote. We then calculate the Cohen's k score [31] between the annotators (Table 1), finding high agreement scores between all the annotators. Based on our labeling, we find that 74.5% of the times if the profile name is the same on the two different platforms, then the account belongs to the same person.

3.5 Classification Performance

After having identified suitable features to identify accounts controlled by the same user across platforms and having established a labeled dataset, we train classifiers to automatically determine if two accounts belong to the same users. We experiment with Random Forest [8], KNN [17], and Decision Tree [55] classifiers, trained using the 400-account annotated dataset discussed above and stratified 10-fold cross-validation. We also use a hold-out validation set of 10 users to fine-tune the parameters of the classifiers. To evaluate performance, we rely on accuracy, along with precision, recall, and F1-score. Table 2 reports the average results on a 10-fold cross validation obtained using different classifier choices. Our best performing model is Random Forest, achieving an average accuracy of 94.6%.

Annotators	Cohen's k
Annotator 1 and 2	0.96
Annotator 1 and 3	0.95
Annotator 2 and 3	0.96

Table 1: Cohen's k score between the different annotators

Classifier	Precision	Recall	Accuracy	F1-Score
KNN	90.2%	92.0%	93.0%	91.1%
Decision Tree	91.4%	90.9%	90.7%	91.1%
Random Forest	**93.2%**	**92.5%**	**94.6%**	**92.8%**

Table 2: Classification scores for the task of predicting if the profile name on the two different platforms belonged to the same person.

We then use Random Forest to classify the accounts on the rest of the dataset. This yields the following results: 3,074 out of 4,790 (64%) suspended accounts on Twitter and 5,216 out of 6,308 (82.7%) on Reddit have an active (matching) Gab account. We next analyze these matched pairs of accounts, to understand how the activity of a user changed after they were suspended on Twitter or Reddit and moved to Gab.

4 ANALYSIS

In this section, we present the results of our analysis aimed to address our four research questions, using the dataset presented above.

4.1 RQ1: Are Accounts Created on An Alternative Platform After Being Suspended?

First, we investigate whether once an account gets suspended on a platform, their owner simply keeps using an alternative account that they already had on another social network, or rather creates a new one.

Unfortunately, neither Twitter nor Reddit provide any information about when accounts get suspended. However, our Gab dataset includes the date when an account was created. Therefore, we can compare this to the last post by the corresponding Twitter/Reddit account on Twitter/Reddit, and use this as a reasonable estimation of whether or not the Gab account was created after suspension.

Since Gab was launched in 2016, to remove any bias in this experiment we only consider users whose Twitter and Reddit accounts

were suspended after January 1, 2017. We find that 58.74% of Twitter users in our dataset (1,152 out of 1,961) created their account on Gab after their last active time on Twitter, presumably after their Twitter account was suspended. For Reddit, we have very similar findings, as 75.88% (3,958 out of 5,216) of the suspended Reddit users create their account on Gab after their last post on Reddit.

Overall, these results show that most Gab accounts in our dataset were created by users after being suspended on the other platform, allowing us to answer RQ1 in the affirmative.

4.2 RQ2: Do Suspended Users Become More Toxic if They Move to Another Platform?

Next, we investigate whether suspended users who move to an alternative platform become more toxic. To do so, we use Google's Perspective API, a free Google service developed by Jigsaw [21] which uses a machine learning model trained on comments (manually) labeled as toxic or non-toxic [16]. The API returns several scores, ranging from 0 to 1, including "Toxicity" and "Severe Toxicity." We use the latter, as prior work [66] shows it to be a more robust indicator of online toxicity.

Platform Toxicity. We first look at the average Severe Toxicity of posts and users on the three platforms. We find that the average Severe Toxicity of *posts* on, respectively, Twitter, Reddit, and Gab is 0.387, 0.42, and 0.498. The average Severe Toxicity of *users* on, respectively, Twitter, Reddit, and Gab is 0.104, 0.3, and 0.334. Overall, this somewhat confirms that there is less moderation on Gab than on Twitter and Reddit.

Also note that on Gab 79% of users had less than 100 posts. This means that most Gab users do not post regularly, which can be explained by the fact that it has been launched in 2016 and did not gain popularity, relatively speaking, until recently [20]. Manual inspection of toxic posts also show that Gab users do complain about other social platforms suspending their accounts, or for not supporting freedom of speech.

Change in Toxicity. We then want to understand whether users become more toxic after they get suspended on a platform and move to an alternative one. Figure 2(a) shows the Cumulative Distribution Function (CDF) of the average Severe Toxicity of accounts on Twitter (before suspension) and Gab (after suspension). As it can be seen, about 60% of users are slightly less toxic after moving to Gab, but 20% of users become much more toxic after moving to Gab. We also perform a two-sample Kolmogorov Smirnov (KS) test on the two curves, which shows that the distributions exhibit statistically significant differences ($D = 0.193$, $p < 0.01$).

We do the same for Reddit, with the CDF shown in Figure 2(b). The figure shows that the toxicity of users increases after moving to Gab. A two-sample KS test confirms statistically significant differences in the distributions ($D = 0.416$, $p < 0.01$).

Overall, our analysis shows that users do tend to become more toxic when they are suspended from a platform and are forced to move to another platform. This is true for users moving from Reddit to Gab, and for 20% of the users moving from Twitter to Gab.

	Twitter	Gab
Average Number of Followers	7,764	167
Average Number of Friends	4,475	201
Average Number of Posts by a User	51,443	485

Table 3: Average user statistics on Twitter and Gab.

4.3 RQ3: Do Suspended Users Become More Active if They Move to Another Platform?

Next, we study how user activity changes once they move to an alternative platform after suspension. We do so by looking at the daily number of posts made by a user on the various platforms.

For Twitter, we compare the activity that a user had on Twitter before being suspended with the activity they have on Gab after suspension. Figure 3(a) shows the CDF of the daily number of posts before and after suspension, showing that activity overall increases. By manually inspecting Gab posts right after suspension, we often find that users complain about being unfairly suspended on Twitter, or including quotes about freedom of speech—i.e., how suspending accounts violates it. A two-sample KS test confirms statistically significant differences in the distributions ($D = 0.242$, $p < 0.01$). We also illustrate the pre-suspension and post-suspension activity of users on Twitter and Gab as a scatter plot in Figure 4(a); this shows that a large number of users were not very active on Twitter but then became so once moving to Gab.

We then perform the same analysis for Reddit; see Figure 3(b). Once again, we find that daily activity increases after users are suspended from Reddit and move to Gab. A scatter plot of the pre-suspension and post-suspension activity of users on Reddit and Gab is shown in Figure 4(b). Again, a two-sample KS test confirms statistically significant differences in the distributions ($D = 0.132$, $p < 0.01$).

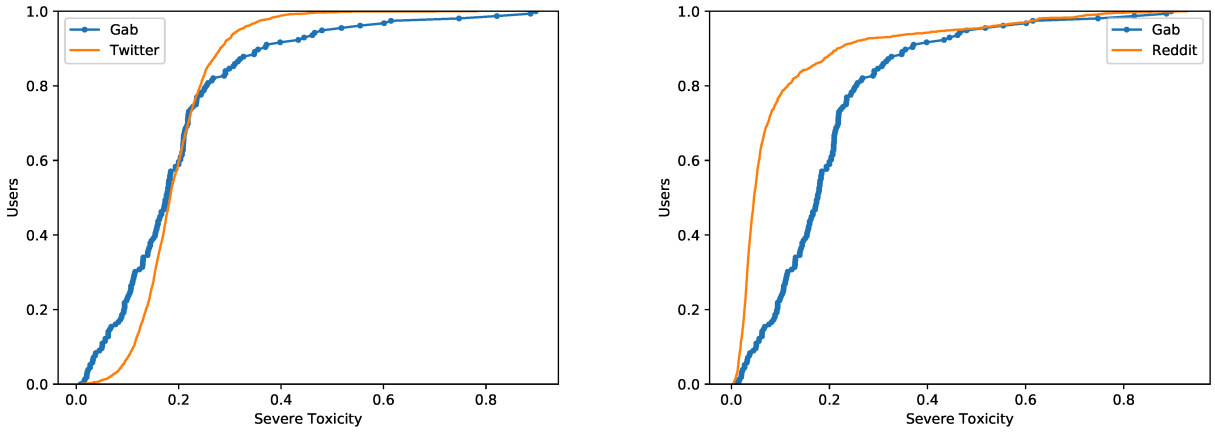
Overall, we show that the activity of suspended users does tend to increase after they move to the alternative platform.

4.4 RQ4: Do Suspended Users Gain More Followers on the Other Platform?

While suspending accounts can arguably curb hateful and toxic content on the platform, there is a chance that this can actually facilitate users moving to an alternative service obtain a broader audience, more appreciative of whatever behavior got them banned from their previous platform.

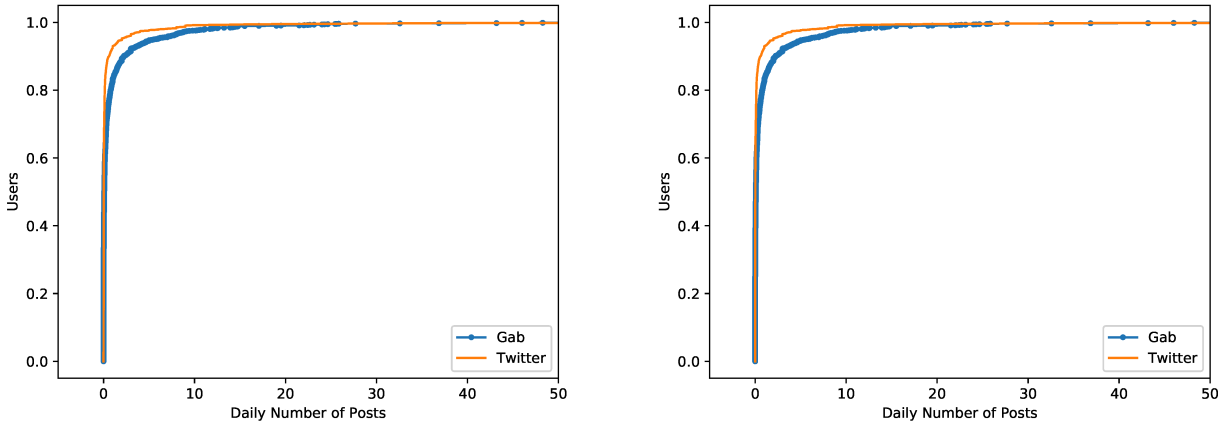
To understand how the size of a user's audience changes after suspension, we look at the number of followers of Twitter and Gab accounts. Table 3 reports the average number of followers, friends, and the number of posts for the users in our dataset on Twitter and Gab. We notice that the number of followers of the suspended users drastically decreased once they moved to Gab.

We also look at the distribution of followers of users on Twitter and Gab. Figure 5 shows the CDF of the number of followers of matched accounts on Twitter and Gab. A two-sample KS confirms statistically significant differences in the distributions ($D = 0.193$, $p < 0.01$). This shows that, even though users tend to become more toxic and more active after they move to the alternative platform, their audience decreases.



(a) (b)

Figure 2: Cumulative Distribution Function (CDF) of Severe Toxicity of a) Twitter users before being suspended and then moving to Gab b) Reddit users before being suspended and then moving to Gab. 60% of users migrating from Twitter become less toxic on Gab, while 20% become much more toxic. In general, users migrating from Reddit become more toxic when moving to Gab.



(a) (b)

Figure 3: Cumulative Distribution Function (CDF) of the daily number of posts of a) Twitter users before being suspended and then moving to Gab. b) Reddit users before being suspended and then moving to Gab. Users become more active after being suspended on Twitter and Reddit and moving to Gab.

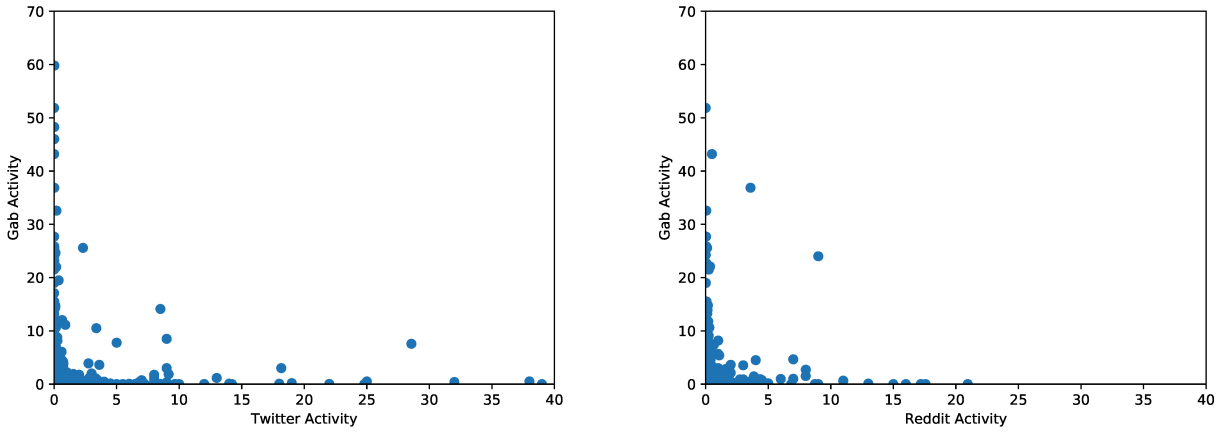
Alas, we cannot perform the same analysis for Reddit because there is no equivalent to the follower concept. On Reddit, users usually join and comment in communities called subreddits, while on Twitter and Gab people follow other users whose posts they are interest in.

5 DISCUSSION AND CONCLUSION

In this paper, we presented a large-scale study of social network users who get suspended from Twitter and Reddit and move to Gab. Overall, we found that users tend to become more toxic and active once they migrate to Gab, but their audience decreases in size. This paper provides a first understanding of user-level migration after being banned on social network platforms, and our results suggest that it is important to understand potentially unintended consequences when designing moderation and suspension mechanisms.

5.1 Implications of Our Results

A common solution to tackle malicious users that send out spam and malware is to block/suspend them. The same solution is adopted by several services to mitigate online toxicity, particularly against



(a) (b)

Figure 4: Scatter plot of the number of posts on a) Twitter before being suspended as compared to their activity on Gab after being suspended. b) Reddit before being suspended as compared to their activity on Gab after being suspended. As it can be seen a number of users became more active after moving to Gab.

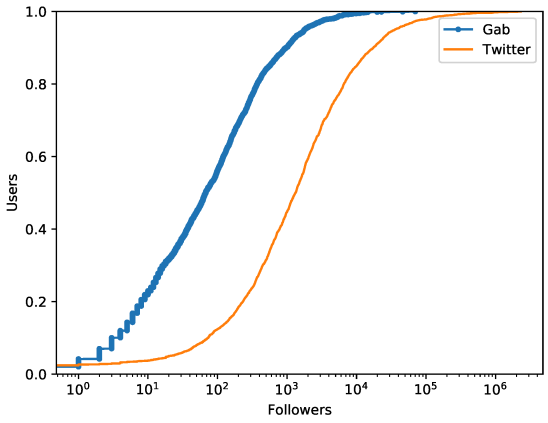


Figure 5: Cumulative Distribution Function (CDF) of the number of followers on Twitter and Gab. After moving to Gab, the number of followers that users reach shrinks.

users who post hateful content. This practice has raised some criticism, with people arguing that users should not be blocked because "selective free speech" is a dangerous precedent [34].

More importantly, suspension practices by online services are usually conducted in isolation, with the goal of keeping the platform safe, and do not take into account broader effects to the online ecosystem, such as the unintended consequences that might arise from suspending users and having them more to alternative communities.

In this paper, we shed some light on these deplatforming dynamics. Next, we discuss the key insights obtained from our analysis, together with some open questions and promising future research directions.

Migration to Alternative Communities. Our study finds that accounts that are suspended on Twitter and Reddit often migrate to Gab. In particular, we can reliably map 2.5% of all Gab accounts to a pre-existing account on Twitter and Reddit that was suspended. This is a lower bound, since our data collection approach for suspended accounts has limitations (see below) and our tracking methodology is only able to identify accounts whose owners maintained the same profile name across platforms. This indicates that users who are suspended for violating the terms of service of a social platform often move to a less moderated community, and the consequences of this action need to be studied.

Our work focused on Gab, but there are other alternative communities where users migrate, like Parler, WrongThink, Voat, and PewTube [4, 40, 64]. Future work should focus on measuring migrations to these communities as well, to paint a better picture of the displacement effect of suspended users and its potential effect on the online ecosystem.

Unexpected Consequences of Suspension. Our results show that users who move to Gab after a suspension tend to become more active and more toxic. At the same time, we find that the audience that these users can reach on the alternative platforms is much smaller than it used to be on Twitter. This quantitative analysis highlights trends that should be studied further, but also leaves many questions that could be answered through qualitative analysis. For example, a manual analysis of the Gab accounts created after a suspension on Twitter showed that these users often wear their suspension as a badge of honor, mentioning it in their profile description and often talking about it in their posts on Gab. It is important to understand how this affects the popularity of these users, in particular in helping them build an engaged following.

Also, our quantitative analysis does not allow us to understand what type of users are following Gab accounts, together with the nature of the toxic speech used by them. Many offline violence instances have been linked with alternative social platforms [3, 37,

51]. As part of future work, qualitative analysis could shed light on how the type of discussion changes after users get suspended and move to alternative platforms, together with understanding whether there is a link between online speech and radicalization.

5.2 Limitations

We now discuss some limitations of our analysis.

Data. One important aspect of our work is being able to understand the role and behavior of users before being suspended on Twitter or Reddit. Twitter's Terms of Service require users to promptly delete all tweets generated by Twitter accounts as soon as they are suspended. To overcome this limitation, we used publicly available datasets from the Wayback Machine Twitter archives and Pushshift. However, this also means that the data we collect is limited, as we could only perform analysis on the tweets and posts that were available from publicly available snapshots. Also, neither Twitter nor Reddit list the suspension date of accounts, and we had to estimate this from the data. Regardless, our data provides us with a lower bound of the number of users that migrated to Gab after being suspended on Twitter or Reddit.

Furthermore, for our dataset we do not take into consideration banned accounts where the usernames are not identical. This biases the dataset to more active users who probably reused their username to gain back old followers.

Google's Perspective API. Our toxicity analysis uses the Perspective API and is thus bounded by its limitations. For instance, [25] discuss how posts can be tampered with in order to reduce the Perspective score, e.g., editing the sentence "Homer Simpson is a moron" to "Homer Simpson is a mor.on" lowers the score from 0.93 to 0.12. Furthermore, the API could have racial biases [10, 13] as, rather than looking at the context of the abusive comment, it may just be looking at the potentially hateful words used. While we acknowledge that the Perspective API has some limitations and biases as reported, we also find that it is the most reliable tool to assess toxicity that is available to the research community.

Reason for Suspension. Due to Twitter's and Reddit's policies, we do not know why users were suspended. This is important, because we expect users getting suspended over hateful conduct to behave differently than those who posted fraudulent links for example. Anecdotally, we find that many Gab posts are about how users were unfairly suspended. In particular, the most toxic posts contain hatred against Twitter and that their free speech and rights have been denied.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their feedback. This work was partially supported by the National Science Foundation under Grants 1827700, 1942610, 2114411, and 2046590.

REFERENCES

[1] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak. Malicious accounts: Dark of the social networks. *Journal of Network and Computer Applications*, 79, 2017.
[2] ADL. When Twitter Bans Extremists, GAB Puts Out the Welcome Mat. https://www.adl.org/blog/when-twitter-bans-extremists-gab-puts-out-the-welcome-mat, 2019.
[3] ADL. Gab and 8chan: Home to Terrorist Plots Hiding in Plain Sight. https://www.adl.org/resources/reports/gab-and-8chan-home-to-terrorist-plots-hiding-in-plain-sight, 2020.
[4] M. Aliapoulios, E. Bevensee, J. Blackburn, E. De Cristofaro, G. Stringhini, and S. Zannettou. An early look at the parler online social network. *arXiv preprint arXiv:2101.03820*, 2021.
[5] W. Alorainy, P. Burnap, H. Liu, A. Javed, and M. L. Williams. Suspended accounts: A source of Tweets with disgust and anger emotions for augmenting hate speech data sample. In *ICMLC*, 2018.
[6] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep Learning for Hate Speech Detection in Tweets. *ArXiv:1706.00188*, 2017.
[7] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn. The Pushshift Reddit Dataset. In *ICWSM*, 2020.
[8] L. Breiman. Random forests. *Machine Learning*, 45(1), 2001.
[9] P. Burnap and M. L. Williams. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet*, 7(2), 2015.
[10] S. Cao. Google's AI Hate Speech Detector Has a âĂŸBlack Twitter' Problem: Study | Observer, 2019.
[11] 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. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), 2017.
[12] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Mean birds: Detecting aggression and bullying on twitter. In *ACM WebSci*, 2017.
[13] D. Coldewey. Racial bias observed in hate speech detection algorithm from Google | TechCrunch, 2019.
[14] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*, 2017.
[15] E. De Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq. Paying for likes? understanding facebook like fraud using honeypots. In *ACM IMC*, 2014.
[16] P. Delgado. How El País used AI to make their comments section less toxic. https://blog.google/outreach-initiatives/google-news-initiative/how-el-pais-used-ai-make-their-comments-section-less-toxic/, 2019.
[17] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang. Efficient kNN classification algorithm for big data. *Neurocomputing*, 195, 2016.
[18] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *WWW Companion*, 2015.
[19] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding. Peer to peer hate: Hate speech instigators and their targets. In *ICWSM*, 2018.
[20] D. Gilbert. Here's How Big Far Right Social Network Gab Has Actually Gotten. https://www.vice.com/en_uk/article/pa7dwg/heres-how-big-far-right-social-network-gab-has-actually-become, 2019.
[21] Google. Perspective API. https://www.perspectiveapi.com, 2020.
[22] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *ACM CCS*, 2010.
[23] H. Habib, M. B. Musa, F. Zaffar, and R. Nithyanand. To Act or React? Investigating Proactive Strategies For Online Community Moderation. *arXiv:1906.11932*, 2019.
[24] G. Hassan, S. Brouillette-Alarie, S. Alava, D. Frau-Meigs, L. Lavoie, A. Fetiu, W. Varela, E. Borokhovski, V. Venkatesh, C. Rousseau, et al. Exposure to extremist online content could lead to violent radicalization: A systematic review of empirical evidence. *International journal of developmental science*, 12(1-2):71–88, 2018.
[25] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective API built for detecting toxic comments. *arXiv:1702.08138*, 2017.
[26] J. Huang, G. Stringhini, and P. Yong. Quit playing games with my heart: Understanding online dating scams. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, 2015.
[27] S. Jhaver, D. S. Appling, E. Gilbert, and A. Bruckman. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), 2019.

[28] J. Kadivar. Online radicalization and social media: A case study of daesh. *International Journal of Digital Television*, 8(3):403–422, 2017.

[29] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi. The Social World of Content Abusers in Community Question Answering. In *WWW*, 2015.

[30] M. Korayem and D. J. Crandall. De-anonymizing users across heterogeneous social computing platforms. In *ICWSM*, 2013.

[31] T. O. Kvalseth. Note on Cohen's Kappa. *Psychological Reports*, 65(1), 1989.

[32] E. Mariconti, J. Onaolapo, S. Ahmad, N. Nikiforou, M. Egele, N. Nikiforakis, and G. Stringhini. What's in a Name?: Understanding Profile Name Reuse on Twitter. In *WWW*, 2017.

[33] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee. Spread of Hate Speech in Online Social Media. In *ACM WebSci*, 2019.

[34] B. Mathew, P. Saha, H. Tharad, S. Rajgaria, P. Singhania, S. K. Maity, P. Goyal, and A. Mukherje. Thou shalt not hate: Countering Online Hate Speech. In *ICWSM*, 2019.

[35] R. McGrath. Twython. https://twython.readthedocs.io/en/latest/, 2020.

[36] M. Mondal, L. Silva, and F. Benevenuto. A Measurement Study of Hate Speech in Social Media. In *WWW*, 2017.

[37] J. Morse. Gab Chat 'likely' to be used by white extremists, according to police. https://mashable.com/article/law-enforcement-documents-violent-white-extremists-encrypted-gab-chat/?europe=true, 2020.

[38] E. Newell, D. Jurgens, H. M. Saleem, H. Vala, J. Sassine, C. Armstrong, and D. Ruths. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *Proceedigns of the Tenth International AAAI Conference on Web and Social Media*, ICWSM, 2016.

[39] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu. Using of Jaccard coefficient for keywords similarity. *Proceedings of the international multiconference of engineers and computer scientists*, 1(6), 2013.

[40] A. Papasavva, J. Blackburn, G. Stringhini, S. Zannettou, and E. De Cristofaro. "is it a qoincidence?": A first step towards understanding and characterizing the qanon movement on voat. co. *arXiv preprint arXiv:2009.04885*, 2020.

[41] R. D. Perera, S. Anand, K. Subbalakshmi, and R. Chandramouli. Twitter analytics: Architecture, tools and analysis. In *Milcom*, 2010.

[42] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils. How unique and traceable are usernames? In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 1–17. Springer, 2011.

[43] H. C. Powell and B. Clifford. De-platforming and the Online Extremist's Dilemma. https://www.lawfareblog.com/de-platforming-and-online-extremists-dilemma, 2019.

[44] Pushshift. Reddit Statistics. pushshift.io, 2020.

[45] Pushshift. Search Twitter Users and Discover Interesting Accounts. https://pushshift.io/twitter-user-search/, 2020.

[46] A. K. Raymond. What We Know About Robert Bowers, Alleged Pittsburgh Synagogue Shooter. https://nymag.com/intelligencer/2018/10/what-we-know-about-robert-bowers-alleged-synagogue-shooter.html, Oct. 2018.

[47] M. Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, E. De Cristofaro, G. Stringhini, and R. West. Does platform migration compromise content moderation?, 2020.

[48] D. Robinson, Z. Zhang, and J. Tepper. Hate Speech Detection on Twitter: Feature Engineering v.s. Feature Selection. In *ESWC Satellite Events*, 2018.

[49] R. Rogers. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, 35(3), 2020.

[50] G. Ryan. Weighing the Value and Risks of Deplatforming. https://gnet-research.org/2020/05/11/weighing-the-value-and-risks-of-deplatforming/, 2020.

[51] E. Schumacher. Far-right social network Gab struggles after Pittsburgh attack. https://www.dw.com/en/far-right-social-network-gab-struggles-after-pittsburgh-attack/a-46065847, 2018.

[52] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber. Analyzing the targets of hate in online social media. *arXiv:1603.07709*, 2016.

[53] D. Spina, J. Gonzalo, and E. Amigó. Learning similarity functions for topic detection in online reputation monitoring. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 527–536, 2014.

[54] G. Stringhini, C. Kruegel, and G. Vigna. Detecting Spammers on Social Networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, 2010.

[55] P. H. Swain and H. Hauska. The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3), 1977.

[56] M. Swartz and A. Crooks. Comparison of emoji use in names, profiles, and tweets. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, pages 375–380, 2020.

[57] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In *ACM IMC*, 2011.

[58] Twitter. Help with locked or limited account, 2020.

[59] S. Volkova and E. Bell. Identifying Effective Signals to Predict Deleted and Suspended Accounts on Twitter Across Languages. In *ICWSM*, 2017.

[60] Y. Wang, J. Qin, and W. Wang. Efficient approximate entity matching using jaro-winkler distance. In *International Conference on Web Information Systems Engineering*, 2017.

[61] Z. Waseem and D. Hovy. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *NAACL Student Research Workshop*, 2016.

[62] Wayback Machine. Wayback Machine APIs – Internet Archive. https://archive.org/help/wayback_api.php, 2020.

[63] J. Weerasinghe, B. Flanigan, A. Stein, D. McCoy, and R. Greenstadt. The pod people: Understanding manipulation of social media popularity via reciprocity abuse. In *The Web Conference*, 2020.

[64] Wired. The wheels are falling off the alt-right's version of the internet. https://www.wired.co.uk/article/alt-right-internet-is-a-ghost-town-gab-voat-wrongthink, 2020.

[65] S. Zannettou, B. Bradlyn, E. De Cristofaro, H. Kwak, M. Sirivianos, G. Stringini, and J. Blackburn. What is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber. In *WWW Companion*, 2018.

[66] S. Zannettou, M. ElSherief, E. Belding, S. Nilizadeh, and G. Stringhini. Measuring and Characterizing Hate Speech on News Websites. In *ACM WebSci*, 2020.