Proceedings of the Sixteenth International AAAI Conference on Web and Social Media (ICWSM 2022)

"I Can't Keep It Up." A Dataset from the Defunct Voat.co News Aggregator

Amin Mekacher, Antonis Papasavva

City University of London, University College London
amin.mekacher@city.ac.uk, antonis.papasavva@ucl.ac.uk

Abstract

Voat.co was a news aggregator website that shut down on December 25, 2020. The site had a troubled history and was known for hosting various banned subreddits. This paper presents a dataset with over $2.3M$ submissions and $16.2M$ comments posted from $113K$ users in $7.1K$ *subverses* (the equivalent of subreddit for Voat). Our dataset covers the whole lifetime of Voat, from its developing period starting on November 8, 2013, the day it was founded, April 2014, up until the day it shut down (December 25, 2020).

This work presents the largest and most complete publicly available Voat dataset, to the best of our knowledge. Along with the release of this dataset, we present a preliminary analysis covering posting activity and daily user and subverse registration on the platform so that researchers interested in our dataset can know what to expect.

Our data may prove helpful to false news dissemination studies as we analyze the links users share on the platform, finding that many communities rely on alternative news press, like Breitbart and GatewayPundit, for their daily discussions. In addition, we perform network analysis on user interactions finding that many users prefer not to interact with subverses outside their narrative interests, which could be helpful to researchers focusing on polarization and echo chambers. Also, since Voat was one of the platforms banned Reddit communities migrated to, we are confident our dataset will motivate and assist researchers studying deplatforming. Finally, many hateful and conspiratorial communities were very popular on Voat, which makes our work valuable for researchers focusing on toxicity, conspiracy theories, cross-platform studies of social networks, and natural language processing.

1 Introduction

Social networks are a primary tool in today's society. They offer countless opportunities for people around the world to connect in various ways, find jobs, entertain themselves, catch up on world happenings, etc. At the same time, social networks sometimes offer a "safe-house" for people that want, among other things, to connect to like-minded individuals towards sharing hate and toxicity (Almerekhi, Jansen, and Kwak 2020; Caffier 2017), discussing controversial matters (Perrigo 2021), and spreading misinformation and disinformation (Hao 2021).

Mainstream social networks suffer from users and communities that organize these conversations on their platforms. A common "solution" the administrators result to is to ban them–*deplatforming*. A social network that is known to have taken this action many times is Reddit, which banned more than $7K$ subreddits (Vincent 2020) from its platform; the first one being in 2014 (Alfonso 2014). Research on deplatforming shows that users that had their communities banned met on forums and even got more toxic than what they used to be (Horta Ribeiro et al. 2021). Other than forums, banned users also move to social networks that allow controversial discussions. One of the platforms that many banned Reddit communities decided to migrate to was Voat.

Voat was a Reddit-esque social network founded in April 2014 and shut down in December 2020 (Robertson 2020). Similar to Reddit, discussions on Voat are divided into various channels–*subverses*–the equivalent of a subreddit. Users can subscribe to as many subverses they wish but cannot moderate more than ten to prevent users gaining undue influence on the platform. User registration on Voat requires only a unique username and a password. Newcomers can upvote, downvote, and comment on existing submissions but cannot create new submissions under subverses until they achieve a certain amount of upvotes on all of their comments.

Since its foundation, Voat gradually gained popularity over its years of operation, especially after every Reddit cleansing (Robertson 2015b; Ohlheiser 2016; Hathaway 2017; Robertson 2018b). Overall, Voat is known for hosting banned extreme communities and users, providing a safe space for like-minded individuals to share their ideas "freely." Voat has attracted the interest of researchers before as it hosted communities like /v/fatpeoplehate, /v/CoonTown, and /v/Nigger (Chandrasekharan et al. 2017), /v/TheRedPill (Saleem et al. 2017), /v/GreatAwakening (Papasavva et al. 2021), etc.

Data Release. In this work, we present, to the best of our knowledge, the largest and most complete dataset of Voat. Along with this paper, we release a dataset (Zenodo 2021) that consists of over $18.6M$ posts from $113K$ users in $7K$ subverses over the lifetime of Voat (November 2013 - December 2020). More specifically, our dataset is *four fold* as it contains the title, body, and metadata of submissions; content and metadata of comments; user profile data; and subverse profile data.

Relevance. Our dataset provides several opportunities to the research community. First, Voat was evidently the place many banned users and communities moved to after being banned from other platforms (Papasavva et al. 2020; Chandrasekharan et al. 2017). To this end, our dataset can assist researchers that focus on deplatforming and user migration. Also, our dataset may aid researchers deepen our understanding on how and when these communities choose their new "home" after a ban. Second, our dataset covers numerous offline events like the 2016 and 2020 US Presidential Elections and debates, Brexit, Epstein's arrest, and various terrorist attacks and unrest around the world that can prove helpful in further analysis of these events. Third, since Voat was a supporter of online freedom of expression for extreme and hateful communities, it contains a variety of slang language and toxic content that can be useful towards understanding hateful communities.

Paper organization. The rest of the paper is organized as follows. First, we briefly explain what Voat is and how it works in Section 2 before going through its history in Section 3. Then, we describe the process of parsing the data released by the Archive team at the Internet Archive Wayback Machine (IAWM) platform, along with the complimentary collection of additional user and subverse data in Section 4. We then describe the structure of our dataset in Section 5 and provide a statistical analysis of the dataset (Section 6), followed by reviewing related work (Section 7). The paper concludes with Section 8.

2 What *was* Voat?

Voat was a Reddit-esque news aggregator launched in April 2014. The mascot of Voat resembles an angry goat, which was designed and freely offered to the website by a user of the site.

Subverses. Discussions on Voat occur in specific groups of interests called "subverses." Users could create subverses on-demand before June 2020, when the administrators disabled this functionality. When a user creates a subverse, they become its owner, hence having complete authority over the subverse: they can deactivate the subverse and appoint other co-owners and moderators. The moderators can delete submissions and comments posted by users and even ban users from posting on the subverse. The owners and moderators can also allow users to post anonymously in their subverse, which replaces the posters' username with a random multi-digit number; not unique to each user. To prevent users from gaining extreme influence on the platform, Voat limits the number of subverses one can own or moderate.

Users. Voat proclaimed itself as a free-speech platform that offered its users anonymity. When newcomers register a new account, Voat does not require any personal details to verify the account, like an email address or phone number. A user can insert a username and a password to register, but if they forget their password, there is no way to recover the account.

After registering a new account, users can subscribe to subverses of interest, comment, upvote, and downvote the comments and submissions but cannot post new submissions. To post a new submission, they first need to acquire ten Comment Contribution Points (CCP). To do so, newcomers post comments on existing submissions trying to collect a *net score* of ten upvotes on all of their comments (one downvote cancels one upvote). The privilege of posting submissions is not guaranteed as users may lose it if their CCP falls below ten. Although this functionality may discourage users from being toxic to each other, it might also prevent users from debating their opinions as others may disagree and downvote them. Voat users often refer to themselves as "goats" due to the platform's mascot.

Submissions and voting system. Voat was a news aggregator platform, hence users can create a new submission by posting a title and a description, accompanied by a link to a news source, optionally. If the poster provides a link, the submission's title becomes a hyperlink to the source website. The domain of the source website appears next to the submission's title, along with the poster's username, and the date and time the submission was posted.

Similar to Reddit, Voat offers a hierarchical, tree-like commenting system: other users can comment on the submission and the comments of other users. Users can upvote or downvote the submission or other users' comments. In contrast with Reddit, Voat displays the total number of upvotes and downvotes a submission or a comment received. Also, the downvote functionality on Voat is not the same as Reddit's: downvoted submissions and comments alert the moderators of spammy or illegal content so they can take action. This functionality enforces the establishment of echo chambers as users usually downvote content that does not align with their beliefs. This usually results in the downvoted user to either losing their submission posting privilege or even being banned from the subverse.

Content visibility. Voat attempted to provide its users with some ephemerality without deleting its content, but hiding it instead. Voat subverses filter submissions under three tabs, namely, hot, new, and top. Each subverse has 500 active submissions in 20 pages (0 to 19). Hot submissions are the ones that are currently active and discussed, new submissions are the ones that were posted most recently, and top submissions are the most popular submissions of the subverse (many comments). Many subverses disabled the functionality of these tabs, and the submissions shown across all three tabs are often the same, just in a different order. We note that our dataset does not contain the tab of submission's as tabs are merely filters and often change based on the status of the submission, e.g., from new to hot.

When a user creates a new submission on a subverse, it would typically appear first on the new tab on page 0. At the same time, the last submission of page 19 is archived but not deleted, meaning if one knows the link to that submission they can still reach it but cannot comment or vote it.

Voat API. Voat supported a JSON API service for some time, but its maintenance stopped in October 2020. To collect the submissions of a subverse, one had to request the API of a specific page number (0 to 19) of a subverse's tab. The response of the API would be the 25 submissions of that page without their comments. To collect the comments, one needs to request them using the submission ID number, in

which the API responds with 25 comments at a time.

Thus, to collect all the submissions from a subverse, one needs to request all 20 pages for the three tabs separately from the API. As explained by (Papasavva et al. 2021), the API does not list the archived subverses and does not respond to requests where the page is above 19. However, if one knows the submission ID and the subverse it was posted in, they can request the API for that specific submission. Since submission IDs on Voat are incremental, one could theoretically collect all of Voat's submissions by requesting the API for each submission ID incrementally for more than 7.5K subverses; that is 7.5K requests, in the worst case, to collect a single submission. To the best of our knowledge, no study or work managed to collect the full Voat dataset.

SearchVoat. A website not associated with Voat, named searchvoat.co, used to collect the Voat submissions and comments for its users to browse.[1] The site does not support an API and does not allow web scraping. After Voat shut down, the site became a news aggregator, similar to Voat.[2]

3 Voat's Troubled History

In this section, we present Voat's history as we believe it highlights the significance of our dataset. WhoaVerse was the original name of the website and it was founded in April 2014. The website was a hobby project of Atif Colo (Voat username @Atko). Justin Chastain later joined Colo as a co-founder (username @PuttItOut). The founders advertised the website as an alternative social network focusing on freedom of expression and speech, which satisfies its users' needs and wants. In December 2014, WhoaVerse changed its name to Voat and marked its mascot as an angry goat.

In June 2015, after Reddit banned various hateful subreddits (Robertson 2015b), including /r/nigger and /r/fatpeoplehate, many Reddit users started registering accounts on Voat. The sudden influx of users overloaded the site, causing temporary down time (Tracy 2015).

On June 19, 2015, Voat's web hosting service, Host Europe,[3] canceled Voat's contract claiming that the site is publicizing abusive, insulting, youth-endangering content, along with illegal right-wing extremist content (Sawers 2015). Some days later, PayPal froze Voat's payment processing services (Pick 2015). In response, Voat shut down four subverses, two of which hosted sexualized images of minors and the founders attributed the shutdown to political correctness (Dewey 2015b). The site moved to a different hosting provider and started accepting cryptocurrency donations.

In July 2015, Reddit banned a popular administrator that caused another influx of Reddit members registering with Voat, leading to more downtime. In an interview, Colo said that they "provide an alternative platform where users would not be censored and still say whatever they want" (Poletti 2015). Voat was the target of DDoS attacks many times and experienced numerous failures during its six years of operation. The most significant attack was in July 2015 (Roberts 2015). Voat, Inc. became a registered corporation in the US in August 2015. Although Voat was based in Switzerland, the U.S. seemed like the best option as explained by Colo in a post: "US law with regards to free speech, by far beats every other candidate country we've researched."

No	Date	Ban
1	May 9, 2014	/r/beatingwomen (Dewey 2015a)
2	Sep 6, 2014	/r/TheFappening (Dewey 2015a)
3	May 7, 2015	/r/nigger (Chandrasekharan et al. 2017)
4	Jun 6, 2015	/r/fatpeoplehate (Chandrasekharan et al. 2017)
5	Nov 23, 2016	/r/pizzagate (Ohlheiser 2016)
6	Nov 7, 2017	/r/incel (Hathaway 2017)
7	Mar 15, 2018	/r/CBTS_Stream (Robertson 2018b)
8	Sep 18, 2018	/r/GreatAwakening (Papasavva et al. 2021)

Table 1: Reddit bans that reportedly affected Voat's activity.

In November 2016, more users relocated to Voat after Reddit banned the /r/pizzagate conspiracy theory subreddit (Ohlheiser 2016). In January 2017, Colo resigned as CEO of Voat due to time availability restrictions and was replaced by Chastain. Chastain ran a fundraiser campaign in May 2017 after announcing that Voat might have to shut down due to financial issues; Voat managed to stay online.

In November 2017, Reddit banned its incel community (/r/incel), and many of its followers reportedly moved to Voat (Hathaway 2017). About a year later, on September 12, 2018, Reddit banned numerous subreddits dedicated to the QAnon conspiracy theory, which again caused many QAnon adherents to migrate to Voat (Papasavva et al. 2021).

In April 2019, Voat's CEO Chastain asked Voat users to stop threatening people as he had been contacted by a "US agency" about the threats posted on the website.[4] In response, Voat users were not pleased to hear that Voat was working with agencies to remove Voat content and "limiting" the site's freedom of expression. Specifically, the first comment on the submission was an anti-Semitic slur, calling for the extermination of Jews (Emerson 2019).

Finally, on December 22, 2020, Voat announced again, now for the last time, that it would shut down due to lack of funding.[5] Chastain explained that he had been funding the site himself since March 2020 but had run out of money. On December 25, 2020, Voat shut down and its last submission was posted by Chastain, noting: "@Atko made the first post to Voat, so I am making the last."[6]

In Table 1, we list some aforementioned Reddit bans that probably affected Voat's activity. Some of these bans previously captured researchers' interest. We use these bans in our analysis in Section 6 to show whether Voat's activity was indeed affected.

4 Data Parsing and Data Collection

This section details the methodology and tools employed for our data collection infrastructure.

Submissions and Comments. Following Voat's shutdown

[1] https://searchvoat.co/search.php

[2] https://searchvoat.co/forum/

[3] https://www.hosteurope.de/en/

[4] https://searchvoat.co/v/Voat/3178819

[5] https://searchvoat.co/v/announcements/4169936

[6] https://searchvoat.co/v/Voat/4174956

	Count	# Users	# Subverses
Submissions	2,334,817	80,063	7,616
Comments	15,731,754	153,827	7,515
Subverses	7,094		
Users	108,451		

Table 2: Number of submissions, comments, user profiles, and subverse profiles in the IAWM dataset.

	Submissions	Comments	Users	Subverses
Total	2,380,262	16,263,309	113,431	7,095

Table 3: Released dataset.

on December 25, 2020, the Archive Team[7] released a set of Voat snapshot captures in Web ARChive (WARC) format (Archive Team 2020), hosted on the Internet Archive Wayback Machine (IAWM). These WARC captures include snapshots the IAWM captured over the lifetime of Voat. A WARC format file consists of single or multiple WARC records (snapshots), and it supports, among other things, the access and scraping of archived data. The files also hold revised and duplicated snapshots (Digital Preservation 2020).

To parse these snapshots into structured data, we set up a Python script to parse the submissions and comments. In our case, every WARC file is a collection of various Voat snapshots the IAWM captured. To facilitate the smooth parsing of the WARC files, we use the *warcio* Python library.[8] This library offers a convenient and reliable way to read a WARC file by streaming every entry included in the file and automatically detecting the *payload*. The payload contains the capture itself, i.e., the HTML DOM tree code of the platform. Each WARC file includes the snapshot of the entire platform for a specific time and date, that is, thousands of submission pages for millions of submissions.

Our parser captures the HTML DOM tree code of each page included in the WARC files serially. Then, it passes the HTML DOM tree to a function that uses the *beautifulsoup* Python library to read and store in JSON format the data and metadata of the submissions and comments, i.e., submission title and content, number of upvotes and downvotes, comments, etc.[9] We ensure that our parser only stores the latest submission version, as WARC files have duplicate data.

We note that although many languages appear in our dataset, the overwhelming majority of posts use the English language. In addition, our parser does not capture or store any visual media, like videos and pictures, since such files are not included in the snapshots. Hence, our dataset is not suitable for researchers focusing on visual media analysis.

User and subverse profiles. To complement our dataset, we also collected user and subverse profiles. A user profile includes data like username and registration date, whereas a subverse profile consists of data like subverse creation date, description, etc. To collect this data, we built a crawler using the IAWM API,[10] beautifulsoup, and HTML requests.[11]

Every user and subverse profile URL is unique, but they all start the same way: *voat.co/u* for the former and *voat.co/v* for the latter. First, we request the IAWM API for all the snapshots whose URLs start like user or subverse URLs. We then collect the responses and parse them into JSON format, storing the latest snapshot the IAWM has in its database for every unique username and subverse profile URL.

The above process results in the dataset summarized in Table 2. We collect a dataset that consists of more than $2.3M$ submissions posted by $80K$ users in $7.6K$ subverses, and over $15.7M$ comments posted by almost $154K$ users. Note that IAWM does not have the profiles of about 500 subverses and hence we only manage to collect the profiles of $7.1K$ subverses (6.8% loss). In addition, we collect almost $108.5K$ unique user profiles.

Data collected via Voat API. In an attempt to complete our dataset, we merge it with the data collected for the (Papasavva et al. 2021) study. For that study, we collected $176K$ submissions and $1.45M$ comments posted from $28K$ users in 241 subverses. Our data collection infrastructure used Voat's API between May 2020 and October 2020, when Voat stopped the maintenance of its API. We find $45.5K$ submissions and $532K$ comments that were missing from the IAWM archive and incorporate them in the released dataset.

Some subverses on Voat offered anonymity to their users by replacing their username with a random eight-digit number (not a unique number for every user). The total number of users that commented or posted a submission (Table 2) does not include anonymous or deleted users. Hence, we assume that Voat's known user base is $155K$ users, at least, based on the data we collect from the IAWM. It is impossible to know the exact Voat user base since Voat never shared the complete list of user profiles, even when it supported a data API service; to collect a user's profile, one needs to know the username. This means that we cannot acquire user profile data of "stalkers." Assuming the total known number of usernames is $155K$, we estimate that about 27.1% of the total users' profile data ($41.6K$) is either missing, or deleted profiles. However, (Papasavva et al. 2021) show that 13% of the $15K$ users being active in QAnon discussions deleted their profiles. Considering that many usernames were deleted every day on Voat, we estimate that this dataset offers the best representation of Voat's user base to date. Incorporating (Papasavva et al. 2021) user data with ours, we find $5K$ additional user and 1 subverse profiles. The final dataset presented and released with this work is detailed in Table 3.

Fair Principles. The data released and presented in this paper aligns with the FAIR guiding principles for scientific data, as described below:[12]

- *Findable*: We assign a unique constant digital object identifier (DOI) to our dataset(Zenodo 2021).[13]

[7]For more details about the Archive team, see wiki.archiveteam.org
[8]https://pypi.org/project/warcio/
[9]https://pypi.org/project/beautifulsoup4/
[10]https://pypi.org/project/waybackpy/
[11]https://pypi.org/project/requests-html/
[12]https://www.go-fair.org/fair-principles/
[13]10.5281/zenodo.5841668

- *Accessible*: Our dataset is openly accessible.
- *Interoperable*: We use JSON format to store our dataset since it is widely used for storing data and can be used in various programming languages. We also provide a detailed description of our dataset's format in Section 5.
- *Reusable*: We provide all the available metadata along with our dataset and we extensively document them in this paper, in Section 5.

Ethical Considerations. The data collected, presented, and released with this paper are available on the Wayback Machine and also used to be accessible (without the need of a registered account) on Voat before it went down. The collection and release of this dataset does not violate Voat's or Wayback Machine's Terms of Service. Although some subverses on Voat allowed users to post anonymously, the overwhelming majority did not offer this functionality. Hence, we collect user profile data of $114K$ users. The only identification of these user profiles is the unique pseudo name, which is not personally identifiable information. Analysis of the activity generated on Voat to other services could potentially be used to de-anonymize users. We note that we followed standard ethical guidelines (Rivers and Lewis 2014) and made no attempt to de-anonymize users.

5 Data Description

We now present the structure of our dataset, available at (Zenodo 2021).

Our dataset consists of submission, comment, user profile, and subverse profile data. We release our data in various newline-delimited JSON files (`.ndjson`).[14] Each line in a `.ndjson` file consists of a JSON object that holds various keys and values. Specifically, we release $7,616$ `.ndjson` files, one for every subverse, that hold the submission data. Similarly, we release $7,515$ `.ndjson` files that have comment data. We inspect our dataset for the missing 101 subverses' comments and find that these subverses have no comment activity, only a small number of submissions. Also, a single `.ndjson` file is released for user profile data, and another for subverse profile data. In total, we release $15,133$ `.ndjson` files. Table 4 lists the keys, value data type, and description of our dataset files.

We choose to release the submission and comment data separately for every subverse as we believe it facilitates researchers that want to focus on specific communities. We also use JSON to release our dataset as it is among the most optimal ways to store and share data as it has extensive documentation and is supported by all popular programming languages.

6 Data Analysis

In this Section we provide statistical analysis and visualization of our dataset.

Posting Activity. First, we show the overall posting activity on Voat. Figure 1 shows the number of submissions and comments per day on the platform. The vertical red dotted lines represent the events listed in Table 1. Although the platform was officially launched in April 2014, the first-ever submission was posted by *@Atko* on November 8, 2013, on the /v/voatdev subverse, that focused on the development of Voat, and at the time, only seven users were posting.

The total number of submissions in 2013 is only 61. These submissions primarily include discussions between @Atko and @PuttItOut in the /v/voatdev subverse. When the platform was launched in 2014, the total number of submissions peaks to $5,268$, then $276K$ in 2015, $324.8K$ in 2016, $397.2K$ in 2017, $382.9K$ in 2018, $439.2K$ in 2019, and for the last year, 2020, $421K$ submissions. Overall, there was no significant increase in activity on the platform after 2016.

The most active day on the site is July 10, 2015, with $5.5K$ submissions. Manual inspection of our dataset indicates that discussions on that day focuses on Donald Trump, vaccine legislation, Reddit's CEO Ellen Pao resigning, and other world happenings. This date is very close to the date Reddit banned hateful communities like /r/fatpeoplehate and /r/nigger (Robertson 2015b). Shortly after Reddit banned these communities, Voat experienced heavy traffic and downtime (Griffin 2015).

Regarding comment activity, only 99 comments were posted in 2013, $13K$ in 2014, $1.6M$ in 2015, $1.8M$ in 2016, $2.1M$ in 2017, $2.4M$ in 2018, $3.8M$ in 2019, and $3.3M$ in 2020 Again, the date with the most comments on the platform is July 10, 2015, with $37.5K$ comments.

In addition, we show the overall activity on Voat in the top ten most subscribed subverses, namely, /v/AskVoat, /v/GreatAwakening, /v/QRV, /v/fatpeoplehate, /v/funny, /v/news, /v/politics, /v/theawakening, /v/videos, and /v/whatever, in Figure 2. We present this analysis to show how active the most popular subverses on Voat were, since we believe that researchers interested in our dataset might consider these findings useful. The vertical red dotted lines on the figure indicate the bans listed in Table 1. When Reddit refugee crowd joined Voat (ban number 1, 3 and 4 from Table 1) many general discussion subverses like /v/AskVoat, /v/news, /v/politics, /v/videos, /v/funny, and /v/whatever became more active, indicating that this new influx of users bolstered the overall activity on the platform.

Interestingly, not all banned subreddits appeared on Voat shortly after a Reddit ban frenzy. The subverse /v/GreatAwakening was created on January 1, 2018, nine months before Reddit banned QAnon subreddits (ban no. 8). This subverse was the 10th most popular subverse when Voat shut down. QAnon discussion on the platform boomed when /v/theawakening and /v/QRV first appeared on Voat on September 12 and September 22, 2018, respectively, with approximately 200 submissions per day on /v/QRV alone. These three subverses turned out to be among the top 5 most active subverses on the platform, with /v/QRV being the most active in both daily submissions and comments on the whole Voat, within only ten days after being banned from Reddit (Ohlheiser 2018).

The figures discussed in this subsection support the reports that Voat was among the main hubs for Reddit migrating communities. In addition, Figure 2 shows that other than general discussion subverses, the most subscribed subverses focused on hate speech (/v/fatpeoplehate)

[14] http://ndjson.org/

Key	Value data type	Description	Key	Value data type	Description
subverse_name_submissions.ndjson (7,616 files)			subverse_name_comments.ndjson (7,515 files)		
title	string	Submission's title	body	string	Comment content
body	string	Submission's content	user	string	Poster's username
user	string	Poster's username	time	string	Time of comment
time	string	Time of submission	date	string	Date of comment
date	string	Date of submission	upvotes	integer	Number of upvotes
upvotes	integer	Number of upvotes	downvotes	integer	Number of downvotes
downvotes	integer	Number of downvotes	comment_id	integer	Comment ID
domain	string	Linked domain	depth	integer	Depth level
link	string	Submission's URL	subverse	string	Subverse's name
submission_id	integer	Submission's ID	root_submission	integer	Parent submission ID
subverse	string	Subverse's name			
user_profiles.ndjson (1 file)			subverse_profiles.ndjson (1 file)		
user	string	Username	subverse	string	Subverse's name
reg_date	string	Registration date	subscriber_count	integer	Subscriber count
moderates	list of strings	Moderated subverses	about	string	Subverse's description
owns	list of strings	Created subverses	date_created	string	Subverse's creation date

Table 4: Description of the keys and data value types.

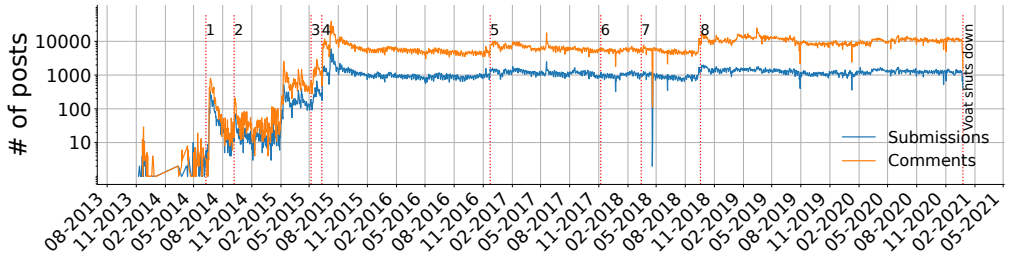


Figure 1: Number of all submissions and comments per day on Voat. Note log scale on y-axis.

and conspiracy theories (/v/QRV, /v/theawakening, /v/TheGreatAwakening).

Submission Engagement. We set to discuss the engagement of the users on the platform. In Figure 3 we plot the Cumulative Distribution Functions (CDF) of the number of comments, upvotes, downvotes, and net votes (upvotes minus downvotes) per submission.

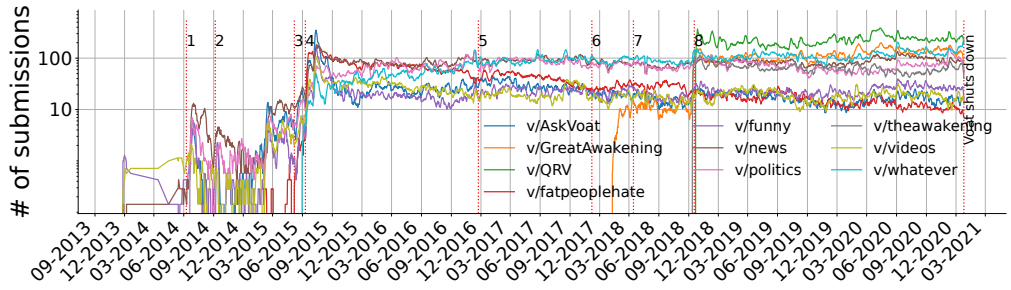
Submissions on Voat get a median number of 3 comments, 7 upvotes, 1 downvotes, and a net score of 7. Comments receive a median 1 and 0 upvotes and downvotes respectively. The most upvoted submission reached over $4K$ upvotes, posted by Atko in /v/announcements in July 2015, explaining that Voat is experiencing heavy traffic due to Reddit bans.The most downvoted submission (392 downvotes) was posted in /v/politics with the title "Dear Media: Please Stop Normalizing The Alt-Right." The most liked comment noted that "someone isn't happy that Voat is succeeding" and reached $1.5K$ upvotes on a submission posted by Atko, discussing the DDoS attacks Voat was experiencing in July 2015. Last, the most disliked comment received 247 downvotes, posted by a user that was asking PuttItOut to reconsider the voting system of the site since they lost their submission posting privileges because of people downvoting them when posting their honest opinion. The user asks the CEOs:

> *[...]ask yourself: Are you fine with a website that caters to some of the most dangerous people currently walking the planet? Take a look at how depraved Trump supporters are, and ask yourself if free speech is worth the cost:[...]*

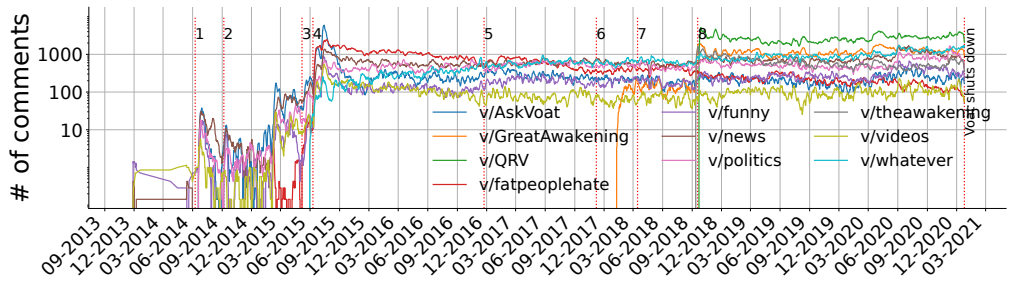
User registration and Subverse creation. In Figure 4 we plot the number of daily user and subverse registrations. The vertical dotted lines mark the bans listed in Table 1.

The first Reddit ban that seemed to have influenced Voat's user base is the one of /r/beatingwomen, on June 9, 2014 (Alfonso 2014) (ban no. 1). Eleven days after the ban, on June 20, Voat had 145 new subverses in a single day, which is the date with the most subverses ever created on the platform. On June 22, there were 112 new subverses.

Moving on to 2015, we find that July 7 is the date with the most users ever registered on Voat (3, 175 registrations), followed by July 5 with 2, 854 registrations. These dates are close to the date Reddit banned various hate subreddits like /r/nigger and /r/fatpeoplehate (bans no. 3 and 4). Also, during the summer of 2015, Reddit changed their free speech and content policy (Robertson 2015a) and the founder noted that "Reddit was not created to be a bastion of free speech." On July 12 and 13, the platform marked two of the five days with the most new subverses created, 125 and 112, respectively. The fifth top date with the most user registrations on



(a) Submissions



(b) Comments

Figure 2: Seven day average number of a) submissions and b) comments per day on the top 10 most subscribed subverses on Voat. Note log scale on y-axis.

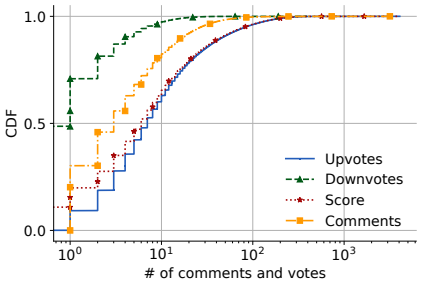


Figure 3: CDF of the number of comments, upvotes, downvotes, and net votes per submission.

Voat is September 13, 2018, with 2,021 users, probably due to Reddit banning QAnon focused subreddits (ban no. 8).

This analysis provides a glance at Voat's user base and subverse changes over the years. It is apparent that Reddit influenced Voat's activity and that the platform was among the preferred Reddit alternatives for banned users.

Links. Since Voat is a news aggregator platform, we analyze the domains the users posted on the site to show what kind of content the userbase of Voat consumed.

For each submission that redirects users to other domains, we retrieve the name of the subverse the submission is posted in, and the external link it redirects to. We count how many times a domain is shared in a community, keeping only the subverse and domain pairs that are the most recurrent in the dataset. The results of this analysis are displayed in an alluvial diagram, which we omit due to space limitations but can be found in the extended version of the paper (Mekacher and Papasavva 2022).

Most of the links that redirect users to Reddit were posted in /v/MeanwhileOnReddit. The subverse focusing on body-shaming, /v/fatpeoplehate, redirected users to Instagram, YouTube, and image sharing services (websites where users can upload images and share the link on other platforms). The /v/news subverse linked YouTube, Voat, online press outlets, and archiving services links. It is known that users in fringe communities avoid sharing the direct link to a website and prefer an archive link instead to avoid monetizing the website (Zannettou et al. 2018). The majority of the alternative news links (Breitbart, GatewayPundit, and Zero Hedge) are posted on /v/news and /v/WorldToday. Most of the Twitter links on the website were posted in /v/QRV and /v/GreatAwakening. Most of the tweets include Donald Trump's tweets and other political discussions on Twitter.

Overall, Voat users shared links to other social networks like Twitter and Instagram. News on the website was shared via legitimate online press outlets and other alternative news outlets, along with archiving services links. Most of the images on the platform were shared on /v/funny, /v/fatpeoplehate, and /v/whatever.

News Aggregators. We now take a deep look into Voat's user ecosystem. We attempt to show how users form clusters based on the subverses they most often engaged with (posted a submission or a comment) to show whether the userbase of Voat is homogeneous or not. Further analysis on Voat's user base may shed light on what content users prefer to see on Voat and whether most Voat's subverses focused on hateful and politically incorrect content.

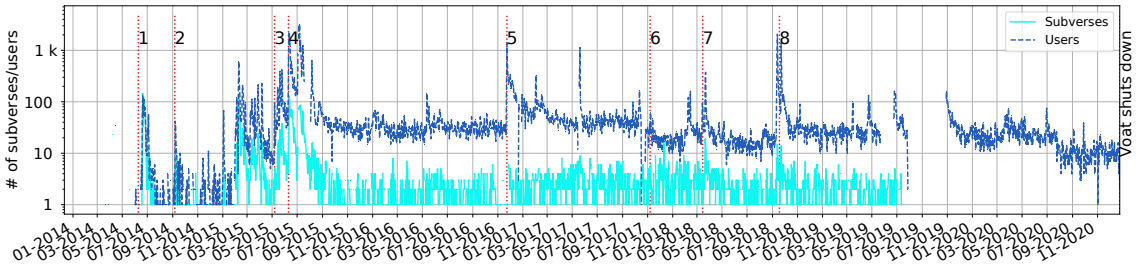


Figure 4: Number of users and subverses registered per day. Note log on y-axis.

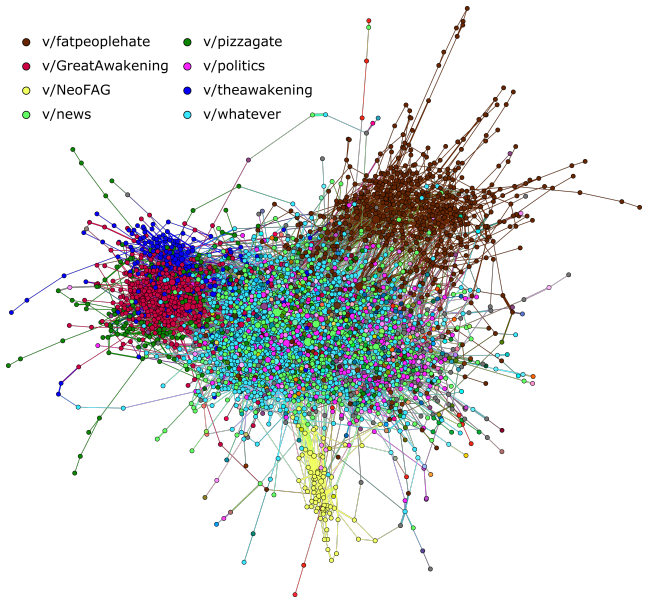


Figure 5: User and subverse interaction ecosystem.

As shown in (Papasavva et al. 2021; Hatemail 2021), some users are responsible for a large amount of content being shared in some communities, leading to imbalances, influencing the content users consume on the platform. By analyzing each user's interactions on Voat, we hope to obverse how all these various communities blended after a mass migration from Reddit, or if Voat was nothing more than an aggregate of small, selective echo chambers.

In Figure 5 we plot a graph network where nodes represent users, and the edges symbolize their interactions. For example, users are linked together if they participated in the same conversation, i.e., they both commented on the same submission, or one of them is the submitter while the other commented. The weight of the edge is given by the number of interactions shared by the two users, and the color represents the subverse where the user participated the most.

The network is composed of a giant cluster, where most subverses are mixed together. This cluster includes /v/politics, /v/news, and /v/whatever, which is expected since these are general discussion subverses, and it is likely that many users meet there for general discussion. However, some subverses are strongly isolated in the network. For example, the /v/NeoFAG (yellow) community shows that most users tend to only engage within that subverse. Similarly, /v/GreatAwakening (red) and /v/theawakening (dark blue) seem to be clustered together and somewhat interacting with /v/pizzagate (dark green). Some users that engage with these three subverses also engage in the general discussion subverses, which is aligned with the findings of (Papasavva et al. 2021). Last, /v/fatpeoplehate (brown) users also seem to form their own cluster while infiltrating the general discussion subverses.

Subverse	EI-Homophily Index
politics	0.50
news	0.40
whatever	0.23
theawakening	0.01
GreatAwakening	-0.25
pizzagate	-0.49
fatpeoplehate	-0.61
NeoFAG	-0.74

Table 5: Average homophily index between subverses and members.

To measure the homophily of these communities, we used the EI homophily index, which is a metric that indicates how many members of a network favor in–group interactions rather than out–group ones. Given a specific node with E external edges, i.e., edges with nodes from the out–group, and I internal edges, i.e., edges with nodes from the in–group, the EI homophily index is given by the equation $EI = (E - I)/(E + I)$.

An index $EI = +1$ indicates that the node only interacts with members of the out–group, whereas $EI = -1$ applies to nodes that only interact within their in–group. Table 5 lists the average EI homophily index of the members of the subverses highlighted in the legend of Figure 5.

Users who are very active on subverses like /v/politics and /v/news have a high average EI homophily index, meaning they mostly interact with users from the out–group. The opposite can be said for the /v/theawakening, /v/GreatAwakening, /v/pizzagate, and especially, /v/NeoFAG and /v/fatpeoplehate. These communities do not converse much outside their social group. The EI index is almost zero for /v/theawakening, meaning its users interact as much with the out–group as with the in–group. By looking at the community, this can be ex-

plained by the fact that users from the communities gravitating around the QAnon narrative, i.e., /v/theawakening, and /v/GreatAwakening, are more connected than other communities. As a result, the external edges can be nothing more than crossovers between these two subverses. The userbases of /v/NeoFAG and /v/fatpeoplehate seem to be the ones that only prefer to interact with members of their community.

We present this analysis to motivate researchers studying user interactions and echo chambers. Further research using our dataset may shed light on whether Voat was a bastion of echo chambers or not, along with what narratives users within these communities exchanged.

7 Related Work

In this section, we present existing work focusing on Voat, and other dataset papers similar to ours. Voat attracted the interest of researchers over the past years, especially after Reddit started banning communities in 2015. Although some papers mention that their dataset is available upon request, these datasets only include data from a couple of subverses that cover a short period of time. To the best of our knowledge, our Voat dataset is: 1) the only one to be openly and publicly available online; and 2) the most complete and largest one, covering the whole history of Voat, along with data of the users that ever posted a submission or a comment on the platform.

Voat research. (Newell et al. 2016) collect data from various platforms, including Voat and Reddit and perform computational analysis to identify the primary motivations that drive users to move to other platforms. (Chandrasekharan et al. 2017) collect data from 4chan, Reddit, MetaFilter, and Voat and build a model to detect abusive content online. Subverses used in this work include /v/CoonTown, /v/Nigger, and /v/fatpeoplehate, all focused on hate towards individuals of specific body or race characteristics, created on Voat shortly after the 2015 Reddit bans (Robertson 2015b). Similarly, (Saleem et al. 2017) collect data from Reddit, Voat, and three online forums to train a classifier that detects hateful speech. Their Voat dataset includes data from /v/CoonTown, /v/fatpeoplehate, and /v/TheRedPill. A study on deepfakes finds that pornographic deepfakes are mainly created for circulation within the community (Popova 2019). The study uses data from Voat's /v/DeepFake and the site mrdeepfakes.com, which both were created after Reddit banned the subreddit /r/DeepFakes in 2018 (Robertson 2018a).

(Khalid and Srinivasan 2020) compare the features of 872K comments from /v/politics, /v/television, and /v/travel, to Reddit and 4chan comments building a classifier that predicts the origin of the comments based on its style and content. (Papasavva et al. 2021) collect 0.5M posts from /v/GreatAwakening, /v/news, /v/politics, /v/funny, and /v/AskVoat to provide an empirical exploratory analysis of the QAnon community on Voat. They find, among other things, that /v/GreatAwakening is not as toxic as the general discussion subverses. Finally, (Papasavva et al. 2022) compare Voat's /v/GreatAwakening and /v/news posts to 4chan, 8kun, Reddit, and Q drops (posts posted by "Q," the mastermind behind the QAnon conspiracy theory) on a large scale study on QAnon. They find that Voat posts are as threatening as Q drops and that content creators on Reddit and Voat only consist of a small portion of the total community.

Other datasets. One of the largest Reddit datasets is the one of (Baumgartner et al. 2020a), which presents an archiving platform that collects Reddit data and makes them available to researchers since 2015. The same platform also published over 27.8K channels and 317M messages from 2.2M users from Telegram (Baumgartner et al. 2020b). (Fair and Wesslen 2019) release a dataset of 37M posts, 24.5M comments, and 819K user profiles collected from Gab. (Aliapoulios et al. 2021) published a dataset consisting of 183M posts and 13.25M user profiles from Parler, a Twitter alternative. Last, (Papasavva et al. 2020) present a dataset with over 3.3M threads and 134.5M posts from the Politically Incorrect board (/pol/) of the imageboard forum 4chan.

8 Conclusion

In this work, we present and release a Voat dataset comprising more than $2.38M$ submissions and $16.2M$ comments posted from $113K$ users in over $7K$ Voat subverses. We combine data collected from Voat API and IAWM released archives to complete the dataset to the best of our ability. Voat shut down on December 25, 2020, and its data are now otherwise inaccessible. In this work we also perform a preliminary analysis of the released dataset so researchers interested in it can know what to expect.

Overall, we hope this work further motivates and assists researchers focusing on deplatforming and how users organize migrations to other platfroms. In addition, our dataset could also help answer numerous questions about how "free–speech" sites operate, e.g., do moderators ban users that express opinions other than the ones aligned with the narratives of a subverse? How do users vote and how toxic are they towards such content? Do sites like these incentivize users to form echo chambers? What kind of content users in these communities consume, etc.? Also, our dataset could assist multi–platform studies to understand similarities and differences of different communities. Last, since Voat was a bastion of free-speech, we are confident that access to our dataset could assist researchers towards training algorithms in natural language processing and detecting hate speech, fake news dissemination, conspiracy theories, etc. Finally, other than quantitative work, we hope that the data can also be used in qualitative studies of specific events, social theories, and communities.

Acknowledgments

This work was partially funded by the UK EPSRC grant EP/S022503/1 that supports the UCL Centre for Doctoral Training in Cybersecurity. Any opinions, findings, and conclusions or recommendations expressed in this work are those of the authors.

References

Alfonso, F. 2014. Reddit bans infamous forum about beating women. https://bit.ly/3LGWaqR. Accessed: 2022-01-14.

Aliapoulios, M.; Bevensee, E.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; and Zannettou, S. 2021. An early look at the parler online social network. In *ICWSM*.

Almerekhi, H.; Jansen, S. b. B. J.; and Kwak, c.-s. b. H. 2020. Investigating toxicity across multiple Reddit communities, users, and moderators. In *WWW*.

Archive Team. 2020. Archive Team: Voat. https://archive.org/details/archiveteam_voat. Accessed: 2022-01-14.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020a. The pushshift reddit dataset. In *ICWSM*.

Baumgartner, J.; Zannettou, S.; Squire, M.; and Blackburn, J. 2020b. The Pushshift Telegram Dataset. In *ICWSM*.

Caffier, J. 2017. Here Are Reddit's Whiniest, Most Low-Key Toxic Subreddits. https://bit.ly/3LxQOy6. Accessed: 2022-01-14.

Chandrasekharan, E.; Samory, M.; Srinivasan, A.; and Gilbert, E. 2017. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *ACM SIGCHI*.

Dewey, C. 2015a. The 'Reddit exodus' is a perfect illustration of the state of free speech on the Web. https://wapo.st/33zyZOc. Accessed: 2022-01-14.

Dewey, C. 2015b. This is what happens when you create an online community without any rules, part 2. https://wapo.st/3qH9aUn. Accessed: 2022-01-14.

Digital Preservation. 2020. Sustainability of Digital Formats: Planning for Library of Congress Collections. https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml. Accessed: 2022-01-14.

Emerson, S. 2019. Founder of Voat, the 'Censorship-Free' Reddit, Begs Users to Stop Making Death Threats. https://bit.ly/3mT77vr. Accessed: 2022-01-14.

Fair, G.; and Wesslen, R. 2019. Shouting into the void: A database of the alternative social media platform gab. In *ICWSM*.

Griffin, A. 2015. Reddit alternative breaks because so many people leave site after harassment scandal. https://www.independent.co.uk/life-style/gadgets-and-tech/news/reddit-alternative-breaks-because-so-many-people-leave-site-after-harassment-scandal-10321474.html. Accessed: 2022-01-14.

Hao, K. 2021. How Facebook and Google fund global misinformation. https://bit.ly/3r4WNTk. Accessed: 2022-01-14.

Hatemail. 2021. Qoup d'état: What QAnon's Reddit Migration Tells Us About Misinformation. https://bit.ly/3eNa8Jq. Accessed: 2022-01-14.

Hathaway, J. 2017. Why Reddit finally banned one of its most misogynistic forums. https://www.dailydot.com/unclick/reddit-incels-ban/. Accessed: 2022-01-14.

Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. In *CSCW*.

Khalid, O.; and Srinivasan, P. 2020. Style Matters! Investigating Linguistic Style in Online Communities. In *ICWSM*.

Mekacher, A.; and Papasavva, A. 2022. "I Can't Keep It Up." A Dataset from the Defunct Voat.co News Aggregator. *arXiv:2201.05933*.

Newell, E.; Jurgens, D.; Saleem, H. M.; Vala, H.; Sassine, J.; Armstrong, C.; and Ruths, D. 2016. User migration in online social networks: A case study on reddit during a period of community unrest. In *ICWSM*.

Ohlheiser, A. 2016. Fearing yet another witch hunt, Reddit bans 'Pizzagate'. https://wapo.st/3FrdXij. Accessed: 2022-01-14.

Ohlheiser, A. 2018. Reddit bans r/greatawakening, the main subreddit for Qanon conspiracy theorists. https://wapo.st/3GsSHtA. Accessed: 2022-01-14.

Papasavva, A.; Aliapoulios, M.; Ballard, C.; De Cristofaro, E.; Stringhini, G.; Zannettou, S.; and Blackburn, J. 2022. The gospel according to Q: Understanding the QAnon conspiracy from the perspective of canonical information. In *ICWSM*.

Papasavva, A.; Blackburn, J.; Stringhini, G.; Zannettou, S.; and De Cristofaro, E. 2021. "Is it a Qoincidence?": An Exploratory Study of QAnon on Voat. In *WWW*.

Papasavva, A.; Zannettou, S.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2020. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *ICWSM*.

Perrigo, B. 2021. Twitter Offers More Transparency on Racist Abuse by Its Users, but Few Solutions. https://bit.ly/3LDZMtJ. Accessed: 2022-01-14.

Pick, R. 2015. PayPal Cuts Off Reddit Clone Voat Over Obscenity. https://bit.ly/32PD4xz. Accessed: 2022-01-14.

Poletti, T. 2015. Creator of surging Reddit rival Voat: We will avoid same mistakes. https://on.mktw.net/3sVWc81. Accessed: 2022-01-14.

Popova, M. 2019. Reading out of context: pornographic deepfakes, celebrity and intimacy. *Porn Studies*.

Rivers, C. M.; and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research*.

Roberts, J. 2015. New Reddit rival Voat hit by DDoS attack. https://bit.ly/3J8kJv9. Accessed: 2022-01-14.

Robertson, A. 2015a. Was Reddit always about free speech? Yes, and no. https://bit.ly/3rjEh8K. Accessed: 2022-01-14.

Robertson, A. 2015b. Welcome to Voat: Reddit killer, troll haven, and the strange face of internet free speech. https://www.theverge.com/2015/7/10/8924415/voat-reddit-competitor-free-speech. Accessed: 2022-01-14.

Robertson, A. 2018a. Reddit Bans 'deepfakes' AI Porn Communities. https://bit.ly/35hS1Wy.

Robertson, A. 2018b. Reddit has banned the QAnon conspiracy subreddit r/GreatAwakening. https://fxn.ws/3IiXYor. Accessed: 2022-01-14.

Robertson, A. 2020. 'Free speech' Reddit clone Voat says it will shut down on Christmas. https://bit.ly/35IJJvm. Accessed: 2022-01-14.

Saleem, H.; Dillon, K.; Benesch, S.; and Ruths, D. 2017. A web of hate: Tackling hateful speech in online social spaces. *TACOS*.

Sawers, P. 2015. Amid censorship brouhaha, Reddit clone Voat has its servers closed by hosting provider. https://bit.ly/3sSXO2C.

Tracy, A. 2015. Web Host Drops Voat After Disgruntled Redditors Flock To The Platform. https://bit.ly/32KhieP. Accessed: 2022-01-14.

Vincent, J. 2020. Reddit reports 18 percent reduction in hateful content after banning nearly 7,000 subreddits. https://www.theverge.com/2020/8/20/21376957/reddit-hate-speech-content-policies-subreddit-bans-reduction. Accessed: 2022-01-14.

Zannettou, S.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; and Stringhini, G. 2018. Understanding web archiving services and their (mis) use on social media. In *ICWSM*.

Zenodo. 2021. Dataset: "I Can't Keep It Up." A Dataset from the Defunct Voat.co News Aggregator. https://zenodo.org/record/5841668. Accessed: 2022-01-14.