

# iDRAMA-Scored-2024: A Dataset of the Scored Social Media Platform from 2020 to 2023

Jay Patel<sup>1</sup>, Pujan Paudel<sup>2</sup>, Emiliano De Cristofaro<sup>3</sup>, Gianluca Stringhini<sup>2</sup>, Jeremy Blackburn<sup>1</sup>

<sup>1</sup>Binghamton University

<sup>2</sup>Boston University

<sup>3</sup>University of California Riverside

jpatel67@binghamton.edu, ppaudel@bu.edu, emiliand@ucr.edu, gian@bu.edu, jblackbu@binghamton.edu

## Abstract

Online web communities often face bans for violating platform policies, encouraging their migration to alternative platforms. This migration, however, can result in increased toxicity and unforeseen consequences on the new platform. In recent years, researchers have collected data from many alternative platforms, indicating coordinated efforts leading to offline events, conspiracy movements, hate speech propagation, and harassment. Thus, it becomes crucial to characterize and understand these alternative platforms. To advance research in this direction, we collect and release a large-scale dataset from Scored – an alternative Reddit platform that sheltered banned fringe communities, for example, c/TheDonald (a prominent right-wing community) and c/GreatAwakening (a conspiratorial community). Over four years, we collected approximately 57M posts from Scored, with at least 58 communities identified as migrating from Reddit and over 950 communities created since the platform’s inception. Furthermore, we provide sentence embeddings of all posts in our dataset, generated through a state-of-the-art model, to further advance the field in characterizing the discussions within these communities. We aim to provide these resources to facilitate their investigations without the need for extensive data collection and processing efforts.

## Introduction

Social media has become more than just a platform for social connection and entertainment, it serves as a platform for the mass consumption of news (Shearer 2018), spreading misinformation and hate speech (Zannettou et al. 2019; Ziems et al. 2020), and propagation of political narratives (Ng, Moffitt, and Carley 2022; Starbird, Arif, and Wilson 2019). As its use continues to rise, this has become an arms race to enhance the existing systems that facilitate a safer space for everyone.

Social platforms like Reddit offer a space for user groups with shared interests, known as communities, to share and discuss content (Porter 2004). However, if communities violate platform policies, Reddit imposes bans, pressing them to migrate elsewhere (e.g., alternative platforms) on the Web. Researchers have studied various alternative platforms, for example, Voat (Papasavva et al. 2021), Poal (Papasavva and

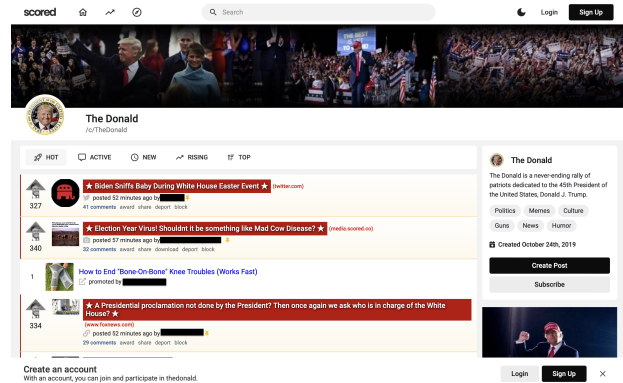


Figure 1: Visual appearance of c/TheDonald’s home page on Scored platform.

Mariconti 2023), and Lemmygrad (Balci, Sirivianos, and Blackburn 2023) (an alternative to Reddit), Bitchute (an alternative to YouTube) (Trujillo et al. 2020), Gab, Gettr, and Parler (alternatives to Twitter) (Zannettou et al. 2018; Paudel et al. 2022; Aliapoulios et al. 2021a). Previous studies uncovered radicalization within migrated communities but also showed connections to real-world events on alternative platforms (Aliapoulios et al. 2021a; McIlroy-Young and Anderson 2019; Balci, Sirivianos, and Blackburn 2023). Therefore, exploring these alternative platforms and understanding their influence is a growing focus of research efforts.

Scored<sup>1</sup>, an alternative to Reddit, is an emerging platform that has become a refuge for many communities banned from mainstream platforms. We focus on collecting data from the Scored platform since its inception. As of December 2023, Scored hosts c/TheDonald (a community for Donald Trump’s supporters) and other communities (e.g., c/ConsumeProduct, c/GreatAwakening, c/FatPeopleHate) created in response to bans from mainstream platforms. Scored remains understudied, and our work contributes to advancing the exploration of banned migrant communities on alternative platforms by providing large-scale data.

**Contributions.** In this work, we release a dataset col-

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>1</sup>Scored is accessible from both <https://scored.co> and <https://communities.win>.

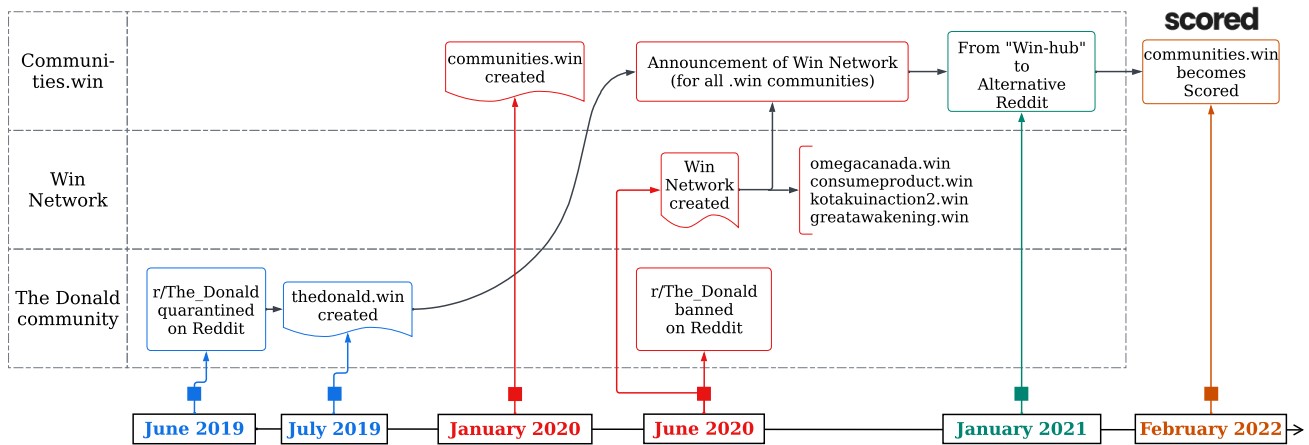


Figure 2: Evolution of Scored.

lected from Scored. We collect approximately 57M posts from Scored, spanning four years. Our contributions go beyond releasing Scored’s data alone; using a state-of-the-art model, we create and release sentence embeddings of the collected data. We adhere to FAIR principles in releasing this data, discussed in a later section.

 [zenodo.org/records/10516043](https://zenodo.org/records/10516043)

 [hf.co/datasets/iDRAMALab/iDRAMA-scored-2024](https://hf.co/datasets/iDRAMALab/iDRAMA-scored-2024)

 [github.com/idramalab/iDRAMA-scored-2024](https://github.com/idramalab/iDRAMA-scored-2024)

**Relevance.** With the release of this dataset, we believe that it will be valuable to the research community in various ways. Scored emerged as an alternative platform to provide shelter to many banned communities after “The Great Ban” event (Reddit banned around 2,000 communities in June 2020) (Cima et al. 2024; Reddit 2020; Newton 2020). Scored hosts the far-right community, c/TheDonald, the QAnon conspiracy community, c/GreatAwakening, and more extremist ideological communities like c/ConsumeProduct, c/IP2Always. This dataset constitutes an essential resource for researchers in understanding the migration pattern and radicalization of banned communities in online spaces.

Moreover, Scored emerged at a crucial time in U.S. history, during the U.S. Capitol riots on January 6th, 2021, an event where protestors stormed the U.S. Capitol in denial of the newly elected president ‘Joe Biden.’ This event fueled discussions among communities like c/TheDonald, c/GreatAwakening, and c/ConsumeProduct. Our dataset includes the period of this event, which we believe will be helpful for researchers and journalists understanding real world events and its interplay with fringe social networks from a data-driven approach perspective. Overall, this dataset provides a unique view to study communities’ platform migration patterns and radicalization in online spaces, constituting historical political events.

## Background

As an emerging platform, Scored has a limited body of literature documenting its evolution. In this section, we discuss the history and outlook of Scored social media platform.

### History

In this section, we discuss Scored’s history, formerly known as “Communities.win,” outlining the key milestones presented as a high-level overview in Figure 2.

**Migration of r/The\_Donald Community (Jun-Jul 2019).** In June 2019, Reddit quarantined the alt-right subreddit r/The\_Donald, due to policy violations involving hate and violence (Ribeiro et al. 2021b). Within a month of the quarantine, the community began to migrate to a self-hosted platform named “thedonald.win.” Records obtained from VirusTotal<sup>2</sup> show the domain (thedonald.win) being registered in July 2019. The migration of r/The\_Donald to a self-hosted platform was somewhat unique, especially compared to other Reddit communities, e.g., r/greatawakening, that opted for Voat.co (Papasavva et al. 2021).

**Rise of Communities.win (Jun 2020).** The Communities.win domain was registered in January 2020, but no webpage references are found on the Wayback Machine until June 28, 2020. The first post<sup>3</sup> on communities.win introduced a Win-network (i.e., a list of Win communities) during the same month when Reddit banned over 2,000 subreddits in June 2020. Additionally, eight Win communities – OmegaCanada.win, GavinMcInnes.win, ConsumeProduct.win, KotakuInAction.win, KotakuInAction2.win, WeekendGunnit.win, GreatAwakening.win, and TuckerCarlson.win – were registered within that same month. This suggests a possible coordinated effort and/or a migration trend, potentially influenced by a strategic move

<sup>2</sup>All domain informations are retrieved through VirusTotal. API reference: <https://developers.virustotal.com/reference/overview>

<sup>3</sup>Announcement of Win communities (June 30, 2020): <https://web.archive.org/web/20200630065948/https://communities.win/p/GIEblc8H/list-of-win-communities/>

made by r/The\_Donald community in response to the Reddit ban. During this period, Communities.win predominantly functioned as an information hub for Win network announcements.

**Community-Oriented Shift (Jan 2021).** Transforming from a central hub for Win network updates, Communities.win responded to user requests<sup>4</sup> by hosting non-political communities (e.g., Funny and Gaming).

**Emergence of Scored (Feb 2022).** Communities.win adopted the name “Scored” and a new logo<sup>5</sup> in February 2022. Simultaneously, the web page of Scored.co, which was registered in 2012, but was for sale until at least July 2019, displayed an interface identical to that of Communities.win.

## What is Scored?

Scored describes itself as “a platform looking to unblur the lines between entertainment and politics.”

**Users & Content Accessibility.** Scored automatically enrolls<sup>6</sup> new users in about ten communities, for example, Memes, Food, Technology, Animals, Sports. Additionally, users can access posts from unjoined communities through search or direct visits. Posts act as discussion threads, allowing users to comment and have the option to include various media types like images, videos, hyperlinks, or text. Posts also display flair, similar to hashtags, and offer the option to be marked as Not Safe For Work (NSFW).

**Moderation.** According to Scored’s content policy<sup>7</sup>, the platform enforces violations through warnings, temporary or permanent removal of users or privileges, and, to some extent, imposing restrictions on the community. Scored also has one or more moderators for each community. However, the content policy does not explicitly mention the ban of a community; instead, they provide options to mark community as unsafe for work or as political.

**Monetization.** Scored offers a pro badge to paid subscribers at USD 6.99 per month. Pro subscribers get ad-free browsing, can give gold badges to posts or comments, and can award a one-week pro subscription to others.

## Dataset

In this section, we discuss the data collected from Scored, detailing the methodology used for collection and addressing potential limitations. Furthermore, we discuss the dataset structure, data release and ethical considerations in collecting data at scale.

<sup>4</sup>Community request thread: <https://web.archive.org/web/20210123025045/https://communities.win/p/11S13qIE19/community-request-thread/>

<sup>5</sup>Appearance of ‘Scored’ name and logo: <https://web.archive.org/web/20220223204351/https://communities.win/>

<sup>6</sup>Automatic enrollment to the listed communities is effective as of June 2023.

<sup>7</sup>Scored content-policy: <https://help.scored.co/knowledge-base/content-policy/>

	Count	#Users	#Communities
Submissions	6,293,980	131,692	975
Comments	50,521,604	194,084	686
<b>Total Posts</b>	<b>56,815,584</b>	<b>226,765</b>	<b>976</b>

Table 1: Overview of dataset

## Data Collection

Using a custom crawler, we collect Scored data ranging from January 1, 2020, to December 31, 2023. This yields a dataset of 50.52M comments and 6.29M posts (see Table 1). Table 1 also provides a comprehensive overview of the count of unique authors contributing at least once during this time-frame and the number of communities within our dataset. Scored’s posts are functionally equivalent to Reddit’s submissions. Throughout the rest of the paper, we use Reddit terminology as readers may be more familiar with it. Therefore, we refer to Scored posts as a combination of Scored submissions and comments.

**Limitations.** It is important to highlight that Scored employs moderation practices leading to users’ posts removal. During our data collection, we observe instances where some posts lack author or comment/submission content, indicating their deletion before our crawler collect it. Another limitation is associated with the attributes *score*, *score\_up*, and *score\_down* (see Section Dataset Structure). These metrics are collected only once through our crawler during the data collection phase without additional updates to the attributes. Hence, dataset users should be careful while utilizing these metrics for analytical purposes.

**Ethical Considerations.** In this work, we collect and analyze publicly available data without engaging with users directly. As a result, our institution’s Institutional Review Board (IRB) does not classify this work as human subject research. Nevertheless, we adhere to established ethical guidelines, as outlined by Rivers and Lewis (Rivers and Lewis 2014), to protect the rights and privacy of users. In the release of this dataset, we hash the actual usernames and plan to make them available only to bona fide researchers upon request. In this paper, we also present aggregated data and refrain from any attempts to re-identify individuals within our dataset.

## Dataset Structure

Our dataset contains various attributes collected through our custom crawler. Below, we outline the attributes available in our dataset for Scored submissions:

- **Identifier:** Unique identifier associated with each submission (*uuid*).
- **Primary Attributes:** Essential information about submissions, including *author*, *community*, *created*, and *date*. Note that the *created* attribute is in Milliseconds since Unix Epoch. We convert this into *date* for ease of analysis for dataset users.
- **Content:** Each submission includes a *title* attribute representing the submission title and a *raw\_content* attribute

containing the submission content.

- **Metadata:** Attributes providing metadata for each submission are: 1) *link*: URL if the submission is a link, 2) *type*: indicating whether the submission is text or a link, 3) *domain*: base domain if the submission is a link, 4) *tweet\_id*: associated tweet id if the submission is a Twitter link, and 5) *video.link*: associated video link if the submission is a video.
- **Flairs:** Similar to Reddit submission flairs, which is a way to tag a submission with a certain keywords, allowing end-users to know what the submission is about. Our dataset includes two attributes: *post\_flair\_text* and *post\_flair\_class*. Examples of *post\_flair\_text* include HIGH ENERGY, SLEEPY JOE, MEME ARMY, and more.
- **Scores:** Metrics providing information about the scores associated with sampled submissions, including *score\_up*, *score\_down*, and *score*. Score attributes reflect user feedback, similar to voting, on submissions or comments. We collect these metrics only once during data collection, with no further updates.
- **Submission Specific Flags:** Attributes indicating specific flags for submissions, such as whether the submission is nsfw (not safe for work), posted by an admin, and more. The available attributes include: *is\_nsfw*, *is\_admin*, *is\_image*, *is\_video*, *is\_deleted*, *is\_twitter*, and *is\_moderator*.

For all Scored comments, the following attributes are available:

- **Identifier:** *uuid*.
- **Content:** *raw\_content*.
- **Primary Attributes:** *author*, *created*, *community*, *date*.
- **Scores:** *score*, *score\_up*, *score\_down*.
- **Comment Specific Flags:** *is\_deleted*, *is\_moderator*.

## Data Embeddings

Using the transformer-based technique, specifically the INSTRUCTOR (Su et al. 2022) model, we generate embeddings<sup>8</sup> of 768 dimensions with a float 32-bit data type. This state-of-the-art model is optimized for diverse downstream tasks, for example, classification, clustering, and semantic similarity. We used a NVIDIA A100 GPU with 80GB of memory to generate these embeddings, running for approximately 40 hours.

**Embedding Methodology:** We generate embeddings for all posts using attributes from our collected dataset. Specifically, we use the *raw\_content* field for comments. However, for submissions, we use two attributes: *title* and *raw\_content*, and combine the content of both before generating sentence embeddings. Then, we perform a preprocessing step, which involves removing hyperlinks from each post. Also, we exclude all posts if their character length becomes zero after this preprocessing step. Finally, using the INSTRUCTOR

<sup>8</sup>Embeddings are generated using INSTRUCTOR-large model; details here: <https://github.com/xlang-ai/instructor-embedding>

Submission-types	Count
<b>Link:</b> Submission with hyperlinks.	2,232,594
<b>Image:</b> Submission includes image.	2,115,019
<b>Video:</b> Submission includes video.	579,150
<b>Twitter:</b> Submission includes link to Twitter.	419,518
<b>Text:</b> Submission with a text content, excludes all other type of submissions.	947,724

Table 2: Distribution of submission types.

model, we generate embeddings for each post, resulting in approximately 48M embeddings. NB: We also release processed posts with the embeddings.

**Data Release & FAIR Principles.** The data is accessible on Zenodo and on Huggingface. The data is freely downloadable and released in a standardized JSON format for interoperability. We adhere to FAIR (Findable, Accessible, Interoperable, Re-usable) guidelines for data release (Wilkinson et al. 2016), ensuring its discoverability through the digital object identifier (DOI). Additionally, we release the code for getting started with our dataset, which is available on our GitHub. All the links to access resources are given in the introduction of this paper.

## General Characterization

In this section, we present basic statistics regarding our dataset, highlighting users’ posting activities on the platform. For the top 15 communities, we discuss community-wise posting activity and also provide a brief background on their migration pattern across various platforms. Finally, we identify migrant communities on Scored that were previously banned on Reddit.

### Posting Activity

We look at users’ posting activity on Scored by creating longitudinal plot of comments and submissions over time. Figure 3a illustrates the longitudinal trends of submissions and comments, while Figure 3b shows the distribution of submission types (hyperlinks, images, videos, text, and Twitter links) over time. We observe notable peaks in the posting activity between 2020-11-01 and 2021-01-25. This period includes consequential real-world political events, for example, the 2020 U.S. presidential election, the 2021 U.S. Capitol attack, and the inauguration of the 46th U.S. President, Joe Biden.

Users can post media, hyperlinks, or text on Scored, with Table 2 showing the distribution of these submission types. Approximately 35% of submissions are hyperlinks, 42% are media (images or videos), 6% are Twitter links and 15% are text submissions. A longitudinal analysis of various submission types reveal distinct patterns during different time periods. As shown in Figure 3b, at the beginning of 2020, most submissions on the platform are hyperlinks. Particularly, during the U.S. Presidential election (around November 2020), we see a sudden surge in text submissions. Between January 1st, 2021, and January 25th, 2021, text-based submissions clearly dominate over other types, driven by

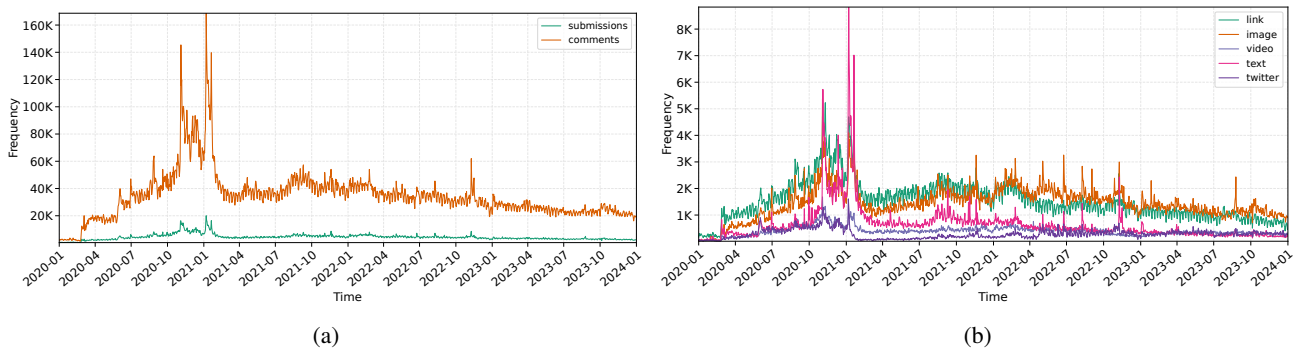


Figure 3: Temporal evolution of daily activity in our dataset: (a) daily number of submissions/comments; and (b) daily number of submissions categorized in different types. X-axis shows daily ticks with a 3-month time interval and Y-axis shows the frequency.

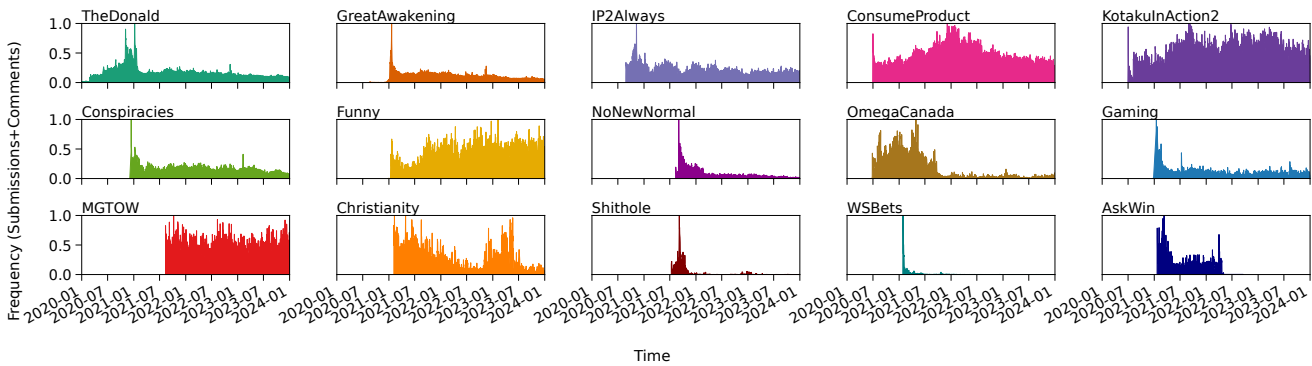


Figure 4: Temporal evolution of daily activity (submissions + comments) in top 15 communities. X-axis shows daily ticks and Y-axis shows the total number of normalized posts (community-wise normalized by dividing max frequency).

discussions sparked by real-world events (e.g., 2021 U.S. Capitol attack).

We analyze community-wise post (i.e., submissions plus comments) distribution. The community with the highest number of posts is *c/TheDonald*, with over 41.74M posts, while approximately 11% of communities (107 out of 976) have only one post (mostly, a single submission). About 75% of communities have 68 or fewer posts, and a mere 3% of communities (30 total) have more than 10K posts.

### Scored Communities

To gain insights per community, we present statistics of the top 15 communities by the total number of posts (Table 3). Notably, 97.02% of the entire Scored user base actively engages within the top 15 communities. All communities, except for *c/TheDonald*, emerged later in the timeline. Moreover, *c/IP2Always* demonstrates a higher posting activity per user than *c/GreatAwakening* and *c/ConsumeProduct*, despite having a smaller user base, with 3.15M posts from 14K users.

Next, we provide a brief background of each community and discuss the longitudinal trend of posting activity within these communities (see Figure 4). Due to space limitations, we only discuss the top 15 communities in depth, highlight-

ing the most interesting details.

***c/TheDonald*.** The largest community on Scored identifies itself as a high-energy rally for supporters of the 45th U.S. President, Donald Trump. Our data collection period begins on January 1st, 2020, allowing us to gather data almost from this community’s inception after migration from Reddit. We discussed the migration history of TheDonald community earlier in the Background section.

***c/GreatAwakening*.** Historically, the Great Awakening originated in the 18th century as a period of religious revival in American Christian history. In recent times, web communities have utilized the QAnon conspiracy theory to promote a new version of the Great Awakening, using phrases like “taking the red pill” to symbolize adopting QAnon awareness. Reddit banned *r/greatawakening* in September 2018, leading to a migration to the Voat platform (Papasavva et al. 2021). Following Voat’s shutdown on December 25th, 2020 (Papasavva and Mariconti 2023), this community potentially migrated to *greatawakening.win*, which co-exists under the ecosystem of *Communities.win* (now Scored). Additionally, Papasavva and Mariconti (Papasavva and Mariconti 2023) confirm that following Voat’s shutdown, users primarily considered migrating to *greatawakening.win* and *Poal*, an alternate platform. As of December 2023, the *c/*

GreatAwakening community on Scored consists of approximately 33K users who have contributed around 6.16M posts.

**c/IP2Always.** IP2Always community engages in watching live streams and creates posts, memes, or short clips from streamed videos. IP refers to “Ice Poseidon,” an In Real Life (IRL) streamer with a notable presence in live streaming.

The IP2Always community was active on Reddit between September 2017 to September 2019. Then, it migrated to SaidIt from September 2019 to August 2020. Later, in August 2020, the community migrated to self-hosted platform, IP2Always.win. Despite facing bans across multiple platforms, including Reddit and SaidIt between 2017 to 2020, c/IP2Always<sup>9</sup> has emerged as the 3rd most active on the Scored after migration. The longitudinal plot of c/IP2Always shows consistent posting activity since its inception, with an earliest post in our dataset dating back to August 26th, 2020, shortly after the ban of /s/Ice\_Poseidon2 on SaidIt<sup>10</sup>.

**c/ConsumeProduct.** Reddit bans ConsumeProduct to remove hate speech-related communities in mid-2020 (Reddit 2020). This community has a controversial history on Reddit, associated with fascist and anti-Semitic ideologies. Within our dataset, the earliest post of c/ConsumeProduct is June 29th, 2020, occurring shortly after the Reddit ban. As of December 2023, c/ConsumeProduct has approximately 18.5K users and about 2.26M posts.

**c/KotakuInAction2.** KotakuInAction2 community identifies itself as a gaming community and actively participates on both Scored and Reddit platforms. r/KotakuInAction is the largest Gamergate community on Reddit and has a controversial history of racism and sexism (Knoop 2018; Cole 2018). On Scored, this community comprises approximately 2.8% of the Scored’s user base.

**c/NoNewNormal.** This community presents itself as a platform for skeptical discussions surrounding the “new normal” resulting from the COVID-19 pandemic. Reddit announced the ban of r/NoNewNormal on September 1st, 2021, mentioning the spread of COVID-19-related misinformation (Security 2021). In our dataset, the first post (August 11th, 2021), hints at the beginning of a migration just 15 days before the Reddit ban. However, c/NoNewNormal shows minimal posting activity approximately after February 23rd, 2022 (see Figure 4).

**c/OmegaCanada.** OmegaCanada has its origins in the formerly known r/metacanada community on the Reddit. This community describes itself as a predominantly conservative community, with many members having faced arbitrary bans on other platforms for their conservative views. Even though there were no apparent ban threats on Reddit, the community migrates from Reddit to omegacanada.win. Our dataset includes posts dating back to June 13th, 2020, and a longitudinal plot shows very high posting activity in the beginning. However, approximately after September 23rd, 2021, we ob-

<sup>9</sup>As of April 1st, 2024, c/IP2Always is accessible through <https://scored.co/c/IP2Always> and <https://IP2alwayswins.com>.

<sup>10</sup>Removal of IP2 sub from SaidIt: [https://saidit.net/s/SaidIt/comments/65k1/we\\_are\\_being\\_forced\\_by\\_our\\_server\\_company\\_to/](https://saidit.net/s/SaidIt/comments/65k1/we_are_being_forced_by_our_server_company_to/).

Community	#Posts	#Users	Min. Date
c/TheDonald	41,745,699	169,520	2020-01-01
c/GreatAwakening	6,161,369	33,008	2020-08-10
c/IP2Always	3,154,741	14,001	2020-08-26
c/ConsumeProduct	2,263,060	18,562	2020-06-29
c/KotakuInAction2	747,215	6,209	2020-06-30
c/Conspiracies	539,164	10,304	2020-12-04
c/Funny	371,081	7,451	2020-12-27
c/NoNewNormal	322,300	7,644	2021-08-11
c/OmegaCanada	249,316	3,546	2020-06-13
c/Gaming	181,469	5,782	2020-12-16
c/MGTOW	175,853	2,006	2021-08-12
c/Christianity	124,866	1,515	2021-02-05
c/Shithole	98,720	1,314	2021-07-12
c/WSBets	66,358	4,427	2021-01-27
c/AskWin	39,308	3,393	2021-01-21
<b>Total</b>	<b>56,240,519</b>	<b>220,013</b>	-

Table 3: Overview of the top 15 communities in our dataset.

serve a substantial decline in posting activity.

**c/MGTOW.** MGTOW (Men Going Their Own Way) believes society is rigged against men, advocating for the abandonment of women and, at times, western society, while also displaying extreme anti-feminist and misogynistic views (Ribeiro et al. 2021a). MGTOW experienced a quarantine on the Reddit platform in January 2021 (Ribeiro et al. 2021a). Within our dataset, the earliest posts date back to August 11th, 2021, emerging after more than six months of the quarantine period. Despite being a small community with only 2K users on Scored, the longitudinal plot shows consistent posting activity.

**c/Conspiracies, c/Funny, c/Gaming, c/Christianity, c/Shithole, c/WSBets, c/AskWin.** These communities on Scored serve as hubs for specific topic discussions, as obvious from their names.

## Migrant Communities

Looking at the history of Scored’s top 15 communities, we see that most migrated from Reddit. Here, we identify Scored communities that once existed on Reddit and were subsequently banned. To do so, first we compile a list of 2,248 Reddit-banned subcommunities by referencing previous research (Phadke, Samory, and Mitra 2022) (NB: Reddit does not provide an official list of such communities). Second, to confirm the banned status of these communities, we crawl each subcommunities’ homepage and verify it through the presence of a Reddit’s ban warning label. Third, we apply the Ratcliff/Obershelp string-matching algorithm to check if a Scored community’s name corresponds to a banned community on Reddit, assuming that community names remain consistent, followed as in prior research (Ali et al. 2021). We find 58 migrant communities with a 0.85 Ratcliff/Obershelp threshold, which accommodates cases where communities change their names. For example, on Reddit, GreatAwakening was known as r/greatawakening, and on Scored, it becomes c/GreatAwakening. Along with dataset release, we also release compiled list



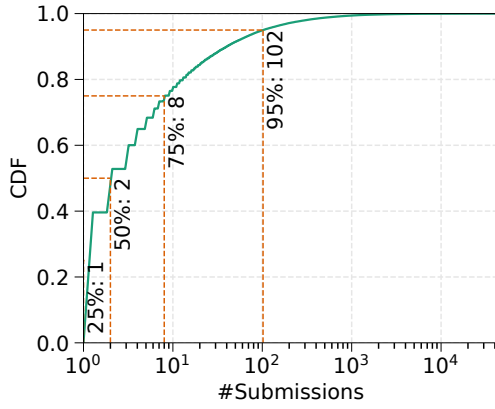


Figure 5: CDF of number of submissions per user

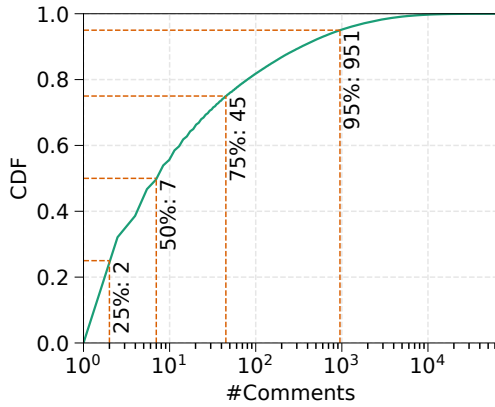


Figure 6: CDF of number of comments per user

of banned Reddit communities and the name of the communities that migrated on Scored, which will help to advance the research related migration movements of banned communities in online space.

### User Analysis

In this section, we look at the influx of new users throughout the dataset collection period and examine the overlap of users within the top 15 communities.

**User-wise Submissions/Comments.** Figure 5 illustrates a Cumulative Distribution Function (CDF) plot showing users' submission frequency on Scored, while Figure 6 displays the CDF plot of users' comment frequency. The CDF plot shows that 95% of Scored users have fewer than 102 submissions and fewer than 1,000 comments. Both distributions of user participation on Scored reveal a long-tail pattern.

**Influx of New Users.** Next, we look at the influx of new users over time, as shown in Figure 7. Interestingly, there are points in time where a substantial number of new users joined the platform, especially during the year 2020 and early 2021. At first, the emergence of Scored attracted new users quickly, with many communities migrating from main-

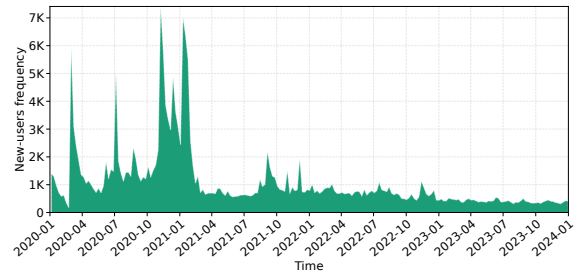


Figure 7: New users over time

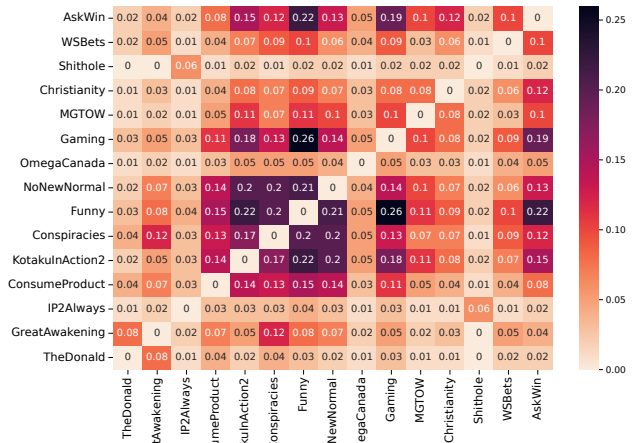


Figure 8: Heatmap of user overlap between communities. Numerical values denote the Jaccard similarity between each pair of communities.

stream platforms. We also observe a surge in new users during The Great Ban event. Another notable event that results in a surge of new users joining Scored is the U.S. election in 2020 and the U.S. Capitol insurrection.

**Inter-community User Overlap.** We create a heatmap to visualize user overlap between the top 15 communities, as depicted in Figure 8. This heatmap is constructed based on common users between pairs of communities, using the Jaccard similarity metric. Notably, *c/Conspiracies* and *c/GreatAwakening* show a user overlap of 0.12 Jaccard value (4,511 unique users in common), while *c/TheDonald* and *c/GreatAwakening* show a user overlap of 0.08 Jaccard value (15,626 unique users in common). Additionally, *c/Funny* overlaps with *c/ConsumeProduct* (3,410 total matches), *c/KotakuInAction2* (2,421 users), *c/Conspiracies* (2,995 users), *c/NoNewNormal* (2,594 users), and *c/Gaming* (2,742 users).

### Scored Web-links Ecosystem

To understand Scored's outgoing connections to the Web, we use regular expressions and extract hyperlinks from all posts. As a result, our dataset contains 6.37M posts (i.e., submission plus comments) with one or more hyperlinks, totaling 7.39M non-unique hyperlinks. Specifically, 4.92M

Community	#Posts	#Links	#Domains
c/TheDonald	36,684,832	9,756,676	101,905
c/GreatAwakening	5,764,811	963,929	45,931
c/IP2Always	2,578,001	556,681	10,743
c/ConsumeProduct	1,973,662	376,741	17,698
c/KotakuInAction2	718,374	110,216	8,168
c/Conspiracies	503,787	125,595	12,300
c/Funny	358,815	81,254	2,969
c/NoNewNormal	302,031	54,502	6,511
c/OmegaCanada	127,041	67,812	5,137
c/Gaming	169,736	15,115	2,135
c/MGTOW	164,717	16,480	2,115
c/Christianity	117,373	24,439	2,306
c/Shithole	60,564	15,372	657
c/WSBets	59,322	8,323	1,165
c/AskWin	33,487	2,474	856

Table 4: Overview of extracted web-Links and unique domains for each community.

hyperlinks are extracted from submissions, while 2.46M hyperlinks are extracted from comments. Then, we extract domains from all 7.39M hyperlinks, resulting in 130K unique domains in our dataset. Additionally, we observe that approximately 98% of domains (approx. 127K out of 130K) have 100 or fewer links.

Table 4 provides community-wise hyperlink posting frequency across top 15 communities. In Table 4, *#posts* refers to the total number of posts with non-empty content (i.e., posts that have not been deleted), while *#links* indicates the number of extracted hyperlinks, and *#domains* denotes the count of unique domains from all extracted links per community. *c/TheDonald* has the highest number of outgoing connections on the Web, with 9.75M hyperlinks posted, linking to 101.9K unique domains. The community with the second-highest number of outgoing connections is *c/GreatAwakening*, with 45.9K domains. *c/OmegaCanada* has the highest ratio (53.38%) of extracted links to the total number of posts, followed by *c/TheDonald* (26.60%) and *c/Conspiracies* (24.93%). *c/MGTOW* displays the lowest ratio (10.01%) of extracted links to total posts. Hyperlink graphs can find semantically similar websites (Hanley, Kumar, and Durumeric 2022), for instance, communities like *GreatAwakening* share connections with other QAnon-related conspiracy websites. In-depth analysis of communities’ outgoing connection can help to understand communities links to the website on broader web and role of these communities to propagate information in and outside the community.

## Related Work

Social networks typically enforce sanctions on users and communities for engaging in antisocial behavior, propagating hate speech, online harassment, misinformation, and conspiracy theories (Trujillo and Cresci 2022; Ribeiro et al. 2021b; Chandrasekharan et al. 2017; Mekacher and Papasavva 2022; Jhaver et al. 2021). As a result, alternative platforms like Gab, Parler, Gettr, Rumble, Poal, and Truth

have emerged in response to mainstream social media platforms. The research community has actively curated publicly available datasets from these alternative social media platforms, emphasizing the necessity of studying them within the broader web ecosystem. Notable contributions include datasets focusing Parler (Aliapoulios et al. 2021a), Gab (Fair and Wesslen 2019), 4chan (Papasavva et al. 2020), Truth Social (Gerard, Botzer, and Weninger 2023), Bitchute (Trujillo et al. 2022), and Telegram (Baumgartner et al. 2020). These datasets have facilitated studies analyzing online coordination in Parler leading to consequential offline events like the Storming of the U.S. Capitol (Jakubik et al. 2023), profiling QAnon supporters on Parler (Bär, Pröllochs, and Feuerriegel 2023), understanding temporal patterns of hate speech in Gab (Mathew et al. 2020), uncovering scams and conspiracy movements on Telegram (La Morgia et al. 2021), and characterizing online incitements to harassment (Aliapoulios et al. 2021b). The utility of these datasets extends beyond the realm of social media behavior analysis, as researchers have leveraged them to identify new cyber threats like zoombombing (Ling et al. 2021), assess toxic behavior in online chatbots (Si et al. 2022), and automatically detect hateful content (González-Pizarro and Zannettou 2023). Henceforth, we hope the public release and availability of Scored dataset will further improve our understanding of alternative platforms, why users migrate to these platforms, and help build better tools to identify and detect problematic content on the internet.

## Conclusion

Over the past few years, there has been a tendency for many online communities, for example, *r/The.Donald* and *r/greatawakening*, to migrate to alternative platforms following bans. Researchers have characterized alternative platforms as problematic due to their propagation of misinformation and influence on offline events. In this effort, we collect approximately 57M posts spanning four years (2020 to 2023) from an alternative Reddit-like social platform, Scored. Our dataset includes over 950 communities created on Scored, from which we identified 58 communities that migrated from the Reddit ban. Additionally, we generate and release the sentence embeddings for all the posts using a state-of-the-art model. Furthermore, we provide a toolkit to facilitate researchers in getting started with our dataset.

Our dataset serves as a valuable resource for researchers interested in studying alternative platforms, platform migration dynamics, and fringe web communities. In conjunction with the dataset release, we present basic statistics about banned communities and discuss the migration history of many communities across various platforms. Leveraging this dataset can further advance research questions about web communities, e.g., do these migrant communities exploit Scored to discuss conspiracies or misinformation? what was the role played by these communities in disseminating information during political events like the 2020 U.S. election and the U.S. Capitol insurrection? And how does the web-link ecosystem of migrant communities look on the broader web? Additionally, researchers can measure changes in overall toxicity or the spread of hate speech and



misinformation over time by investigating pre- and post-migration activity and discussions. Moreover, generated embeddings are useful for seamless topic generation or clustering techniques, or training algorithms in natural language processing, ultimately aiding in characterizing these fringe communities.

We hope that this dataset will help in understanding web communities' migration dynamics, their role in disseminating misinformation, conspiracies, or narratives, and the development of models capable of detecting hate speech propagated in these communities.

## Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CNS-2247868 and CNS-2247867.

## References

- Ali, S.; Saeed, M. H.; Aldreabi, E.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Understanding the Effect of Deplatforming on Social Networks. In *13th ACM Web Science Conference 2021*, 187–195. ISBN 978-1-4503-8330-1.
- Aliapoulos, M.; Bevensee, E.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; and Zannettou, S. 2021a. A Large Open Dataset from the Parler Social Network. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1): 943–951.
- Aliapoulos, M.; Take, K.; Ramakrishna, P.; Borkan, D.; Goldberg, B.; Sorensen, J.; Turner, A.; Greenstadt, R.; Lauinger, T.; and McCoy, D. 2021b. A large-scale characterization of online incitements to harassment across platforms. In *Proceedings of the 21st ACM Internet Measurement Conference*, 621–638.
- Bacı, U.; Sirivianos, M.; and Blackburn, J. 2023. A Data-driven Understanding of Left-Wing Extremists on Social Media. *arXiv preprint arXiv:2307.06981*.
- Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023. Finding Qs: Profiling QAnon supporters on Parler. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 34–46.
- Baumgartner, J.; Zannettou, S.; Squire, M.; and Blackburn, J. 2020. The pushshift telegram dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 840–847.
- Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW).
- Cima, L.; Larios, A. T.; Avvenuti, M.; and Cresci, S. 2024. The Great Ban: Efficacy and Unintended Consequences of a Massive Deplatforming Operation on Reddit. *arXiv preprint arXiv:2401.11254*.
- Cole, S. 2018. The Creator of the Largest Gamergate Subreddit Rage Quits, Says it's 'Infested With Racism and Sexism'. <https://www.vice.com/en/article/43pv3g/gamergate-subreddit-kotaku-in-action-shut-down-by-creator>. Accessed: 2024-01-15.
- Fair, G.; and Wesslen, R. 2019. Shouting into the void: A database of the alternative social media platform gab. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 608–610.
- FORCE11. 2020. The FAIR Data principles. <https://force11.org/info/the-fair-data-principles/>. Accessed: 2024-01-15.
- Geburu, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- Gerard, P.; Botzer, N.; and Weninger, T. 2023. Truth Social Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 1034–1040.
- González-Pizarro, F.; and Zannettou, S. 2023. Understanding and detecting hateful content using contrastive learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 257–268.
- Hanley, H. W.; Kumar, D.; and Durumeric, Z. 2022. No calm in the storm: Investigating qanon website relationships. In *Proceedings of the international AAAI conference on Web and social media*, volume 16, 299–310.
- Jakubik, J.; Vössing, M.; Pröllochs, N.; Bär, D.; and Feuerriegel, S. 2023. Online emotions during the storming of the US Capitol: evidence from the social media network Parler. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 423–434.
- Jhaver, S.; Boylston, C.; Yang, D.; and Bruckman, A. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2).
- Knoop, J. 2018. Reddit brought back a Gamergate forum after its creator tried to destroy it. <https://www.dailydot.com/unclick/reddit-kotakuinaction/>. Accessed: 2024-01-15.
- La Morgia, M.; Mei, A.; Mongardini, A. M.; and Wu, J. 2021. Uncovering the dark side of Telegram: Fakes, clones, scams, and conspiracy movements. *arXiv preprint arXiv:2111.13530*.
- Ling, C.; Bacı, U.; Blackburn, J.; and Stringhini, G. 2021. A first look at zoombombing. In *2021 IEEE symposium on security and privacy (SP)*, 1452–1467. IEEE.
- Mathew, B.; Illendula, A.; Saha, P.; Sarkar, S.; Goyal, P.; and Mukherjee, A. 2020. Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–24.
- McIlroy-Young, R.; and Anderson, A. 2019. From “welcome new gabbers” to the pittsburgh synagogue shooting: The evolution of gab. In *Proceedings of the international aai conference on web and social media*, volume 13, 651–654.
- Mekacher, A.; and Papasavva, A. 2022. ” I Can't Keep It Up.” A Dataset from the Defunct Voat. co News Aggregator. In *AAAI ICWSM*, volume 16.
- Newton, C. 2020. Reddit bans r/The\_Donald and r/Chapo-TrapHouse as part of a major expansion of its rules. <https://www.theverge.com/2020/6/29/21304947/reddit->

- ban-subreddits-the-donald-chapo-trap-house-new-content-policy-rules. Accessed 2024Y01M15DD.
- Ng, L. H. X.; Moffitt, J.; and Carley, K. M. 2022. Coordinated through a web of images: Analysis of image-based influence operations from china, iran, russia, and venezuela. *arXiv preprint arXiv:2206.03576*.
- Papasavva, A.; Blackburn, J.; Stringhini, G.; Zannettou, S.; and Cristofaro, E. D. 2021. "Is It a Coincidence?": An Exploratory Study of QAnon on Voat. In *Proceedings of the Web Conference 2021, WWW '21*, 460–471. ISBN 978-1-4503-8312-7.
- Papasavva, A.; and Mariconti, E. 2023. Waiting for Q: An Exploration of QAnon Users' Online Migration to Poal in the Wake of Voat's Demise. *arXiv preprint arXiv:2302.01397*.
- 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 *Proceedings of the international AAAI conference on web and social media*, volume 14, 885–894.
- Paudel, P.; Blackburn, J.; Cristofaro, E. D.; Zannettou, S.; and Stringhini, G. 2022. A Longitudinal Study of the Gettr Social Network. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*.
- Phadke, S.; Samory, M.; and Mitra, T. 2022. Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 770–781.
- Porter, C. E. 2004. A typology of virtual communities: A multi-disciplinary foundation for future research. *Journal of computer-mediated communication*, 10(1): JCMC1011.
- Reddit. 2020. Update to Our Content Policy. [https://www.reddit.com/r/announcements/comments/h13oht/update\\_to\\_our\\_content\\_policy/](https://www.reddit.com/r/announcements/comments/h13oht/update_to_our_content_policy/). Accessed: 2024-01-15.
- Ribeiro, M. H.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; Long, S.; Greenberg, S.; and Zannettou, S. 2021a. The evolution of the manosphere across the web. In *Proceedings of the international AAAI conference on web and social media*, volume 15, 196–207.
- Ribeiro, M. H.; Jhaver, S.; Zannettou, S.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; and West, R. 2021b. Do Platform Migrations Compromise Content Moderation? Evidence from r/The\_Donald and r/Incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–24.
- Rivers, C. M.; and Lewis, B. L. 2014. Ethical Research Standards in a World of Big Data. *F1000Research*, 3: 38.
- Security, R. 2021. COVID denialism and policy clarifications. [https://www.reddit.com/r/redditsecurity/comments/pfyqqn/covid\\_denialism\\_and\\_policy\\_clarifications/](https://www.reddit.com/r/redditsecurity/comments/pfyqqn/covid_denialism_and_policy_clarifications/). Accessed: 2024-01-15.
- Shearer, E. 2018. Social media outpaces print newspapers in the US as a news source. <https://www.pewresearch.org/short-reads/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/>. Accessed: 2024-01-15.
- Si, W. M.; Backes, M.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; Zannettou, S.; and Zhang, Y. 2022. Why so toxic? measuring and triggering toxic behavior in open-domain chatbots. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2659–2673.
- Starbird, K.; Arif, A.; and Wilson, T. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–26.
- Su, H.; Shi, W.; Kasai, J.; Wang, Y.; Hu, Y.; Ostendorf, M.; Yih, W.-t.; Smith, N. A.; Zettlemoyer, L.; and Yu, T. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.
- Trujillo, A.; and Cresci, S. 2022. Make reddit great again: assessing community effects of moderation interventions on r/the\_donald. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2).
- Trujillo, M.; Gruppi, M.; Buntain, C.; and Horne, B. D. 2020. What Is BitChute? Characterizing the "Free Speech" Alternative to YouTube. [arxiv:2004.01984](https://arxiv.org/abs/2004.01984).
- Trujillo, M. Z.; Gruppi, M.; Buntain, C.; and Horne, B. D. 2022. The MeLa BitChute Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1342–1351.
- Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.
- Zannettou, S.; Bradlyn, B.; De Cristofaro, E.; Kwak, H.; Sirivianos, M.; Stringini, G.; and Blackburn, J. 2018. What Is Gab: A Bastion of Free Speech or an Alt-Right Echo Chamber. In *Companion Proceedings of the the Web Conference 2018, WWW '18*, 1007–1014. ISBN 978-1-4503-5640-4.
- Zannettou, S.; Sirivianos, M.; Blackburn, J.; and Kourtellis, N. 2019. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. *Journal of Data and Information Quality*, 11(3): 1–37.
- Ziems, C.; He, B.; Soni, S.; and Kumar, S. 2020. *Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis*. Georgia Institute of Technology.

## Paper Checklist

- For most authors...
  - Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes**
  - Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? **Yes**
  - Do you clarify how the proposed methodological approach is appropriate for the claims made? **NA**

- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? *NA*
  - (e) Did you describe the limitations of your work? *Yes, please see Section “Limitations.”*
  - (f) Did you discuss any potential negative societal impacts of your work? *NA*
  - (g) Did you discuss any potential misuse of your work? *NA*
  - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? *Yes, please see Section “Data Release & FAIR Principles.”*
  - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? *Yes*
2. Additionally, if your study involves hypotheses testing...
- (a) Did you clearly state the assumptions underlying all theoretical results? *NA*
  - (b) Have you provided justifications for all theoretical results? *NA*
  - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? *NA*
  - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? *NA*
  - (e) Did you address potential biases or limitations in your theoretical framework? *NA*
  - (f) Have you related your theoretical results to the existing literature in social science? *NA*
  - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? *NA*
3. Additionally, if you are including theoretical proofs...
- (a) Did you state the full set of assumptions of all theoretical results? *NA*
  - (b) Did you include complete proofs of all theoretical results? *NA*
4. Additionally, if you ran machine learning experiments...
- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? *Yes, we make data available to reproduce the results, please see Section “Data Release & FAIR Principles.”*
  - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? *Yes, please see Section “Data Embeddings.”*
  - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? *NA*
  - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? *Please see Section “Data Embeddings.”*
  - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? *NA*
  - (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? *NA*
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
- (a) If your work uses existing assets, did you cite the creators? *Yes, please see Section “Data Embeddings.”*
  - (b) Did you mention the license of the assets? *Yes, please see Section “Data Embeddings.”*
  - (c) Did you include any new assets in the supplemental material or as a URL? *Yes, please see Section “Data Release & FAIR Principles.”*
  - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? *No. We only use publicly available data for our work. Please see Section “Ethical Considerations.”*
  - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? *NA*
  - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? *Yes, please see Section “Data Release & FAIR Principles.”*
  - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? *No. We make all the details available in this paper. Please see Section “Dataset Structure.”*
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
- (a) Did you include the full text of instructions given to participants and screenshots? *NA*
  - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? *NA*
  - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? *NA*
  - (d) Did you discuss how data is stored, shared, and de-identified? *NA*