

The MeLa BitChute Dataset

Milo Z. Trujillo,¹ Maurício Gruppi,² Cody Buntain,³ and Benjamin D. Horne⁴

¹ Vermont Complex System Center, University of Vermont, Burlington, VT USA

² Computer Science, Rensselaer Polytechnic Institute, Troy, NY USA

³ College of Information Studies, University of Maryland, College Park, MD USA

⁴ School of Information Sciences, University of Tennessee Knoxville, Knoxville, TN, USA
milo.trujillo@uvm.edu, gouvem@rpi.edu, cbuntain@umd.edu, bhorne6@utk.edu

Abstract

In this paper we present a near-complete dataset of over 3M videos from 61K channels over 2.5 years (June 2019 to December 2021) from the social video hosting platform BitChute, a commonly used alternative to YouTube. Additionally, we include a variety of video-level metadata, including comments, channel descriptions, and views for each video. The MeLa-BitChute dataset can be found at: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KRD1VS>.

1 Introduction

Wilson and Starbird (2021) define “alt-tech” spaces as alternative, non-mainstream platforms that exist largely in reaction to perceived risks of censorship in mainstream spaces. These alt-tech platforms have received significant attention from researchers, practitioners, and even policy makers, due to their role in producing, spreading, and conserving anti-social content. This anti-social content ranges from political disinformation to health-related conspiracy theories to violent hate speech.

To this end, researchers have striven to study alt-tech platforms and to build large datasets around those platforms; e.g., Gab (Zannettou et al. 2018), Gettr (Paudel et al. 2021), Parler (Aliapoulos et al. 2021), Dissenter (Rye, Blackburn, and Beverly 2020), and Telegram (Júnior et al. 2021). However, one platform in this ecosystem that has, for the most part, lacked study is BitChute, an alternative to YouTube. Studies on BitChute are rare because data collected from the platform is rare, as BitChute does not have a publicly available API like other social media platforms. This limitation makes data collection a significant hurdle for researchers.

Despite this difficulty, the platform is deserving of study. Just as other major alt-tech platforms, BitChute plays a critical role in harboring anti-social content and communities (Trujillo et al. 2020). Most famously, BitChute was a safe haven for the viral, COVID-19 conspiracy theory film, *Plandemic*, which was quickly removed from Facebook, YouTube, and Twitter (Kearney, Chiang, and Massey 2020; Buntain et al. 2021a). As shown in Rogers (2020), these alternative spaces do not exist in isolation; rather, users

of and audiences in alt-tech spaces operate across several of these platforms simultaneously. Hence, these spaces need to be studied holistically to understand how extremists and far-right individuals leverage the many affordances available across these alt-tech platforms. This need is apparent, as Doesburg (2021) demonstrates that BitChute is already one of the most popular alt-tech domains shared in Telegram, and Rogers (2020) shows that BitChute is one of the main destinations for internet celebrities who have been de-platformed on the mainstream platforms. As the broader information ecosystem cannot be fully understood without YouTube, so too must one understand BitChute’s role to understand the alt-tech ecosystem.

In this paper, we present the MeLa-BitChute dataset to help fill this gap. The dataset contains data from 3,036,190 videos, 61,229 channels, and 11,434,571 comments between June 28th, 2019 and December 31st, 2021. This dataset provides timestamped activities and estimates on views for the majority of channels and videos on the platform, allowing researchers to align BitChute videos with behavior on other platforms. Therefore, this dataset can facilitate both studies of BitChute in isolation and studies of BitChute’s role in the larger ecosystem.

In the remainder of this paper, we describe the data collection methodology behind MeLa-BitChute, publicly-available data formats and documentation, evaluations of the dataset’s completeness, and an extensive discussion on use cases. While we do discuss some characteristics of BitChute in this paper, we recommend reading Trujillo et al. (2020) for a more complete examination of the platform’s history and early content.

2 Data Collection and Infrastructure

BitChute does not have a published API. Therefore, we build a custom collection engine that uses web-scraping and parsing at regular intervals. Broadly, using this infrastructure, we collect data on four entities within BitChute: videos, video views, comments, and channels. Figure 1 shows the high-level flow chart of this collection infrastructure.

2.1 Videos

The core part of the data collection is video metadata. To collect video data, we utilize a web-page on BitChute that displays newly uploaded videos in a stack (see Figure 1 for an

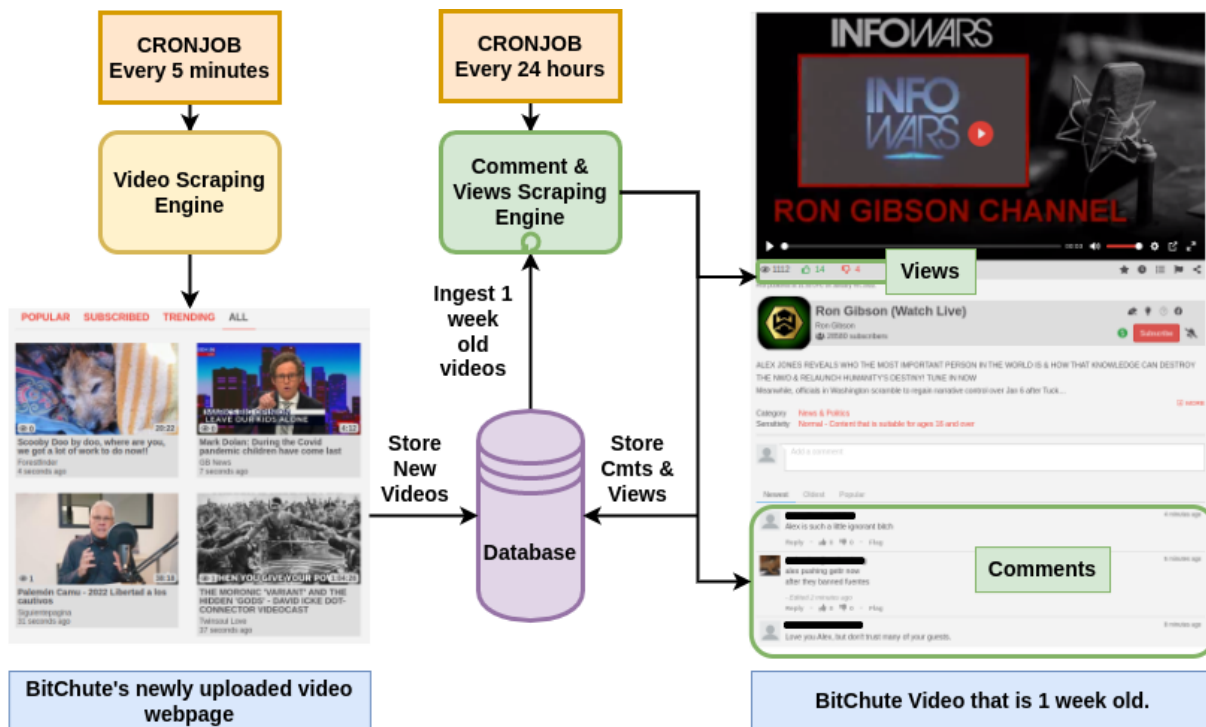


Figure 1: The core data collection is done in two parallel parts: 1. Every 5 minutes we collect newly uploaded videos from the BitChute video stack. 2. Every 24, hours we loop through all videos that are at least one week old, go to the video page, and scrape views- and comments data. The third component, not depicted here, is the collection of channel description, which is scraped once for every new channel in the database.

example of this page). This page is scraped every five minutes, recording the URLs of videos we have not yet seen. For each newly-uploaded video, we visit the video URL, parsing the web interface to extract the video title, description, uploader, channel, video category, sensitivity rating, and exact upload date. We record these results in a PostgreSQL database. An example of the video interface that data is scraped from can be found in Figure 3.

2.2 Comments and Views

Since we collect videos within 5 minutes of being uploaded to the platform, we must wait to collect audience engagement data. To this end, we have a second, concurrent process, which queries the database for a list of videos uploaded at least one week ago that have not yet been re-examined. This script then visits each video URL again, this time recording the number of views and scraping all of the comments that the video has received. If the video is no longer available, we record whether it has been retracted by the uploader or removed by BitChute moderation.

When scraping the comments, we collect the comment author, comment text, creation date, and scrape date.

2.3 Channels

Lastly, to collect description data on the channels, we have a third concurrent process that queries the database for a list of channels that have yet to have their description information

collected. This script then visits each channel 'About' page to scrape the HTML for an author-provided description, if available.

2.4 Collecting Dynamically Loaded Data

While most of the metadata we are interested in is available through HTML parsing, the view counts and comments are loaded dynamically. We initially overcame this by automating a web browser with Selenium, so that BitChute's JavaScript could run without modification. Using this method, we could extract the view count and comments from the browser's DOM. However, automating a browser scaled poorly, so we instead determined what HTTP requests the site JavaScript was making, and automated those requests ourselves to retrieve views and comments.

Collecting Dynamically Loaded Views First, to retrieve views, we visit the video URL and record a CSRF token¹. Then when visiting `{video-URL}/counts/`, we present the CSRF token and collect the loaded views data.

Collecting Dynamically Loaded Comments Second, to retrieve comments, we must rely on BitChute's comment infrastructure, which changed throughout our collection timeline. Originally, from 2017 to around September 2020,

¹Cross-Site Request Forgery Token, a type of cookie typically used to prevent attackers from tricking a browser into making malicious HTTP requests

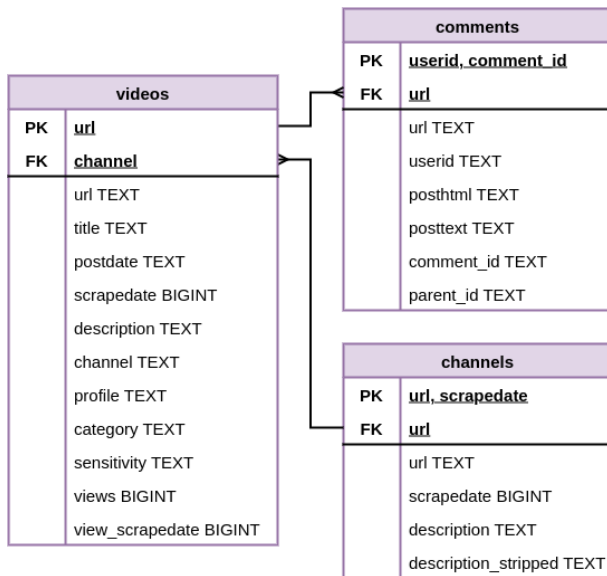


Figure 2: Database schema for SQLite3 format. The video URLs are used as the primary key for the videos table, while a composite key of the commenter user ID and the individual comment ID are used as the primary key of the comments table, and a composite key of the channel URL and the timestamp the channel description was scraped are used as the primary key of the channels table.

BitChute contracted the third-party blog comment hosting service, Disqus, to provide commenting infrastructure for their website. After Disqus terminated the contract, BitChute implemented their own commenting software, “Comment-Freely.” CommentFreely is open source, and can be examined at <https://github.com/BitChute/commentfreely>.

When Disqus was being used, we derived the Disqus URL from the BitChute URL, then visited the Disqus page, which included comments as both HTML and JSON for easy parsing. Once BitChute switched to their own CommentFreely infrastructure, we retrieved comments as follows: 1. Parse the CommentFreely JavaScript embedded in the video page to find a unique “video token.” 2. Make a POST request to https://commentfreely.bitchute.com/api/get_comments/ containing the video token. 3. Parse the response JSON.

3 Publicly Available Data Formats

In order to accommodate the largest audience possible, we provide two widely-used data formats.

3.1 SQLite3 Database

The first format is an SQLite3 database with three tables: videos, comments, and channels. The schema for this database can be found in Figure 2. While our collection engine stores data in a PostgreSQL database, we convert it to an SQLite3 database to allow researchers to use the data without any database server setup.

The primary table in the database is the videos table, which includes the video URL, title, date the video was

posted, timestamp when video was scraped, a description of the video if provided by the uploader, the profile of the user who posted the video, the channel that posted the video (see Appendix Table 1 for description of profile vs. channel), the user picked topical category for the video, the user picked sensitivity for the video, the number of times the video is viewed, and a timestamp of when the views were scraped.

The second table is the comments table, which contains the video URL that the comment is under, the user ID of the commenter, the comment ID for the individual comment, the parent comment ID if it is a nested comment, the comment HTML, and the comment text.

The third table is the channels table, which contains the URL to the channel, a the HTML of the channel description if the channel owner provides one, the text of the channel description if the channel owner provides one, and a timestamp of when the channel data was scraped.

A detailed description of each data column can be found in Appendix A.

3.2 CSV

The second format in which we provide the dataset is a set of Comma-Separated Value (CSV) files. We provide three CSV files, one for each table in the database: videos, comments, and channels. The columns in each CSV file are the same as the columns in each corresponding SQLite3 database table.

3.3 FAIR Principles

We are careful to ensure that the MeLa-BitChute dataset follows FAIR principles².

- Findable - The dataset is persistently stored on Harvard Dataverse, is documented, and described with rich metadata.
- Accessible - The dataset is freely and publicly accessible through Harvard Dataverse’s GUI, is stored in two widely-used formats, and comes with an example Python script for data extraction and use.
- Interoperable - The dataset can be parsed automatically using standard languages (Python, SQL, R), and can be parsed by human annotators using the CSV formats provided.
- Re-usable - The data can be reused for many types of studies, given the breadth of the collection. The data can be paired and augmented with other social media datasets for rich studies of alt-tech, deplatforming, disinformation spread and more. See Section 5 for an in-depth discussion of these use cases. Given that the URLs are stored and rich metadata is well-documented, provenance is maintained.

3.4 Ethical Considerations of Stored Data

The key ethical consideration with the collection and sharing of this dataset is the privacy of the platform’s producers and consumers. While we considered anonymizing the content producers (channels/profiles), we choose not to as their identity is important to understand both activity within

²<https://www.force11.org/group/fairgroup/fairprinciples>

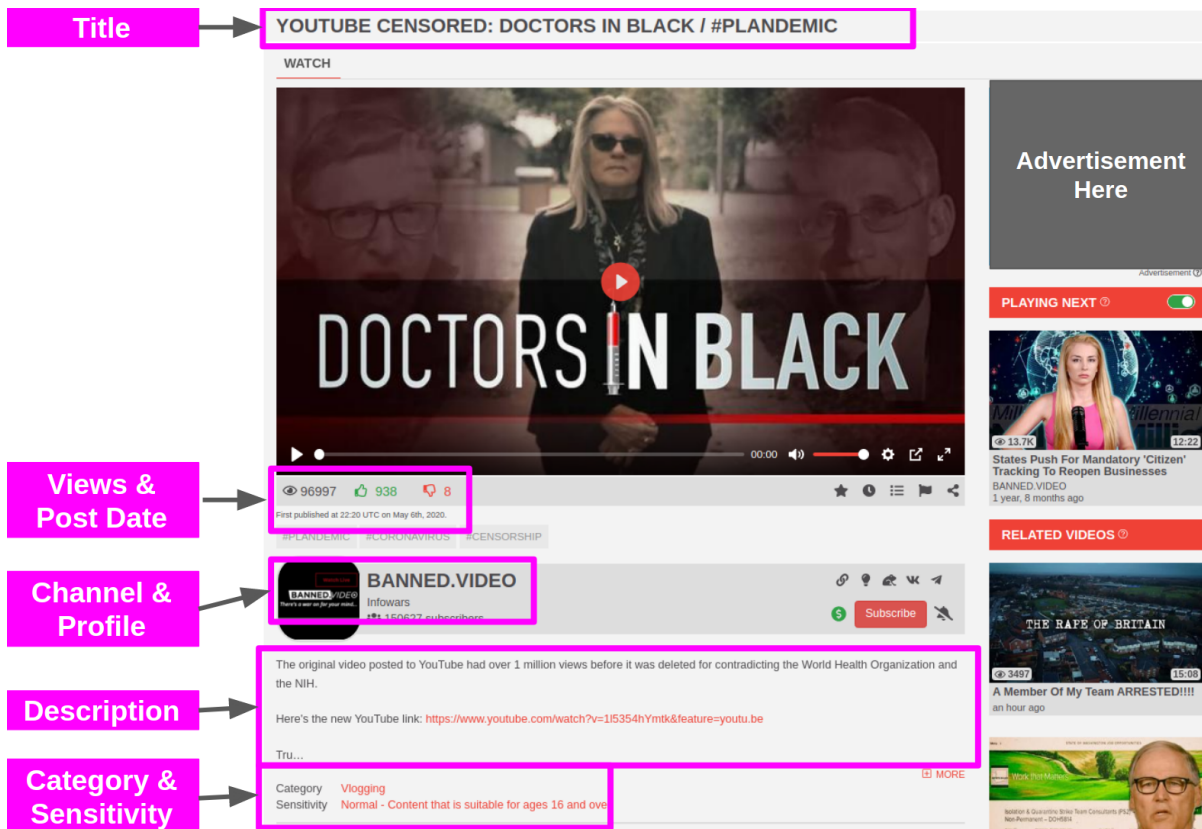


Figure 3: Example of metadata collected for each video. Note, the channel of the video is “BANNED.VIDEO” while the profile is “Infowars.” The Infowars profile has several channels, including an Infowars channel and “The Alex Jones Show.”

the platform and studies across other platforms. Censoring the channel and video URLs would reduce the dataset’s usefulness in the research community and would destroy the provenance of the data. Furthermore, since the content producers on BitChute post videos for public consumption, they should not have the expectation of anonymity.

On the other hand, we do choose to anonymize the comment posters. We have assigned each commenter a unique ID, by creating a salted hash of their account information. This allows researchers to identify all comments made by a particular author, without revealing the username of the author. We do this because commenters have a greater expectation of privacy than content publishers, and the right to be forgotten. Furthermore, the importance of tracking influential content creators across multiple platforms does not apply to commenters.

4 Evaluation of Data Completeness

Given the complexity of collecting this data, it is important to clearly document where we are confident in the data completeness and where we are not. To this end, below we discuss several known data outages and caveats.

Known collect server outages During the collection of this data, we documented several collection outages due to either BitChute itself or due to issues at our server location,

which changed several time during the timeline. Below is a list of documented shutdowns of our data collection:

- October 11th, 2019 - Our collection server’s IP address was blocked from accessing BitChute.
- October 23rd, 2019 and November 18th, 2019 - Our collection server lost connectivity during the Pacific Gas & Electric preemptive power shutdowns.³ Our sever was located in California at the time.
- November 22nd, 2019 - BitChute itself was down. The cause is unknown.
- November 3rd, 2020 - BitChute itself was down due to one of their service providers cutting their account off.
- May 22nd, 2021 to May 24th, 2021 - A power outage in the block of the University of Tennessee Knoxville’s campus where our server was located. The cause of the outage was a mouse entering switch gear at the iconic American football stadium, Neyland Stadium.
- June 12th, 2021 to June 14th, 2021 - Another power outage in the block of the University of Tennessee Knoxville’s campus where our server was located.

In most of these cases, we were able to recover videos published during the outages, but we cannot guarantee we

³www.latimes.com/california/story/2019-11-18/another-power-outage-pge-may-shut-down-grid-northern-california

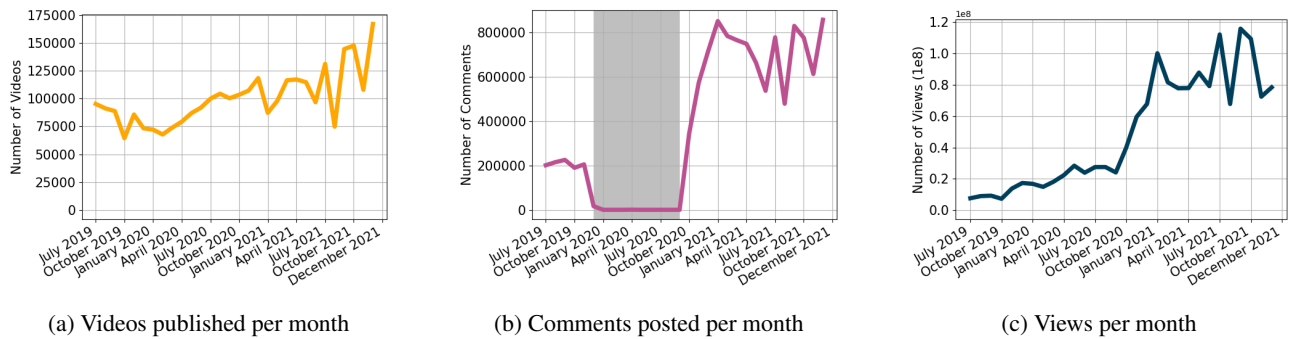


Figure 4: In (a) we show the number of videos published each month. In (b) we show the number of comments posted in each month. Note, the grey background indicates dates where we were not able to collect comments. Please see discussion in Section 4. In (c) we show the number of video views per month. Note both the growth in videos and views collected per month (a and c) and the growth in Google Trends interest in BitChute in Figure 5.

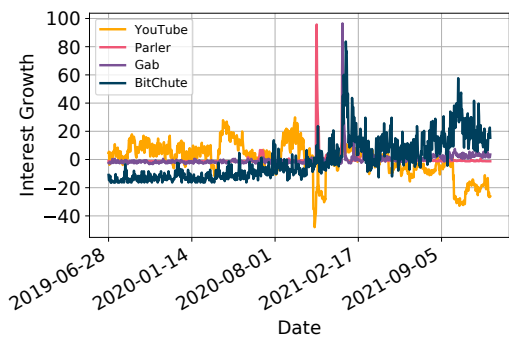


Figure 5: Google Trends interest across the dataset’s timespan, where the mean is shifted to 0 for comparison.

recovered them all. Given how short these outages were, we are confident that the number of videos collected over-time is near-complete.

Known missing comment data As discussed in Section 2.4, BitChute’s comment infrastructure changed throughout our collection timeline. Since our collection method is highly dependent on the structure of BitChute’s platform, these changes directly impacted our ability to collect comments. Originally, BitChute used third-party Disqus, then created their own called CommentFreely after Disqus terminated its contract with BitChute. During this transition, multiple structural changes by BitChute made it impossible to collect comments. Furthermore, we were unable to recover comment data because the comments made during the Disqus era were removed by BitChute and no longer exist. In Figure 4b, dates where comment data is missing due to this are shown in grey. Based on the average number of comments per video, we estimate that we missed at maximum 8.8M comments. This estimate is likely an overestimate given the platform growth during this time-span. In total, 22.9% (697K) of the videos collected were published during the time-span where comments could not be collected. Therefore, we are confident in conclusions being

drawn from the temporal patterns of the videos throughout the collection timeframe but not in conclusions drawn from the temporal patterns of comments during 2020.

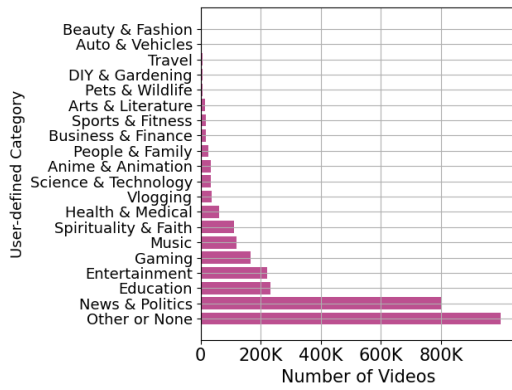
Caveats on views and comments data Again, due to BitChute’s multiple commenting systems, and each system having a different comment ID format. Comment IDs should be assumed to be text, not numeric, and may be NULL. Author IDs cannot link authors from before and after the Disqus to CommentFreely change.

Unfortunately, due to the software changes on BitChute, some of the views and comments are gathered later than one week after video publication. Since our comment data includes the creation date of each comment, researchers can choose to filter out comments more than a week newer than their corresponding videos to obtain a consistent “one week later” view of the dataset.

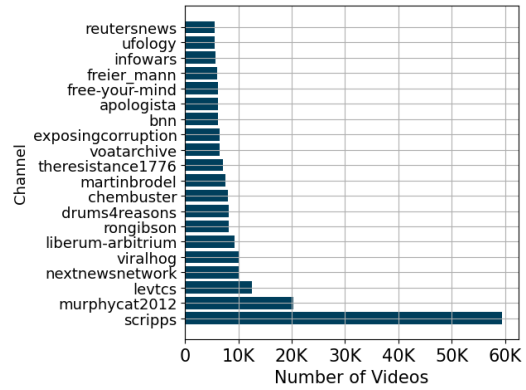
5 Dataset Use Cases

As one of the main video-sharing platforms in the growing alt-tech ecosystem, researchers are increasingly interested in BitChute and its users as lenses for studying these alt-tech spaces. This interest is evident in Google Scholar data, where the number of articles mentioning BitChute has doubled annually over the past two years. Despite this interest, limited data resources exist to support these studies. Given the limited availability of such data, the MeLa-BitChute dataset can facilitate such studies in a multitude of ways, which we outline below.

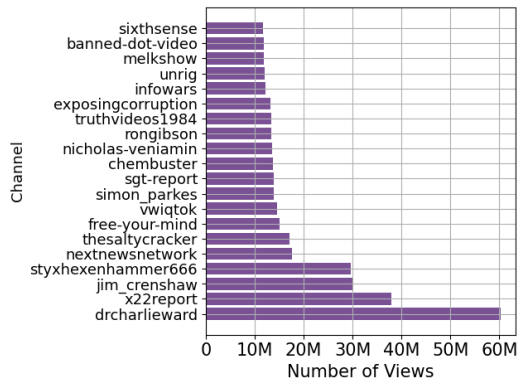
Content Moderation and Deplatforming As social media spaces have become central sources of information despite the prevalence of misinformation and efforts to manipulate audiences (Tucker et al. 2018), the moderation tools and interventions used to ensure the safety of online audiences have likewise become critical aspects of the information ecosystem. “Deplatforming”, or removing/suppressing content or individuals from an online platform, has emerged as a popular moderation intervention across both the mainstream social media platforms (e.g., Twitter, Facebook, and Reddit) (Van Dijck, de Winkel, and Schäfer 2021; Rogers



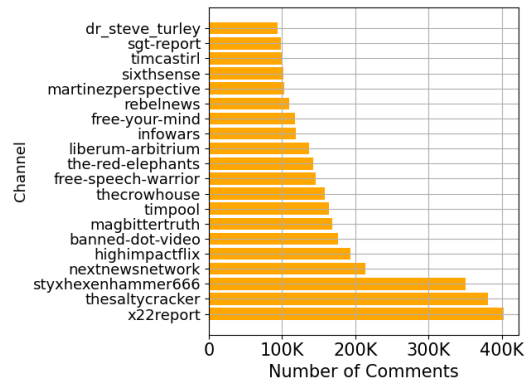
(a) Top 20 user-defined categories by number of videos



(b) Top 20 channels by number of videos



(c) Top 20 channels by number of views



(d) Top 20 channels by number of comments

Figure 6: In (a) we show the top 20 user-picked video categories by number of videos. Note, just as described in (Trujillo et al. 2020), ‘Other’ was originally the default category for all uploaded videos. However, in September 2020 BitChute changed the ‘Other’ category to ‘None’. We show the combined number of videos across those two categories. In (b) we show the top 20 channels by number of videos, in (c) by number of views, and in (d) by number of comments.

2020) and infrastructure platforms (e.g., Amazon Web Services and Google/Apple app stores) (Buckley and Schafer 2021). Assessing impact of these interventions can no longer be done within the context of a single platform, however, as the proliferation of alt-tech platforms has provided new spaces for creators to circumvent moderation. Of particular concern is whether deplatforming pushes individuals from mainstream spaces to these more extreme spaces, thereby increasing exposure to toxic and extreme content in spaces like BitChute.

Researchers have already begun investigations into the cross-platform effects of deplatforming, especially between YouTube and BitChute: Buntain et al. (2021a) examines YouTube’s de-recommendation strategy, finding that removing recommendations to misinforming content appears to suppress its sharing on Twitter and Reddit, but little is known about whether this change simply moves that de-recommended content to alt-tech spaces like BitChute. Similarly, in the aftermath of the “Great Deplatforming” around the January 6 attack on the US Capitol⁴, Buntain et al.

(2021c) shows interest in alt-tech platforms like Gab and BitChute has increased. Rauchfleisch and Kaiser (2021), on the other hand, show that, only about 20% of far-right channels on YouTube that were deplatformed between 2018-2019 had a BitChute presence at the end of 2019. The ecosystem has evolved since these studies, however, and the MeLa-BitChute dataset enables these and new studies of whether and how BitChute creators and audiences respond to these deplatforming efforts. Such questions include whether creators push new content to their BitChute channels or whether BitChute audiences grow and videos are seen or shared more in the aftermath of deplatforming – as suggested with the Plandemic video (Bellemare, Nicholson, and Ho 2020; Buntain et al. 2021a).

As a simple example of the platform’s evolution since these studies, one can compare the popular channels in Figures 6c and 6d (across 2019-2021) to those in 2019, as documented in Trujillo et al. (2020). We see some overlap in the channels with high engagement from late-2019 (*infowars*, *rongibson*, *x22report*, *styxhexenhammer666*, *nextnewsnet-*

⁴[https://www.npr.org/2022/01/06/1070763913/kicked-off-](https://www.npr.org/2022/01/06/1070763913/kicked-off-facebook-and-twitter-far-right-groups-lose-online-clout)

[facebook-and-twitter-far-right-groups-lose-online-clout](https://www.npr.org/2022/01/06/1070763913/kicked-off-facebook-and-twitter-far-right-groups-lose-online-clout)

work), but new channels have emerged as highly engaged in this dataset, such as *drcharlieward*, *chembuster*, and *sixth-sense*. Many of these new top channels are focused on health and vaccination conspiracy theories, where as the top channels in late-2019 were mostly focused on far-right politics.

Conspiracy and Political Misinformation Like other alt-tech spaces, BitChute has a high proportion of political content, much of which focuses on conspiracy theories and otherwise politically extreme (Trujillo et al. 2020). These alt-tech spaces are especially popular among far-right audiences, who often use these platforms to share and amplify right-wing conspiracies and misinformation (Freelon, Marwick, and Kreiss 2020). For example, despite moderation action by the mainstream platforms to suppress election-related misinformation in the aftermath of the January 6 attack, such conspiracies have flourished in BitChute⁵. The QAnon constellation of conspiracy theories is similarly popular on BitChute (Trujillo et al. 2020), as early research is already examining QAnon’s use of the platform for sharing its increasingly popular messages (Forberg 2021). The MeLa-BitChute dataset is valuable in this space, as much of the content contained therein would be removed or suppressed on other platforms, and the alt-tech spaces available are primarily text-oriented. Consequently, the content in BitChute and captured by the MeLa-BitChute dataset represents a unique multi-modal resource that can provide insight into current and emerging topics of conspiracy and political misinformation.

Supply, Demand, and Health Misinformation Following the release of the misinformation-laden *Plandemic* film on social media, Facebook, YouTube, and Twitter responded quickly to limit its spread (Kearney, Chiang, and Massey 2020). The film contained many unfounded conspiracy theories about the origin of the COVID-19 pandemic, leading these mainstream platforms to ban or suppress it as it violated policies on misinformation related to public health. Despite these interventions, the film remained widely available on alt-tech spaces and on BitChute in particular (Belle-mare, Nicholson, and Ho 2020), and analysis in Buntain et al. (2021a) suggests BitChute received more traffic in response, as interest in the film drove viewers to platforms that were willing to host it. Viewed through the supply-and-demand framework in Munger and Phillips (2019), as the mainstream platforms limit COVID-19 misinformation in their spaces, the supply of this content moves to these alt-tech spaces. If demand remains constant, interventions by the mainstream platforms may then push audiences toward these alternative spaces, where both misinformation and extreme rhetoric are more common. Two questions then emerge: First, is health misinformation becoming more popular in BitChute and related spaces as Facebook, YouTube, and others increasingly suppress that content locally, and second, how are content producers on BitChute, Gab, etc. responding to this influx of demand – e.g., are they producing more such content? The MeLa-BitChute dataset

⁵<https://www.vox.com/recode/22240408/facebook-youtube-twitter-qanon-misinformation-inauguration>

provides insight into these questions, first by allowing researchers to evaluate trends in engagement and viewership, and second through longitudinal data of content creators and the videos they post to their channels over time.

Alternative Monetization Related to the supply and demand questions above, creators have multiple ways to monetize their supply of content. YouTube’s Partner Program, for example, pays creators a portion of advertising revenue based on views (Kopf 2020). When YouTube deplatforms or otherwise suppresses content, however, creators are “de-monetized” and lose out on this revenue stream. Despite these interventions, creators have found ways to circumvent deplatforming by posting their more violative content on spaces like BitChute and sharing trailers to this content on mainstream platforms (Trujillo et al. 2020). To this end, an ecology of alternative monetization schemes now exists, allowing creators to monetize their content through other means, such as donation, cryptocurrency, affiliate marketing, or merchandise (Chu et al. 2022). BitChute also supports a variety of these monetization options, both through on-platform advertising and integration with donation-based platforms (e.g., Patreon, PayPal, and others). Warreth (2021) details how the far-right and extremist groups use these alternatives, especially cryptocurrency as funding sources. Via the the MeLa-BitChute dataset, researchers can study how BitChute’s alternative content is monetized through these alternative means.

Hate and Online Extremism A rich body of work has examined hateful and radicalizing content in the mainstream platforms. Evidence shows such content on Facebook has contributed to violence and radicalization; e.g., Jihadist groups have used the platform to radicalize potential recruits (Thompson 2011), anti-refugee sentiment predicts criminal acts targeting refugees in otherwise similar communities (Müller and Schwarz 2021), and anti-Muslim sentiment has been used to stimulate fear and violence against Muslim communities (Fink 2018). Alt-tech spaces like Gab are known to have high proportions of hate speech (Zannettou et al. 2018), and BitChute is no exception, with much of its content containing hateful and extreme rhetoric, often antisemitic or racist in nature (Trujillo et al. 2020; Papadamou 2021). Hate speech need not be confined to textual modalities either, as the Anti-Defamation League has shown through its database of hateful symbols⁶, and the data contained in the MeLa-BitChute dataset may yield data for studying hateful imagery, as the platform’s core affordance is video sharing. Prior work has also shown such propensity towards hate is both indicative and predictive of violent acts (Abdalla, Ally, and Jabri-Markwell 2021). Coupling these works with the MeLa-BitChute dataset can provide a lens through which this hateful content can be studied and assessed for potential harm or as an indicator for new violent attacks. Likewise, the multi-year timeframe covered by the MeLa-BitChute dataset may allow researchers insight into radicalization processes among BitChute content creators and their audiences. Research enabled by this

⁶<https://www.adl.org/hate-symbols>

dataset those studying terrorism, who have expressed a desire for more data-driven analyses of hateful and extreme online content, actors, and audiences (Pelzer 2018).

6 Related Datasets

Many available social media datasets are typically focused – e.g., covering disasters (Buntain et al. 2021b), COVID-19 (Qazi, Imran, and Ofli 2020), and hate/harassment (Davidson et al. 2017) – with new datasets regularly released from spaces like SemEval⁷ or the annual Text Retrieval Conference (TREC). While these datasets are valuable resources for understanding specific phenomena, they provide limited general insight into trends across the full platforms. To study new questions not covered by these topic-specific datasets, researchers often need to build new datasets, which introduces confounders when these datasets need to be collected retrospectively. Such problems include limits on search timeframes (i.e., one can only go back so far), memory-hole problems (Marshall 2020) (i.e., relevant information may have been deleted from the target platform, especially a risk for anti-social behavior), or API changes. The MeLa-BitChute dataset outlined herein solves these issues by providing a reusable, general collection that characterizes the full BitChute platform over a multi-year period, facilitating research questions in a consistent context.

These issues can be addressed with sufficiently large samples of the target platform, and it is this type of sample that the MeLa-BitChute dataset provides. Similar platform-wide datasets are available for other platforms, most notably the Pushshift.io Reddit dataset (Baumgartner et al. 2020). Unlike Reddit, BitChute is part of the alt-tech space and fits into a constellation of recent work studying these alternative spaces, including releases from Parler (Aliapoulos et al. 2021), Gab (Fair and Wesslen 2019), and Mastodon (Zignani et al. 2019). Paralleling YouTube’s centrality in the mainstream information ecosystem, BitChute and its video-hosting is likewise a core element of the alt-tech space and is often a highly shared domain in other fringe platforms and Telegram channels. Taken together, these collections provide a crucial cross-platform view into the ecosystem.

Separate from the above published datasets, “hacktivists” have released several large-scale datasets from alt-tech spaces, including Gab and Parler⁸. These datasets, such as those hosted on DDoSecrets.com and Wikileaks, provide more insight into these platforms but at significant ethical and intellectual risk. In particular, these leaked datasets often contain private data, such as direct messages, that users would not intend for public distribution. The MeLa-BitChute dataset instead exclusively contains public-facing data, which may miss out on important activity like collusion, brigading, radicalization, or other anti-social behaviors. These two sources come with different insights and ethical considerations, and we leave questions about which source is most appropriate to future researchers.

⁷<https://semeval.github.io/>

⁸<https://arstechnica.com/information-technology/2021/09/anonymous-leaks-gigabytes-of-data-from-epik-web-host-of-gab-and-parler/>

7 Conclusion

In this paper, we presented a dataset covering 3M+ videos, 61K+ channels, and 11.4M+ comments from the alt-tech, social, video-hosting platform BitChute. We provided an in-depth description of our custom built data collection infrastructure, documentation on the stored data, and a discussion of potential use cases for the dataset. We argued due to the difficulty of data collection, the academic literature is lacking diverse, large-scale studies of BitChute and its role in the alt-tech ecosystem. By filling this gap, researchers can gain a holistic-view of the alt-tech environment and the potential public harms fueled by BitChute. The MeLa-BitChute dataset and sample code can be found at: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KRD1VS>.

A Data Column Descriptions

In this appendix, we provide detailed descriptions of each data column in the MeLa-BitChute dataset. Below are tables for each table in the database (videos, comments, channels).

Column Name	Description
url	URL to video
title	Title of the video
postdate	Date video was uploaded to BitChute. Unparsed from the website, like “First published at 17:31 UTC on August 28th, 2020.”
scrapedate	Date video was added to our database, Unix Epoch time, accuracy in seconds
description	Uploader-chosen description of their video
channel	URL to the channel
profile	URL to the uploader’s profile. Note, a profile can have multiple channels, but a channel belongs to one profile.
category	Uploader picked category of the video. Note, the default category is ‘None’ or ‘Other’ depending on if the video comes before or after September 2020.
sensitivity	Uploader chosen sensitivity score, chosed from: “Normal”, “NSFW”, and “NSFL”
views	Integer number of video views at data collection time. If the views data is -1 , the video was removed by the uploader. If views is -3 , the video was removed by BitChute.
view_scrapedate	Unix Epoch time when views and comments were added to the database. Guaranteed to be at least one week after <code>scrapedate</code> .

Table 1: videos data description

Column Name	Description
url	URL to video that the comment falls under
userid	A SHA256 hash that uniquely identifies each commenter
posthtml	The full HTML of the comment
posttext	The body text of the comment (a pre-processed version of <code>posthtml</code>)
comment_id	A text ID identifying a comment on a video
parent_id	If non-NULL, refers to the <code>comment_id</code> of the parent comment

Table 2: comments data description

Column Name	Description
url	URL to the channel
scrapedate	UTC timestamp of the date on which the description data was collected
description	Full HTML of channel description on the ‘About’ page of the channel. These range from very long descriptions to short or blank descriptions. If no description was found, the value of the column will be ‘Null’. However, note that occasionally we found channels with descriptions that were multiple blank characters, making the stored value not ‘Null’
description_stripped	Text of description stripped from the HTML

Table 3: channels data description

References

- Abdalla, M.; Ally, M.; and Jabri-Markwell, R. 2021. Dehumanisation of ‘Outgroups’ on Facebook and Twitter: towards a framework for assessing online hate organisations and actors. *SN Social Sciences*, 1(9): 1–36.
- Aliapoulios, M.; Bevensee, E.; Blackburn, J.; Bradlyn, B.; Cristofaro, E. D.; Stringhini, G.; and Zannettou, S. 2021. A Large Open Dataset from the Parler Social Network. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1): 943–951.
- Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1): 830–839.
- Bellemare, A.; Nicholson, K.; and Ho, J. 2020. How a debunked COVID-19 video kept spreading after Facebook and YouTube took it down.
- Buckley, N.; and Schafer, J. S. 2021. “Censorship-free” platforms: Evaluating content moderation policies and practices of alternative social media. *OSF preprint 10.31235/osf.io/yf9qz*.
- Buntain, C.; Bonneau, R.; Nagler, J.; and Tucker, J. A. 2021a. Youtube recommendations and effects on sharing across online social platforms. *PACM on Human-Computer Interaction*, 0(0).
- Buntain, C.; Buntain, C.; Mccreadie, R.; and Soboroff, I. 2021b. Incident Streams 2020 : TRECIS in the Time of COVID-19. *18th International Conference on Information Systems for Crisis Response and Management*, (April).
- Buntain, C.; Innes, M.; Mitts, T.; and Shapiro, J. N. 2021c. Effects of the Post-January 6 Deplatforming on Social Media Discourse.
- Chu, A.; Arunasalam, A.; Ozmen, M. O.; and Celik, Z. B. 2022. Behind the Tube: Exploitative Monetization of Content on YouTube. In *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association.
- Davidson, T.; Warmesley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 512–515.
- Doesburg, J. V. 2021. *An alternative rabbit hole? An analysis of the construction of echo chambers within coronavirus activism groups on Telegram*. Masters, Utrecht University.
- Fair, G.; and Wesslen, R. 2019. Shouting into the void: A database of the alternative social media platform Gab. *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, (Icwsml): 608–610.
- Fink, C. 2018. Dangerous Speech, Anti-Muslim Violence, and Facebook In Myanmar. *Journal of International Affairs*, 71(1.5): 43–52.
- Forberg, P. L. 2021. From the Fringe to the Fore: An Algorithmic Ethnography of the Far-Right Conspiracy Theory Group QAnon. *Journal of Contemporary Ethnography*, 1–27.
- Freelon, D.; Marwick, A.; and Kreiss, D. 2020. False equivalencies: Online activism from left to right. *Science*, 369(6508): 1197–1201.
- Júnior, M.; Melo, P.; Da Silva, A. P. C.; Benevenuto, F.; and Almeida, J. 2021. Towards Understanding the Use of Telegram by Political Groups in Brazil. *ACM International Conference Proceeding Series*, 237–244.
- Kearney, M. D.; Chiang, S. C.; and Massey, P. M. 2020. The Twitter origins and evolution of the COVID-19 “plandemic” conspiracy theory. *Harvard Kennedy School Misinformation Review*, 1(October): 1–18.
- Kopf, S. 2020. “Rewarding Good Creators”: Corporate Social Media Discourse on Monetization Schemes for Content Creators. *Social Media and Society*, 6(4).
- Marshall, J. M. 2020. The Modern Memory Hole: Cyberethics Unchained. *Athenaeum Review*, 1(3).
- Müller, K.; and Schwarz, C. 2021. Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association*, 19(4): 2131–2167.
- Munger, K.; and Phillips, J. 2019. A Supply and Demand Framework for YouTube Politics. In *OSF Preprint*. OSF.
- Papadamou, K. 2021. *Characterizing Abhorrent, Misinformative, and Mismatched Content on YouTube*. Doctoral thesis, Cyprus University of Technology.
- Paudel, P.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. An Early Look at the Gettr Social Network. *arXiv preprint arXiv:2108.05876*.
- Pelzer, R. 2018. Policing of Terrorism Using Data from Social Media. *European Journal for Security Research*, 3(2): 163–179.
- Qazi, U.; Imran, M.; and Ofli, F. 2020. GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information. *SIGSPATIAL Special*, 12(1): 6–15.
- Rauchfleisch, A.; and Kaiser, J. 2021. Deplatforming the Far-right: An Analysis of YouTube and BitChute. *SSRN Electronic Journal*, (108).
- Rogers, R. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication*, 35(3): 213–229.
- Rye, E.; Blackburn, J.; and Beverly, R. 2020. Reading In-Between the Lines: An Analysis of Dissenter. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC*, 133–146.
- Thompson, R. 2011. Radicalization and the Use of Social Media. *Journal of Strategic Security*, 4(4): 167–190.
- Trujillo, M.; Gruppi, M.; Buntain, C.; and Horne, B. D. 2020. What is BitChute? Characterizing the “Free Speech” Alternative to YouTube. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media, HT ’20*, 139–140. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370981.
- Tucker, J.; Guess, A.; Barber, P.; Vaccari, C.; Nyhan, B.; Seigel, A.; Sanovich, S.; and Stukal, D. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. Technical Report March, Hewlett Foundation.
- Van Dijck, J.; de Winkel, T.; and Schäfer, M. T. 2021. Deplatformization and the governance of the platform ecosystem. *New Media and Society*, (January).
- Warreth, S. 2021. *Comparing Far Right and Jihadi Use of Crowdfunding, Cryptocurrencies, and Blockchain Technology: Accessibility, Geography, Ideology*. Masters thesis, University of Glasgow.
- Wilson, T.; and Starbird, K. 2021. Cross-platform Information Operations: Mobilizing Narratives and Building Resilience through both ‘Big’ and ‘Alt’ Tech. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2).
- 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*, volume 12, 1007–1014. ISBN 9781450356404.
- Zignani, M.; Quadri, C.; Galdeman, A.; Gaito, S.; and Rossi, G. P. 2019. Mastodon content warnings: Inappropriate contents in a microblogging platform. *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, (Icwsml): 639–645.