iDRAMA-rumble-2024: A Dataset of Podcasts from Rumble Spanning 2020 to 2022

Utkucan Balci[*1], **Jay Patel**[*1], **Berkan Balci** [2], **Jeremy Blackburn** [1]

[1] Binghamton University
[2] Middle East Technical University
ubalci1@binghamton.edu, jpatel67@binghamton.edu, e252615@metu.edu.tr, jblackbu@binghamton.edu

Abstract

Rumble has emerged as a prominent platform hosting controversial figures facing restrictions on YouTube. Despite this, the academic community's engagement with Rumble has been minimal. To help researchers address this gap, we introduce a comprehensive dataset of about 6.7K podcast videos from August 2020 to December 2022, amounting to over 5.6K hours of content. Besides covering metadata of these podcast videos, we provide speech-to-text transcriptions for future analysis. We also provide speaker diarization information, a collection of 168K unique representative images from podcast videos, and face embeddings of more than 400K extracted faces. With the rise of the influence of podcasts and populist figures, this dataset provides a rich resource to identify challenges in cyber social threats in a relatively underexplored space.

Introduction

We have witnessed the rise of alternative social media platforms that target users dissatisfied with content moderation policies of more established platforms (e.g., Twitter and YouTube). These platforms, e.g., Parler, Gab, and Voat, frequently market themselves as bastions of "free speech (Zannettou et al. 2018; Goodwin 2021; Robertson 2015)," a claim that resonates with a broad audience that values unrestricted expression. However, beneath the surface lies a more complex and concerning reality. These platforms can inadvertently transform into fertile grounds for the proliferation of extremist ideologies and the widespread dissemination of misinformation (Aliapoulios et al. 2021a; Papasavva et al. 2020a). This phenomenon unmasks a big gap between the promise of unrestricted discourse and the potential for harmful echo chambers that these environments can promote.

zenodo.org/records/10515991

hf.co/datasets/iDRAMALab/iDRAMA-rumble-2024

github.com/idramalab/iDRAMA-rumble-2024

[*]These authors contributed equally.

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

We have also witnessed the rise of podcasts. The aftermath of the COVID-19 pandemic has seen a surge in podcast popularity (Alesi 2021). In 2021, 82M (25%) of the U.S. population listened to podcasts (Götting 2024). This popularity has also led to an increase in the popularity of podcast videos. By the end of 2023, more than half of the top 30 podcasts are available as videos (Escandon 2024). This rise also come with increased popularity of controversial podcasters that are often right-wing (e.g., Dan Bongino (Culliford and Dave 2022)), who have faced restrictions from YouTube due to their misleading or hateful content. These restrictions resulted in the migration to another online platform, Rumble [1], that promotes itself as a platform that protects freedom of speech (Gillespie 2024). Rumble is recognized for hosting controversial personalities, including Donald Trump Jr., Andrew Tate, and Alex Jones (Marcus 2023; Farah 2023; McCluskey 2022), and reached an average of 58M monthly active users in 2023 (Rumble 2023).

Contributions. In this work, we compile a large-scale dataset from Rumble. Specifically, we collect 6,735 podcast videos along with their corresponding metadata, spanning 5,612 hours of content. Our contributions extend beyond metadata release; leveraging state-of-the-art models, we extract information across three modalities: 1) text, 2) audio, and 3) video. We detail the methodology for extracting information from podcast videos and release a first-of-its-kind dataset including data from different modalities:

- **Metadata:** Details about podcast videos, e.g., channel name, video name, video description, and more.
- **Text:** Transcription (i.e., speech-to-text) of podcast videos.
- **Audio:** Speaker diarization information providing speaker detection over time for each video.
- **Video:** Sampled representative video frames from each video, totaling 168K images. We also detect more than 400K non-unique faces from these images and release face embeddings.

Due to copyright concerns, we refrain from releasing raw audio and videos from our dataset. However, we will make them available to bonafide researchers upon request. In summary, we release metadata, generated video transcripts, rep-

[1]Rumble is accessible through: https://rumble.com

resentative images of the video, extracted speaker information, and detected face embeddings for all 6,735 podcast videos.

Relevance. Rumble is a video-sharing platform established in 2013 as an alternative to YouTube (McCluskey 2022). This platform hosts a wide variety of podcast content, including political discussions, and is known for hosting controversial personalities that faced restrictions from YouTube, including Donald Trump Jr., Andrew Tate, Fresh & Fit, and Alex Jones (Marcus 2023; Farah 2023; Horowitz; McCluskey 2022). As platform migrations can result in an increased toxicity and radicalization in online networks (Horta Ribeiro et al. 2021; Ali et al. 2021), our dataset can help the research community broaden its focus on the effects of deplatforming events in two underexplored domains on cyber social threats: 1) podcast content, and 2) podcast hosts, which also contain public figures that are known for their hateful speeches (Wilson 2022; Mathes and Kaplan 2023).

The release of this dataset will also provide benefit to a broader audience within the research community. As an emerging platform, researchers can leverage our dataset to understand political polarization in podcast videos, track the evolution of topics on Rumble, and conduct data-driven comparisons with mainstream platforms.

Related Work

A variety of datasets have been released to analyze social media platforms. Datasets on Twitter focus on sociopolitical issues, e.g., US Elections (Chen, Deb, and Ferrara 2022), climate change (Effrosynidis et al. 2022), Russian invasion of Ukraine (Haq et al. 2022; Shevtsov et al. 2022), COVID-19 pandemic (Alqurashi, Alhindi, and Alanazi 2020; Naseem et al. 2021; Hayawi et al. 2022). For Instagram, (Zarei et al. 2021) presented a multilingual COVID-19 dataset that contains 25.7K posts and 829K comments. Additionally, (ALBayari and Abdallah 2022) and (Hamlett et al. 2022) released datasets that aim to help combat cyberbullying, and include human-annotated comments. There have also been several datasets released using Facebook data (Dragan and Zota 2017; Menon 2012; Rieder 2013; Santia and Williams 2018) and YouTube comments (Chakravarthi et al. 2021; Ashraf et al. 2022). Although not as widely studied as these platforms, (Steel, Parker, and Ruths 2023) published a TikTok dataset on content related to the Russian invasion of Ukraine.

Datasets related to online platforms extend beyond the mainstream ones, covering platforms like BitChute (Trujillo et al. 2022), 4chan (Papasavva et al. 2020b), Gab (Fair and Wesslen 2019), Parler (Aliapoulios et al. 2021b), Gettr (Paudel et al. 2021), and Voat (Mekacher and Papasavva 2022), which are particularly insightful for understanding extremist or niche online communities.

However, the development of datasets for podcast-related research remains relatively limited. (Lea et al. 2021) presented the SEP-28k dataset, comprising over 28K audio clips to aid research related to stuttering. (Schmidt, Pons, and Miron 2022) contributed the PodcastMix dataset, focusing on the separation of background music from foreground speech in podcasts. Additionally, (Clifton et al. 2020) released the Spotify Podcasts Dataset, a collection of 100,000 audio only podcast episodes that comprise 60K hours of speech, accompanied by ASR (Automatic Speech Recognition) transcripts using Google Cloud's Speech-to-Text API. (Saha, Nayak, and Baumann 2022) compiled a dataset of Angela Merkel's podcast videos that spans 16 years. However, to date, there has not been a large-scale podcast dataset of an online platform that includes text, audio, and visual content. Therefore, in this work, we fill this gap by releasing a dataset that helps researchers to focus on exploring video podcast in online settings.

Data Collection

In this section, we provide an overview of our dataset and crawling methodology. We also present a summary analysis of the dataset, including a monthly distribution of podcast videos and statistics of the top-20 channels, ranked by video count.

Crawling Methodology. We create a specialized crawler to collect video information from the "Podcasts" section on Rumble. This crawler systematically moves through the URLs, scanning the pages in the podcast section until it encounters no new pages. Initially, we deploy our crawler during October 2022 and conduct a subsequent run in early 2023 to cover the entirety of 2022. In early July 2023, we updated the metadata for the video pages in our dataset and limited out dataset to the videos that are available by this date. This method was chosen to ensure that each video's metadata, e.g., views, upvotes, and dislikes, had at least six months to stabilaze and accurately represent their actual statistics. Through this process, we compile a dataset containing 6,761 videos across 246 channels, covering the period from August 27, 2020, to December 31, 2022.

Language verification for podcasts. Building on previous work (Clifton et al. 2020), we run language detection on podcast video descriptions to filter out content other than English. We use the langdetect library (Danilak 2021), a Python implementation of Google's language detection library. Prior to our analysis, we remove hyperlinks from the descriptions. During our analysis, we observe that some videos are detected as non-English, mainly due to having short video descriptions (e.g., social media platforms and their URLs). Consequently, we conduct a manual inspection of the podcast videos that are detected as non-English and videos with no description, and remove podcast videos of the "Monarky" channel as it produces content other than English. In the end, our dataset consists of 6,735 podcast videos.

Data Overview. Table 1 shows Top 20 channels by podcast video counts in our dataset, where we can see the inclusion of notable right-wing content creators (e.g., Dan Bongino, Tim Pool, Charlie Kirk, and Steven Crowder). "The Dan Bongino Show" is the most prolific content creator with a total of 576 videos in our dataset, along with the most followers across all channels in our dataset. This is followed

Channel Name	Count	# Followers
The Dan Bongino Show	576	2.79M
TimcastIRL	326	398K
Ben Shapiro	297	1.02M
Timcast	219	381K
The Charlie Kirk Show	215	1.22M
Steven Crowder	212	1.36M
Dinesh D'Souza	208	1.73M
Liz Wheeler	207	48.1K
The Trish Regan Show	190	248K
vivafrei	178	358K
The Rubin Report	174	466K
The Clay Travis & Buck Sexton Show	169	147K
HodgeTwins	152	742K
AMERICA First with Sebastian Gorka	144	383K
Michael Knowles	139	73.6K
Joe Pags	134	174K
The Jimmy Dore Show	125	202K
Matt Walsh	106	131K
Diamond and Silk	105	607K
phetasy	100	44.6K

Table 1: Top 20 channels by their total number of videos in our dataset and follower count for each channel. NB: Follower count may have changed as these numbers are from July 2023.

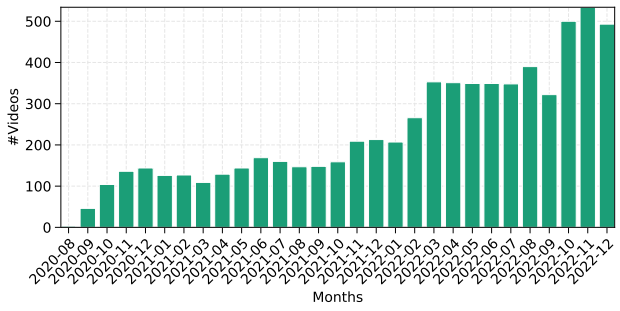

Figure 1: Number of videos posted per month in our dataset.

by "TimcastIRL" and "The Ben Shapiro Show," which account for 326 and 297 podcast videos, respectively. Figure 1 plots the number of videos per month within our dataset. The data clearly indicates an increase in video counts for the year 2022, with approximately two-thirds of our entire collection originating from this period.

Ethical Considerations

In this work, we do not engage with users directly, collecting only publicly available data. As a result, our institution's Institutional Review Board (IRB) does not classify our work as human subject research. Public figures, e.g., podcast channel owners on Rumble, are not anonymized and transcriptions may include hate speech.

Data Release & FAIR Principles

Curating this dataset involved using both CPU compute and NVIDIA A100 GPUs with 80GB of memory, in total, taking approximately 11 compute days (284 compute hours) for transcript generation, 22 days (528 hours) for speaker diarization, and about 5 days (125 hours) for face embeddings generation. We release approximately 36GB of curated podcast data from Rumble, accessible on Zenodo and on Huggingface. We follow FAIR (Findable, Accessible, Interoperable, Re-usable) guidelines for data release, making sure it is discoverable, accessible, and freely available through digital object identifier (DOI) at Zenodo. Additionally, we release code that allows researchers to use our data, which is available on our GitHub.

	Input/Output Modality	**#Data-points**
Metadata	-	6,735
Transcript	Audio/Text	6,735
Representative Images	Video/Image	168K
Speaker diarization	Audio/Text	6,735
Face embeddings	Repr. Images/Vector	400K

Table 2: ***"iDRAMA-rumble-2024"*** dataset release information.

Data Curation Methodology

In this section, we outline the data curation methodology we used. We then discuss the data structure and address the size and format of the released data, offering guidance to researchers on what to expect during the download.

Our data curation pipeline, depicted in Figure 2, begins with the use of a custom crawler to collect podcast videos from Rumble. Then, we process all collected videos to generate the transcripts and speaker diarization via the audio modality. We extract representative frames from each video, using them for face detection and generating corresponding face embeddings. A comprehensive overview of the dataset is presented in Table 2.

Metadata

Our dataset contains a variety of metadata providing unique perspectives on the content, including video durations, publication dates, engagement metrics (such as comments, followers, views, upvotes, and dislikes), and content descriptors like video descriptions and tags. This metadata can help researchers conduct analyses to better understand the dynamics of podcast video content on Rumble, described below:

- **v_id:** Unique "id" of the video.
- **publication_date:** Publication date of the video.
- **rumble_url:** URL of the video on "rumble.com."
- **title:** Video title provided by channel.
- **channel_name:** Name of the channel the video belongs to.
- **video_duration:** Length of the podcast video in seconds.

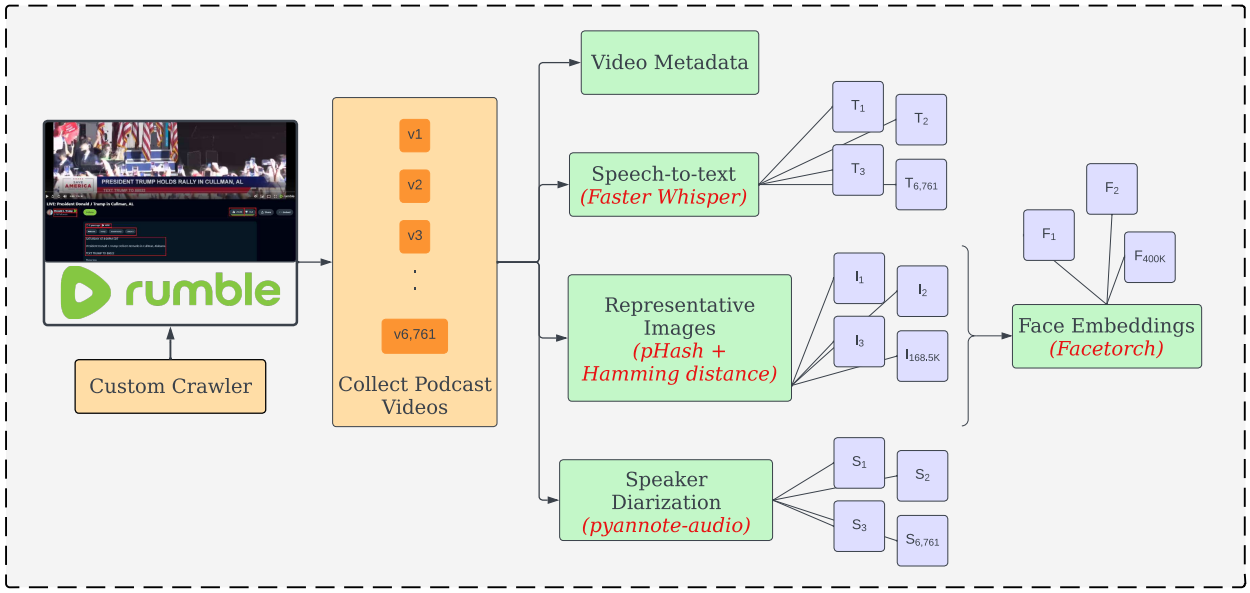

Figure 2: Pipeline of our data curation methodology: 1) Custom crawler collects data from Rumble (6,735 podcast videos), 2) Extracts video metadata, for example, follower counts, comment counts, and more, 3) Using 'faster-whisper' generate transcriptions (6,735 transcriptions), 4) Samples representative images of podcast videos using our methodology (168K unique images), 5) Using 'pyannote-audio' for speaker diarization (6,735 files), 6) Using 'Facetorch' to detect and generate face embeddings from representative images (400K non-unique faces).

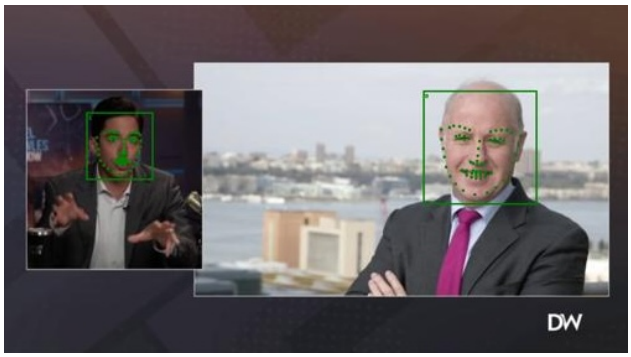

Figure 3: Multiple face detection through applied methodology on a sampled representative video frame from one of the podcast videos in our dataset.

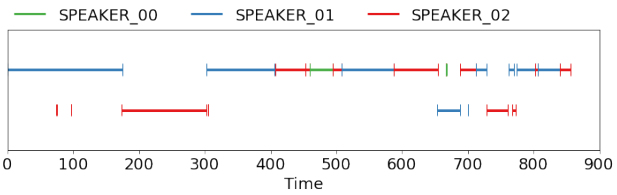

Figure 4: Speaker detection with temporal information using speaker diarization methodology on a sampled 870-second video. The X-axis represents time in seconds.

- **follower_count:** Total number of subscribers to the channel that uploaded the podcast video at the time of collection.
- **view_count:** Total number of how many times the podcast video has been viewed at the time of collection.
- **upvote_count:** Total number of positive ratings or 'likes' received by the podcast video at the time of collection.
- **dislike_count:** Total number of negative ratings or 'dislikes' registered for the podcast video at the time of collection.
- **comment_count:** Total number of comments posted on the podcast video at the time of collection.
- **tags:** Keywords or phrases provided by the channel, specifically selected to encapsulate the main themes, subjects, or topics relevant to the podcast video.
- **video_description:** Official textual synopsis or summary of the podcast video, as provided by the content creator or channel. We observe that Rumble typically formats video descriptions in two segments. The primary segment offers a concise synopsis or summary of the video's content, while the secondary segment often includes supplementary information, e.g., social media links. Although our dataset encompasses these additional video descriptions (**video_description_more**), the analysis in this paper is conducted based on the information provided in the primary video description segment.

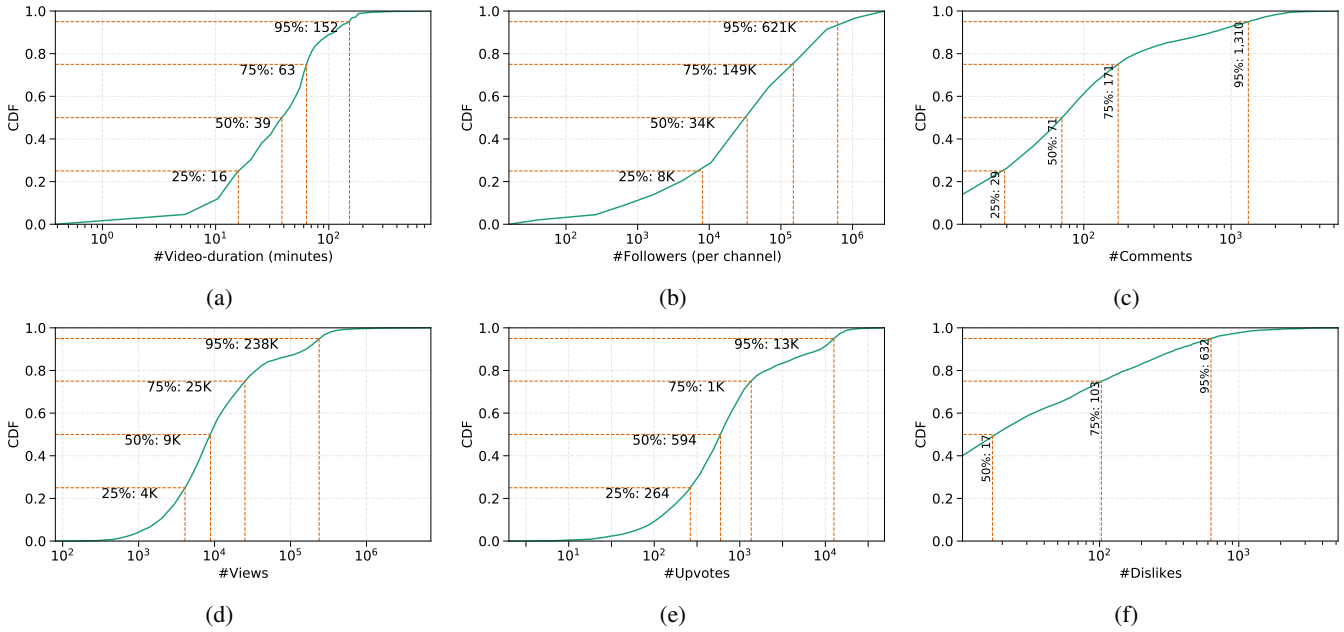

Figure 5: CDF of metadata attributes.

Transcription

For transcribing podcast videos, we use faster-whisper (Klein 2023), a reimplementation of OpenAI's Whisper through CTranslate2, along with Silero's Voice Activity Detection (VAD) (Team 2021). This approach proves especially adept at addressing common issues in many videos in our dataset, e.g., extended pauses and background music. In our data curation, we use the "large-v2" model of Whisper to transcribe the audio content of the podcast videos. We set the language to English, enabled VAD, and configure our model to generate word-level timestamps. In our dataset we provide two different structure for transcriptions:

- **transcription:** Raw transcription segments provided by faster-whisper which include all details, e.g., probability of the word predictions.
- **transcription-lite:** Processed segments that include only text and word-level timestamps.

Named Entities

To facilitate in-depth examination of titles, video descriptions, and transcripts, we use the *en_core_web_lg* model from SpaCy. This tool is commonly used by the research community (Balcı, Sirivianos, and Blackburn 2023; Filgueira et al. 2020; Papasavva et al. 2020b) for its effectiveness in named entity recognition, having been trained on multiple datasets, including WordNt (Miller 1995) and Common Crawl (Repository 2023), using multi-task CNN and GloVe vectors. In our analysis, we have intentionally excluded commonplace labels, i.e., cardinal, ordinal, and date to concentrate on entities of greater significance, but they remain in the raw dataset. Our dataset provides named entities for titles, video descriptions, and speech-to-text transcriptions.

Representative Images (Sampling Unique Frames)

To sample unique frames per podcast video, we extract images at a rate of one frame per second and apply perceptual hashing (pHash) to each frame. Then, we calculate the similarity between images using Hamming distance. Finally, we select images with a similarity greater than a set threshold ($\theta = 20$) to identify frames with meaningful visual differences. This threshold is selected by three authors of the paper after examining 20 sampled videos for thresholds 5 to 50, incrementing by 5, aiming to maximize visually distinct images while minimizing loss of information. In total, we sampled around 168K unique images to represent podcast videos available in the dataset.

Speaker Diarization

To extract meaningful information from podcast videos, for example, the number of speakers and the temporal aspects at which an individual speaker is speaking in it (if more than one speaker exists), we use the Pyannote-audio library (Plaquet and Bredin 2023; Bredin 2023). Specifically, we use the "pyannote/speaker-diarization@2.1" model[2] to detect speakers and their temporal information through speaker diarization technique.

Face Embeddings

While discussing real-world events or politics in podcasts, podcasters often host personalities, celebrities, or political figures. Detecting faces from podcast videos can be

[2]https://huggingface.co/pyannote/speaker-diarization.

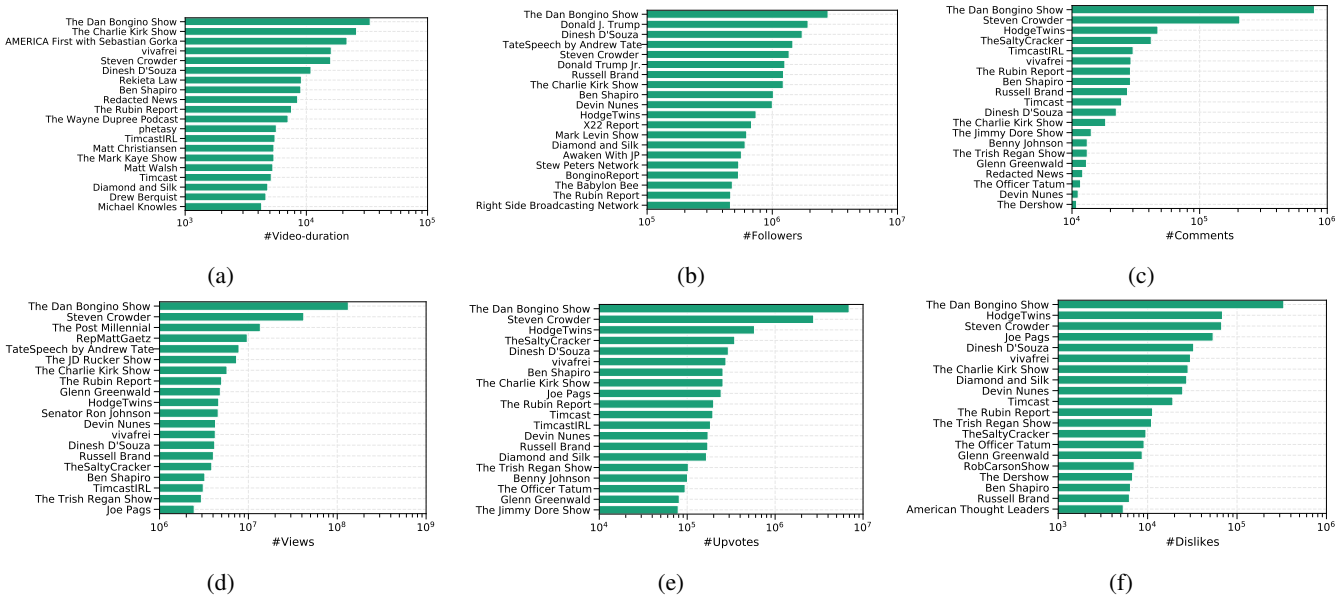

Figure 6: Top 20 channels of the metadata attributes.

helpful for a comprehensive understanding of the visual themes and discussed topics. To extract faces, we leverage AdaFace (Kim, Jain, and Liu 2022) and RetinaFace (Deng et al. 2020) models, using the Facetorch package. [3] This approach generates 512-dimensional embeddings of the detected faces. We also release images with bounding boxes and locations of detected faces along with face embeddings. Figure 3 illustrates a sample image showcasing the detected faces with bounding boxes.

Metadata Analysis

In this section, we characterize the metadata extracted from the podcast videos within our dataset.

Video Durations. Our dataset consists of videos totaling 5,612 hours, which translates to 237 days. We plot the CDF of video duration in minutes in Figure 5a. The average duration of a video is 50 minutes, with a median of 38 minutes and 32 seconds, and a standard deviation of 49 minutes and 12 seconds. It is important to note that we have not excluded any videos based on duration in our selection process, as we rely on self-identified labels of "Podcasts." This means our dataset also includes videos ranging from 0 to 5 minutes. We believe these shorter videos, though not typical podcasts, are still relevant and could provide valuable insights for future research due to their podcast-related content or labeling.

User Engagement. In this section, we analyze user engagement metrics, including comments, views, upvotes, and dislikes, alongside the follower count for Rumble channels. We plot the CDF of these metrics in Figure 5. In general, our findings corroborate earlier studies (Zarei et al. 2020; Zannettou et al. 2019; Aliapoulios et al. 2021b) on social network metadata and exhibit a characteristic long-tailed distribution.

- **# Followers.** The follower count per channel, depicted in Figure 5b, has a mean of 152,861.13, a median of 33,800, and a standard deviation of 328,519.57. This wide distribution suggests a diverse range of channel popularity, with some channels having amassed a predominantly larger follower base, potentially due to longer presence on the platform or more engaging content. "The Dan Bongino Show" has the most followers (2.7M), where Donald Trump's channel ("Donald J. Trump") is second with 1.9M followers, and Dinesh D'Souza is third with 1.7M followers.
- **# Comments.** We plot the CDF of comment counts for the podcast videos in Figure 5c. The mean comment count is 240.86, the median is 71, and the standard deviation is 475. This skewed distribution suggests that while a few videos attract a high volume of comments, the majority receive a relatively modest number. This could indicate varying levels of viewer engagement or controversy among different videos. "The Dan Bongino Show" leads the total number of comment counts with 793,598 comments, followed by "Steven Crowder" (205,707), and "HodgeTwins" (47,016).
- **# Views.** Figure 5d presents the CDF for the total number of views for the podcast videos. The mean view count is 47,709.65, the median is 8,840, and the standard deviation is 165,711.43, suggesting a wide range in viewership among the videos. The large disparity between the mean and median values indicates the presence of some highly popular videos that elevate the average view count, a common phenomenon in content distribution platforms. "The Dan Bongino Show" has the most cumulative views (133.2M), followed by "Steven Crowder" (42.1M), and

[3] Facetorch models and implementation: https://github.com/tomas-gajarsky/facetorch.

Video Title Named Entities			Video Description Named Entities			Transcript Named Entities			Tags	
Entity	**#**	**Label**	**Entity**	**#**	**Label**	**Entity**	**#**	**Label**	**Tag**	**#**
Biden	346	PERSON	Biden	692	PERSON	Trump	40,571	PERSON	Podcasts	6,735
The Dan Bongino Show	168	ORG	Joe Biden	465	PERSON	Biden	29,417	PERSON	politics	639
Trump	145	PERSON	Dinesh	458	PERSON	America	28,476	GPE	news	557
Democrats	142	NORP	Democrats	418	NORP	Democrats	26,315	NORP	Fox News	553
FBI	120	ORG	Trump	393	PERSON	American	22,848	NORP	trump	527
Sebastian Gorka	116	PERSON	America	277	GPE	Joe Biden	19,699	PERSON	Podcast	514
Joe Biden	83	PERSON	Ukraine	231	GPE	Republicans	19,017	NORP	donald trump	512
Ukraine	81	GPE	FBI	194	ORG	the United States	17,341	GPE	Conservative	506
Elon Musk	73	PERSON	CNN	189	ORG	Russia	16,451	GPE	Politics	503
Russia	72	GPE	Americans	183	NORP	Ukraine	16,065	GPE	president trump	489
China	68	GPE	Russia	180	GPE	China	15,910	GPE	white house	484
Putin	64	PERSON	Dave Rubin	179	PERSON	FBI	15,650	ORG	bongino	483
US	64	GPE	China	178	GPE	Republican	13,394	NORP	fox news channel	482
GOP	55	ORG	Elon Musk	171	PERSON	Donald Trump	13,156	PERSON	opinion	482
America	51	GPE	Buck Sexton	168	PERSON	Americans	12,543	NORP	Democrats	469
Clayton Morris	48	ORG	Republicans	160	NORP	Joe	11,411	PERSON	Liberal	440
The Dan Bongino	46	PERSON	The Rubin Report	159	WORK OF ART	Florida	11,276	GPE	Republicans	433
Matt Gaetz	44	PERSON	Clay Travis	158	PERSON	Twitter	10,676	ORG	News	433
CNN	43	ORG	American	154	NORP	Democrat	10,020	NORP	Trump	358
Democrat	43	NORP	US	152	GPE	Congress	9,223	ORG	Journalism	298

Table 3: Top 20 named entities (based on their occurrences) in video titles, descriptions, and speech-to-text transcriptions within our dataset. We also present tags and their number of occurrences. For each video, we only account for unique tags as multiple tags are mistakenly placed by the channel owner. NB: Tags are user-defined and case sensitive.

"The Post Millennial" (13.6M).

- **# Upvotes.** The upvote distribution, illustrated in Figure 5e, shows a mean of 2,266.67, a median of 594, and a standard deviation of 4,405.57. This indicates a positive skew in user reactions, with a few videos receiving exceptionally high approval but most garnering a moderate number of upvotes. This pattern may reflect the variations in content quality or viewer preferences. Similar to total number of comments, "The Dan Bongino Show" has the most total number of upvotes (6.9M), followed by "Steven Crowder" (2.7M), and "HodgeTwins" (584K).
- **# Dislikes.** Dislikes follow a similar, albeit less pronounced, trend as seen in Figure 5f. The mean dislike count is 130.85, with a median of 17 and a standard deviation of 322.38. This trend can be attributed to specific content triggering negative reactions, but generally, videos tend to have fewer dislikes compared to likes. "The Dan Bongino Show" has the most dislikes (333K), followed by "HodgeTwins" (68K), and "Steven Crowder" (67K).

Content Descriptors. Content descriptors can provide insights into the nature and focus of the videos in our dataset. We analyze two key aspects, tags and video descriptions, to understand the thematic trends of the content. We also release the extracted named entities of video titles and descriptions per video.

- **Tags.** Table 3 presents the top 20 most frequently used tags in our dataset. The most frequent tag, by definition, "podcasts," appears in all videos, followed by "politics" (639) and "news" (557). We also see that the prevalence of politics- and news-oriented tags indicates that the videos in our dataset primarily focus on these subjects.
- **Titles.** Table 3 presents the top 20 most frequent named entities used in the titles in our dataset. The most frequent entity is "Biden" (346), followed by "The Dan Bongino Show" (639), and Trump (557). Similar to tags, the most frequent named entities are politics and news oriented, with the addition of the channel owners.
- **Video Descriptions.** Table 3 highlights the top 20 named entities extracted from the video descriptions in our dataset. The frequent occurrence of entities (e.g., "Biden" (694 mentions), "Joe Biden" (467), and "Trump" (393)) corroborates the political and news-centric orientation of the content, mirroring the findings from the tag and title analyses.

Text, Audio, and Video Analysis

In this section, we perform descriptive analysis of information extracted from text, audio and video modalities through the methodology discussed earlier.

Transcription Analysis

We first look at the CDF of word count per video in Figure 7. The average word count per video is 8,408.62, with a median of 6,934 and a standard deviation of 7,650.14. The figure reflects the diversity in length of verbal content across the podcast videos within our dataset.

Table 3 presents the top 20 named entities identified in the transcripts. Similar to our previous findings, the top named entities mentioned in the video transcripts primarily revolve around political figures and organizations. "Trump" is the

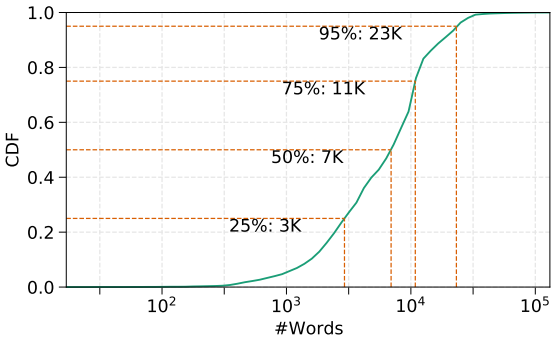

Figure 7: CDF of number of words per speech-to-text transcript (i.e., per video).

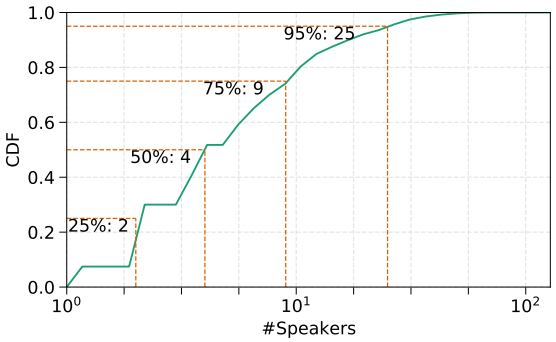

Figure 8: CDF of number of extracted speakers per podcast video.

most frequent named entity with over 40K mentions, followed by "Biden" with 29K, and "America" with 28K.

Channel Name	**# Faces**	**# Speakers** (Median)
The Dan Bongino Show	23,832	8
TimcastIRL	6,978	4
Ben Shapiro	13,974	6
Timcast	5,169	2
The Charlie Kirk Show	15,193	15
Steven Crowder	33,824	16
Dinesh D'Souza	12,595	5
Liz Wheeler	3,484	2
The Trish Regan Show	995	1
vivafrei	7,712	2
The Rubin Report	20,140	10
The Clay Travis & Buck Sexton Show	1,658	3
HodgeTwins	6,113	3
AMERICA First with Sebastian Gorka	22,127	33
Michael Knowles	9,228	5
Joe Pags	801	2
The Jimmy Dore Show	2,671	3
Matt Walsh	7,710	7
Diamond and Silk	8,051	5
phetasy	3,402	5

Table 4: Top 20 channels (by their video count) with the number of extracted speakers and the number of detected faces. NB: The number of faces is aggregated across all podcast videos per channel and is non-unique. The number of speakers are shown without applying any filtering or removal of false positives.

Speaker Analysis

We process all podcast videos to detect speakers (i.e., speaker diarization), excluding videos longer than 3 hours in duration. In total, we extract speakers information from 6,714 videos (out of 6,735). Figure 8 plots the CDF for the number of speakers detected in each podcast video. Across the dataset, videos feature a minimum of 1 and a maximum of 128 speakers, with an average count of 7 speakers (median: 4). Table 4 displays the median count of total speakers for the top 20 channels, with "AMERICA First with Sebastian Gorka" having the highest speakers (per median statistics) per video, while "The Trish Regan Show" consistently features only one speaker per video.

Face Detection Analysis

One use-case of representative images is to detect the appearances of celebrities or political figures. Using around 168K representative images from all podcast videos, we extract 400,791 non-unique faces. Table 4 presents an aggregated view of non-unique faces for the top 20 channels, sorted by their video count. As we see in Table 4, the channels "Steven Crowder," "The Dan Bongino Show," "AMERICA First with Sebastian Gorka," and "The Rubin Report" feature the highest counts of non-unique faces.

Conclusion

This paper presents a comprehensive dataset of 6,735 podcast videos from Rumble, totaling 5,612 hours. Our dataset extends beyond the metadata of these podcast videos to include multiple layers of data across text, audio, and video modalities. These include speech-to-text transcriptions, speaker diarizations, and face embeddings created with state-of-the-art machine learning models. We also present an analysis that provides insights into the data we provide. In our analysis, we reveal that the content of these podcast videos predominantly revolves around politics and news. Our dataset presents a unique resource for researchers aim to explore the depths of podcast content and its implications on broader socio-political research. We believe that the content of this dataset will help improve research on cyber social threats from a unique perspective, where researchers can analyze the implications of controversial figures using the content created by themselves. Given that our dataset includes numerous hosts who have been deplatformed, often associated with right-wing ideologies or the Manosphere, we anticipate it will serve as a fruitful resource in understanding cyber social threats.

Limitations

Our dataset is subject to certain limitations. First, it is important to acknowledge that this data was not collected in real-

time, which may result in missing certain podcast videos if they were published and deleted before our crawler could collect them. The identification of videos as podcasts is based solely on labels provided by content creators, without further verification on our part. Additionally, the precision of our speech-to-text transcriptions, face embeddings, speaker diarization, and named entity recognition is limited by the performance of the models, e.g., Whisper is known for hallucinating content (Mittal et al. 2024). Researchers are encouraged to undertake additional postprocessing steps regarding the specific objectives of their research.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2046590.

References

ALBayari, R.; and Abdallah, S. 2022. Instagram-Based Benchmark Dataset for Cyberbullying Detection in Arabic Text. *Data*, 7(7): 83.

Alesi, T. 2021. The podcast explosion: Who's listening?

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 *Proceedings of the 13th ACM Web Science Conference 2021*, 187–195.

Aliapoulios, M.; Bevensee, E.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; and Zannettou, S. 2021a. An early look at the parler online social network. *arXiv preprint arXiv:2101.03820*.

Aliapoulios, M.; Bevensee, E.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; and Zannettou, S. 2021b. A large open dataset from the Parler social network. In *ICWSM*, volume 15, 943–951.

Alqurashi, S.; Alhindi, A.; and Alanazi, E. 2020. Large Arabic Twitter Dataset on COVID-19. arxiv:2004.04315.

Ashraf, N.; Rafiq, A.; Butt, S.; Shehzad, H. M. F.; Sidorov, G.; and Gelbukh, A. 2022. YouTube Based Religious Hate Speech and Extremism Detection Dataset with Machine Learning Baselines. *Journal of Intelligent & Fuzzy Systems*, 42(5): 4769–4777.

Balcı, U.; Sirivianos, M.; and Blackburn, J. 2023. A Data-driven Understanding of Left-Wing Extremists on Social Media. *arXiv:2307.06981*.

Bredin, H. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Chakravarthi, B. R.; Priyadharshini, R.; Ponnusamy, R.; Kumaresan, P. K.; Sampath, K.; Thenmozhi, D.; Thangasamy, S.; Nallathambi, R.; and McCrae, J. P. 2021. Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments. arxiv:2109.00227.

Chen, E.; Deb, A.; and Ferrara, E. 2022. #Election2020: The First Public Twitter Dataset on the 2020 US Presidential Election. *Journal of Computational Social Science*, 5(1): 1–18.

Clifton, A.; Reddy, S.; Yu, Y.; Pappu, A.; Rezapour, R.; Bonab, H.; Eskevich, M.; Jones, G.; Karlgren, J.; Carterette, B.; and Jones, R. 2020. 100,000 Podcasts: A Spoken English Document Corpus. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 5903–5917.

Culliford, E.; and Dave, P. 2022. YouTube permanently bans Fox News Host Dan Bongino — reuters.

Danilak, M. M. 2021. Language detection library ported from Google's language-detection. https://pypi.org/project/langdetect/.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 5203–5212.

Dragan, I.; and Zota, R. 2017. Collecting Facebook Data for Big Data Research. In *2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet)*, 1–3.

Effrosynidis, D.; Karasakalidis, A. I.; Sylaios, G.; and Arampatzis, A. 2022. The Climate Change Twitter Dataset. *Expert Systems with Applications*, 204: 117541.

Escandon, R. 2024. "video podcasting" growing in popularity.

Fair, G.; and Wesslen, R. 2019. Shouting into the void: A database of the alternative social media platform gab. In *ICWSM*, volume 13, 608–610.

Farah, H. 2023. What is rumble, the video-sharing platform "immune to cancel culture"?

Filgueira, R.; Grover, C.; Terras, M.; and Alex, B. 2020. Geoparsing the historical gazetteers of scotland: Accurately computing location in mass digitised texts. In *Proceedings of the 8th Workshop on Challenges in the Management of Large Corpora*, 24–30. European Language Ressources Association.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2024-01-15.

Gebru, 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.

Gillespie, B. 2024. Rumble makes major announcement in effort to combat censorship, ensure "free and open internet".

Goodwin, J. 2021. GAB: Everything you need to know about the fast-growing, controversial social network — CNN business.

Götting, M. C. 2024. Topic: Podcasting industry.

Hamlett, M.; Powell, G.; Silva, Y. N.; and Hall, D. 2022. A Labeled Dataset for Investigating Cyberbullying Content Patterns in Instagram. *ICWSM*, 16: 1251–1258.

Haq, E.-U.; Tyson, G.; Lee, L.-H.; Braud, T.; and Hui, P. 2022. Twitter Dataset for 2022 Russo-Ukrainian Crisis. arxiv:2203.02955.

Hayawi, K.; Shahriar, S.; Serhani, M. A.; Taleb, I.; and Mathew, S. S. 2022. ANTi-Vax: A Novel Twitter Dataset for COVID-19 Vaccine Misinformation Detection. *Public Health*, 203: 23–30.

Horowitz, J. ???? Rumble is profiting from creators who spread antisemitism.

Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–24.

Kim, M.; Jain, A. K.; and Liu, X. 2022. Adaface: Quality adaptive margin for face recognition. In *CVPR*, 18750–18759.

Klein, G. 2023. faster-whisper. https://github.com/guillaumekln/faster-whisper.

Lea, C.; Mitra, V.; Joshi, A.; Kajarekar, S.; and Bigham, J. P. 2021. SEP-28k: A Dataset for Stuttering Event Detection from Podcasts with People Who Stutter. In *ICASSP*, 6798–6802.

Marcus, J. 2023. Rumble inks seven-figure podcast deal with Donald Trump Jr for "triggered" show.

Mathes, N.; and Kaplan, A. 2023. Rumble promoted and profited from an unhinged, harrowingly antisemitic and homophobic six-hour livestream.

McCluskey, M. 2022. Rumble offers Joe Rogan $100 million to switch platforms.

Mekacher, A.; and Papasavva, A. 2022. " I Can't Keep It Up." A Dataset from the Defunct Voat. co News Aggregator. In *ICWSM*, volume 16, 1302–1311.

Menon, A. 2012. Big data facebook. In *Proceedings of the 2012 workshop on Management of big data systems*, 31–32.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41.

Mittal, A.; Murthy, R.; Kumar, V.; and Bhat, R. 2024. Towards understanding and mitigating the hallucinations in NLP and Speech. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, 489–492.

Naseem, U.; Razzak, I.; Khushi, M.; Eklund, P. W.; and Kim, J. 2021. COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE Transactions on Computational Social Systems*, 8(4): 1003–1015.

Papasavva, A.; Blackburn, J.; Stringhini, G.; Zannettou, S.; and De Cristofaro, E. 2020a. " Is it a Qoincidence?": A First Step Towards Understanding and Characterizing the QAnon Movement on Voat. co. *arXiv:2009.04885*.

Papasavva, A.; Zannettou, S.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2020b. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *ICWSM*, volume 14, 885–894.

Paudel, P.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. An early look at the Gettr social network. *arXiv:2108.05876*.

Plaquet, A.; and Bredin, H. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Repository, C. C. 2023. https://commoncrawl.org/.

Rieder, B. 2013. Studying Facebook via Data Extraction: The Netvizz Application. In *WebSci*, WebSci '13, 346–355. ISBN 978-1-4503-1889-1.

Robertson, A. 2015. Welcome to voat: Reddit Killer, Troll Haven, and the strange face of internet free speech.

Rumble. 2023. Rumble reports third quarter 2023 results.

Saha, D.; Nayak, S.; and Baumann, T. 2022. Merkel Podcast Corpus: A Multimodal Dataset Compiled from 16 Years of Angela Merkel's Weekly Video Podcasts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2536–2540.

Santia, G.; and Williams, J. 2018. BuzzFace: A News Veracity Dataset with Facebook User Commentary and Egos. *ICWSM*, 12(1): 531–540.

Schmidt, N.; Pons, J.; and Miron, M. 2022. PodcastMix: A dataset for separating music and speech in podcasts. *arXiv:2207.07403*.

Shevtsov, A.; Tzagkarakis, C.; Antonakaki, D.; Pratikakis, P.; and Ioannidis, S. 2022. Twitter Dataset on the Russo-Ukrainian War. arxiv:2204.08530.

Steel, B.; Parker, S.; and Ruths, D. 2023. The Invasion of Ukraine Viewed through TikTok: A Dataset.

Team, S. 2021. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad.

Trujillo, M. Z.; Gruppi, M.; Buntain, C.; and Horne, B. D. 2022. The MeLa BitChute Dataset. In *ICWSM*, volume 16, 1342–1351.

Wilson, J. 2022. The downfall of Andrew Tate and its implications.

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*, 1007–1014.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In *Companion proceedings of the 2019 world wide web conference*, 218–226.

Zarei, K.; Farahbakhsh, R.; Crespi, N.; and Tyson, G. 2021. Dataset of Coronavirus Content From Instagram With an Exploratory Analysis. *IEEE Access*, 9: 157192–157202.

Zarei, K.; Ibosiola, D.; Farahbakhsh, R.; Gilani, Z.; Garimella, K.; Crespi, N.; and Tyson, G. 2020. Characterising and detecting sponsored influencer posts on Instagram. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 327–331. IEEE.

Paper Checklist to be included in your paper

1. For most authors...

(a) 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

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) 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 "Ethical considerations and Data Release & FAIR Principles" sections.

(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 Curation Methodology."

(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 Release & FAIR Principles."

(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 Curation Methodology."

(b) Did you mention the license of the assets? Yes, please see Section "Data Curation Methodology."

(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 "Data Curation Methodology."

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, please see Section "Ethical considerations."

(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 "Data Structures and Data Release & FAIR Principles" sections.

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 deidentified? NA