

TUBERAIDER: Attributing Coordinated Hate Attacks on YouTube Videos to their Source Communities*

Mohammad Hammas Saeed¹, Kostantinos Papadamou², Jeremy Blackburn³,
Emiliano De Cristofaro^{2,4}, and Gianluca Stringhini¹

¹Boston University, ²University College London, ³Binghamton University,
⁴University of California, Riverside

Abstract

Alas, coordinated hate attacks, or *raids*, are becoming increasingly common online. In a nutshell, these are perpetrated by a group of aggressors who organize and coordinate operations on a platform (e.g., 4chan) to target victims on another community (e.g., YouTube). In this paper, we focus on attributing raids to their source community, paving the way for moderation approaches that take the context (and potentially the motivation) of an attack into consideration.

We present TUBERAIDER, an attribution system achieving over 75% accuracy in detecting and attributing coordinated hate attacks on YouTube videos. We instantiate it using links to YouTube videos shared on 4chan’s /pol/ board, r/The_Donald, and 16 Incels-related subreddits. We use a *peak detector* to identify a rise in the comment activity of a YouTube video, which signals that an attack may be occurring. We then train a machine learning classifier based on the community language (i.e., TF-IDF scores of relevant keywords) to perform the attribution. We test TUBERAIDER in the wild and present a few case studies of actual aggression attacks identified by it to showcase its effectiveness.

1 Introduction

Coordinated hate attacks are a nefarious online phenomenon whereby bad actors organize on a social platform to orchestrate attacks disrupting other users or communities [7, 16, 26]. A high-profile example is the *GamerGate* campaign, when 4chan users coordinated and recruited others to raid the #GamerGate hashtag on Twitter, targeting women and genderqueer people in the gaming industry with relentless hate speech, death threats, doxxing, and swatting attempts.

Anecdotal evidence suggests that some online communities, e.g., KiwiFarms [4], are almost exclusively dedicated to orchestrating these attacks; others, e.g., 4chan [23], are notorious for that. Since this type of abuse is generated by humans rather than automated programs, techniques to detect content produced by bots [33, 47, 48, 60] are not effective.

In this paper, we focus on coordinated hate attacks on YouTube videos. YouTube is one of the most popular video-

sharing platforms and, alas, it is increasingly targeted by cyber-aggression campaigns [21]. In fact, prior work shed light on fringe communities routinely targeting YouTube videos with raids [23].

Motivation. Past research on mitigating cyber-aggression has mostly neglected the *attribution* of attacks, rather focusing on identifying and removing hateful messages [9, 13, 15], or detecting hateful users [16, 44]. This prompts the need to analyze coordinated aggression attacks from the lens of the online community where the attackers coordinate and the target platform where the attack occurs.

Furthermore, knowing which community is responsible for a coordinated attack could assist platforms in designing more targeted mitigation techniques, e.g., by factoring in context and motivation into moderation decisions.

Technical Roadmap. This paper presents TUBERAIDER, a system that attributes coordinated hate attacks on YouTube to the online community that organized them. We instantiate it using data from three toxic communities: 1) 4chan’s politically incorrect (/pol/) board, 2) the r/The_Donald subreddit, and 3) a collection of 16 Incels subreddits identified by Papadamou et al. [37]. We collect all links to YouTube videos posted on these communities and the comments on those videos on YouTube.

Our analysis shows that videos exhibit a peak in comment activity once they are shared on a platform, possibly indicating that a raid is taking place. Thus, we make TUBERAIDER model the comment activity and detect peaks in the comments of the video when the link is posted on a community. We identify important keywords for each community using TF-IDF and use those as features to train a classifier. This approach allows us to perform a more comprehensive attribution of aggression attacks; rather than focusing on what is being said in the comments to a YouTube video, which is context-dependent and changes over time, we rely on the typical language used by hateful online communities.

TUBERAIDER achieves accuracy above 75% in attributing a coordinated attack on a given video. We also run TUBERAIDER in the wild and identify 700 videos that were likely targeted by coordinated attacks. We then compare the commenting activity on identified videos with those that are not attributed by TUBERAIDER and regular YouTube videos

* Accepted for publication at the 18th International AAAI Conference on Web and Social Media (ICWSM 2024). Please cite accordingly.

across a variety of inflammatory markers to demonstrate the increased toxicity, abuse, and targeted hate in raided videos. We also report case studies of actual attacks that exhibit hate speech and target different individuals.

Implications. Overall, our work paves the way for more effective content moderation. TUBERAIDER enables approaches that take into account the context and the motivation of online attacks by attributing them to the community that orchestrated them. This can help identify terms of service violation (e.g., attacking people based on their race or gender) and make online communities safer.

Due to the topic we focus on, this paper contains examples of misogynistic, hateful, and toxic content; reader discretion is advised. We discuss broader implication of our work, in more detail, in Section 8.1.

2 Related Work

Online Aggression. Research on online aggression mainly focuses on analyzing the toxic content posted on various platforms [9, 13, 15]. Salminen et al. [45] detect hateful comments in the context of online news media, while Olteanu et al. [34] analyze hateful speech on Twitter and Reddit in relation to extremist violence. Zannettou et al. [59] study the spread of hateful memes on the Web with a focus on 4chan’s /pol/ board and r/The_Donald subreddit. Cheng et al. [10] predict accounts that will engage in antisocial behavior for popular websites and detect antisocial behavior in comments, while Chelmiss and Yao [8] predict if a hateful comment on an Instagram post will be followed by further hateful comments. Jaki et al. [24] focus on the Incels.me forum and propose a deep learning classifier that analyzes the users’ language and detects instances of misogyny, homophobia, and racism. Another line of work focuses on characterizing users who violate terms of service, e.g., aiming to distinguish users that post hateful content on Twitter from others [16, 44].

YouTube. Specific to YouTube is the work by Kwon and Gruzd [27], who analyze swearing comments against Donald Trump on YouTube. Moor et al. [32] study “flaming” (i.e., the use of hostile/offensive language) on YouTube videos. Agarwal and Sureka [1] detect YouTube videos *promoting* hatred using user and video features, while Giannakopoulos et al. [18] detect violence using a variety of features like audio, video, and text. Papadamou et al. [37] study the Incels community on YouTube and how to detect Incels-related videos based on a dictionary of Incel-related terms, while Papadamou et al. [38] assess the effects of watch history on YouTube’s pseudoscientific video recommendations.

Sureka et al. [49] find hateful videos on YouTube via social network analysis, while Weaver et al. [52] show that violence in YouTube videos has more realistic consequences and a more negative nature than television violence. Ottoni et al. [35] observe that right-wing YouTube channels feature hateful and discriminatory content, Wotanis and McMillan [55] that female YouTubers receive more negative feedback in terms of sexist and hateful comments as compared to male

YouTubers, and Tucker-McLaughlin [51] that a quarter of the most-viewed YouTube videos include misogynistic language, violence, or both, while the primary actors are male.

Alshamrani et al. [3] investigate the correlation of toxic behaviors like identity hate and obscenity in users’ interactions with popular videos. Tahir et al. [50] detect inappropriate videos targeting children on YouTube.

Other research efforts focus on detecting cyberbullying on YouTube; e.g., Marathe and Shirsat [30] develop a semi-automated system to identify cyber-bullying in videos, while Dadvar et al. [12] use machine learning to identify cyber-bullies on YouTube.

Coordinated Aggression. As part of an exploratory study of 4chan, Hine et al. [23] shed light on raids coordinated on /pol/ targeting communities on Twitter and YouTube. This work motivates ours by showing how a video’s surge in comment activity follows it being shared on another platform.

Some work also focuses on detecting accounts involved in coordinated campaigns. Pacheco et al. [36] use an unsupervised network-based method to discover groups of accounts participating in coordinated influence campaigns, while Sharma et al. [46] model account activity and hidden group behaviors to separate coordinated accounts from normal social media users. Hernandez et al. [22] attributes fraudulent user accounts in online peer-opinion systems to a set of known fraudsters; their system takes in input a seed set of known fraudster profiles and iteratively attributes more users controlled by the same fraudster using graph deep learning. Conversely, our system uses TF-IDF-based approach to model the language of online communities and attributes aggression attacks on YouTube videos to a source community without focusing on whether the accounts posting comments are controlled by the same actor.

Mariconti et al. [31] use ensemble learning to predict whether a YouTube video will be raided, while Kumar et al. [26] investigate *brigading* on Reddit, whereby sub-communities form alliances and perpetrate hate crimes against competitor communities. By contrast, we develop a generalizable model that can assign coordinated aggressive attacks to the community that planned and executed them.

3 Dataset

We now provide an overview of the communities we study as well as the data we collect. Overall, we gather 1,143,988 youtube videos linked from 4chan and Reddit.

3.1 Background

/pol/. 4chan is an imageboard created in 2003. As of January 2023, it features 76 boards covering topics ranging from video games to Japanese culture, politics, and adult content. Users create new threads by posting an image to a board along with a message; other users can reply by posting messages and/or images. We focus on the Politically Incorrect board (/pol/), which is known for the high volume of offensive content and very loose moderation. We do so as prior

Source	Dates	#Posts	#YouTube Links	#YouTube Comments
/pol/	2016–2019	134.5M	850,523	1B
r/The_Donald	2015–2019	4.8M	278,849	63.7M
r/Braincels	2017–2019	216,806	7,902	16.4M
r/ForeverAlone	2010–2019	134,723	4,427	5.3M
r/Incels	2014–2017	54,218	2,020	2.9M
r/IncelTears	2017–2019	61,765	923	583,057
r/IncelsWithoutHate	2017–2019	16,217	491	402,075
r/ForeverUnwanted	2016–2019	1,898	58	19,666
r/BlackPillScience	2018–2019	1,129	44	6,187
r/gymcels	2018–2019	226	34	21,472
r/MaleForeverAlone	2017–2018	619	19	17,917
r/foreveraloneteens	2011–2019	322	18	3,592
r/Incelselfies	2018–2019	6,385	18	15,552
r/Truecels	2015–2016	364	17	5,900
r/ForeverAloneDating	2011–2019	76,976	11	605
r/askanincel	2018–2019	2,465	10	2,056
r/IncelDense	2018–2019	254	7	3,410
r/SupportCel	2017–2019	352	6	2,253

Table 1: Number of videos retrieved from each data source.

work [23] has uncovered attempts of coordinated aggression attacks targeting YouTube users.

r/The_Donald was a subreddit created in 2015 in support of Donald Trump’s 2016 U.S. Presidential Election campaign. It was broadly linked to the alt-right movement, and rife with racist and sexist content [11]. In 2019, it was quarantined and restricted before being banned.¹ We choose this subreddit as previous work [17] analyzing the behavioral patterns of its active participants revealed frequent calls to action that ultimately lead to disruptive behavior.

Incels Subreddits. Incels (an abbreviation of Involuntary Celibates) are an online subculture of people who identify as unable to get a romantic or sexual partner despite desiring one. They are part of a larger collection of groups loosely organized around “men’s rights,” known as the Manosphere [19], which are often associated with promoting masculism [5]. Incels mainly focus on sexual deprivation, which they blame on their unattractive appearance. They believe that women are attracted to men with specific facial attributes and racial backgrounds, a lack of which leads to celibacy.

We choose the Incels community as prior work studying the Manosphere [43] shows it is highly engaged and produces a high volume of hateful speech. We start from the 19 Incels subreddits identified by Papadamou et al. [37], but ultimately work with 16 as the remaining three do not have any YouTube links.

3.2 Data Collection

Our methodology involves two steps: 1) retrieving all links to YouTube videos posted on /pol/, r/The_Donald, and the 16

¹Quarantined subreddits do not generate revenue and, among other things, require users to explicitly opt-in to viewing the content; subreddits placed in restricted mode prevent most users from creating new posts.

Incels subreddits, and 2) gathering all comments and replies to these YouTube videos. For /pol/, we use the dataset released by Papasavva et al. [39], which contains 3,397,911 threads and 134,529,233 posts posted between June 2016 to November 2019. For Reddit, we first gather all public data posted on Reddit from 2005 to 2020, which includes 600M posts and 5B comments from 2.8M subreddits, using the Pushshift dumps [58]. Then, we filter out all comments and posts made on r/The_Donald and the 16 Incels subreddits.

We extract all links to YouTube videos posted on these communities using regular expressions.² For /pol/, we use the ‘com’ field of the JSON object containing all posts from a single thread, while, for Reddit, the ‘url’ field from the JSON object corresponding to a post. The total number of videos posted on 4chan is 850,523, 278,849 on r/The_Donald, and 14,616 on the Incels subreddits.

Finally, we use the YouTube Data API [56] to collect all comments and replies. We obtain 1B YouTube comments from /pol/ links, 63.7M from r/The_Donald, and 24.4M from the Incels subreddits. An overview of our dataset is available in Table 1.

4 Characterizing Coordinated Attacks

Before we can build machine learning models for coordinated hate attacks attribution, we need to first of all “understand” them. We start by formulating two hypotheses, namely, that coordinated online attacks will share two distinctive characteristics: 1) once a YouTube link is shared on a platform, there might be a spike in its commenting activity on YouTube, which is a possible indication of a coordinated attack occurring; and 2) each community has its lingo and slang words (e.g., /pol/’s slang includes characteristic terms like “cuck” and “libtard”), which can be modeled and used as features in a machine learning classifier. If these assumptions hold, coordinated hate attacks could be traced back to a source community by looking for spikes in comments and matching the language of the community with the video comments.

To test the validity of our hypotheses, we run an experiment with videos posted on /pol/, r/The_Donald, and the 16 Incels subreddits, between January and June 2019 (when r/The_Donald got quarantined). We use a dataset consisting of 17,023 videos from 4chan’s /pol/ board, 2,499 from r/The_Donald, and 248 from the 16 Incels subreddits. We pick videos that were only linked in one community to avoid overlaps (i.e., to have a single ground truth). We do this to exclude confounding variables and treat each community as a distinct class. This approach allows us to establish clear relationships between each source community and their possible

²We find 5 different kinds of YouTube links: 1) youtube.com/watch?v=video_id, 2)youtu.be/video_id, 3) m.youtube.com/watch?v=video_id, 4) m.youtu.be/video_id, 5) youtube.com/embed/video_id, which are captured by the regular expressions.

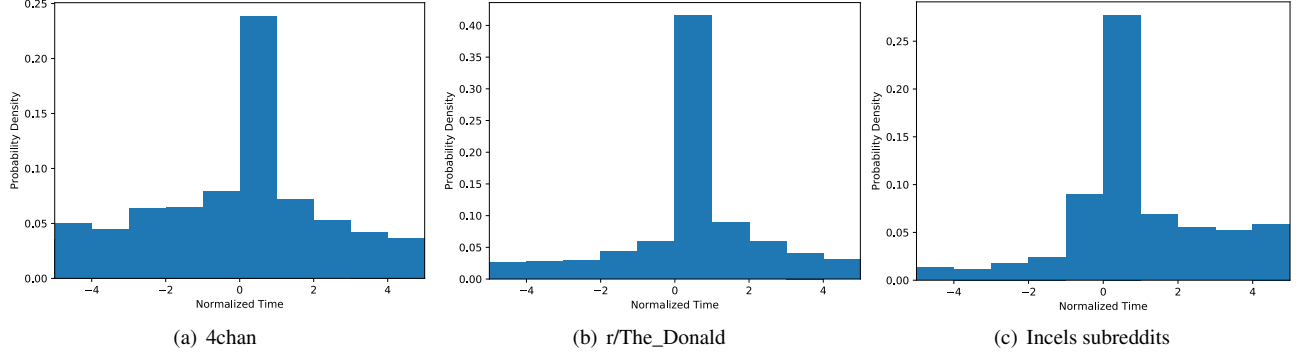


Figure 1: Probability Density Functions (PDFs) of the activity peak in YouTube comments and the source community thread where the YouTube video is linked from. The time is normalized to the thread’s lifetime, where $t = 0$ denotes the time when the video was first mentioned, and $t = 1$ is the last post in the thread.

attacks.

4.1 Commenting Activity

We first want to measure whether after a URL to a third platform (e.g., a YouTube video) is posted on an online community (e.g., 4chan’s /pol/), this third-party content observes a peak in commenting activity. Preliminary observations in this direction were made by Hine et al. [23]. If a video linked to on a Web community receives a surge of comments during the lifetime of the discussion thread, this might indicate that a coordinated hate attack taking place.

More formally, let x be the thread on which the link to the YouTube video is posted, and y the set of comments on the video. We denote the timestamps in x and y as $\{t_x^i \mid 1, \dots, N_x\}$ and $\{t_y^j \mid 1, \dots, N_y\}$. We then normalize the time frame for $\{t_x^i\}$ and $\{t_y^j\}$ so that $t = 0$ represents the time the YouTube link is posted on the source community, and $t = 1$ the time the last post is made on that thread. The normalized time frame is then calculated as: $t = \frac{t - t_{yt}}{t_{last} - t_{yt}}$.

In Figure 1, we plot the Probability Density Function (PDF) of the (normalized) YouTube comment timestamps for each of the three communities. The distribution shows the YouTube commenting activity with respect to the thread lifespan from where the video is linked. For each community, the highest peak in comment activity occurs between $t = 0$ and $t = 1$, which is a possible indication that an attack occurs once the link is posted on a community. This aligns with the results in Hine et al. [23], who find that a rise in commenting activity of a YouTube video is a possible indication of a raid taking place.

4.2 Language

Our second hypothesis is that, since online communities are characterized by their own jargon, the comments originating from a certain community will present linguistic features that are closer to that community. To validate this assumption, we first pre-compute the TF-IDF (Term Frequency – Inverse Document Frequency) scores of all the words in the three source communities. We calculate the TF-IDF score for each word by computing TF on the given community and the IDF

Keywords	Accuracy	Keywords	Accuracy
Top-10	278/1,176 (23.6%)	Top-12	290/1,176 (24.7%)
Top-14	292/1,176 (24.8%)	Top-16	293/1,176 (24.9%)
Top-18	287/1,176 (24.4%)	Top-20	294/1,176 (25.0%)
Top-22	287/1,176 (24.4%)	Top-24	278/1,176 (23.6%)

Table 2: Percentage of videos posted on /pol/ for which the language in their comments is closest with the source community, with different keyword combinations.

on the other two communities. This method ensures that the score for any word used in a community shows its importance relative to the other two communities. For each video, we extract comments between $t = 0$ and $t = 1$ and calculate a TF-IDF score for each word.

Next, to understand if the comments in the video are closest to their source community, we use the keywords with the highest TF-IDF score. I.e., we calculate the average score of the top keywords, and for each of the three source communities, we compute the average score of the same words, finding the community with the closest language.

We experiment with various thresholds to identify the ideal number of keywords, ranging from Top-10 to Top-24. For this experiment, we use videos linked from /pol/ because we expect it to be the most complex and varied in terms of language. In Table 2, we report the fraction of videos posted on /pol/ for which the language of the comments in the peak is closest to the one by the community itself, using various threshold selections. We stop at Top-24 keywords because the accuracy does not increase any further, and find that Top-20 yields the highest accuracy.

After fixing the keyword threshold to 20, we run this approach on all videos in our dataset, reporting our results in Table 3; while 85% of the videos posted on Incel subreddits have a language that is closest to the source community, there is definitely room for improvement on r/The_Donald (45%) and /pol/ (25%).

Model selection. We also investigate whether more sophisticated language learning models would lead to better results. More precisely, we experiment with BERT (Bidirectional En-

Community	TF-IDF	S-BERT
/pol/	294/1,176 (25%)	624/1,176 (53%)
r/The_Donald	427/985 (43%)	97/985 (10%)
Incels subreddits	69/81 (85%)	35/81 (43%)

Table 3: Accuracy of attributing attacks on videos linked from each source community.

coder Representations from Transformers) [14], since it is considered the state-of-the-art model for a wide range of language processing tasks like text classification [14], named entity recognition [28], text summarization [29], etc. We use PyTorch’s SentenceTransformer library to load a pre-trained Sentence-BERT (SBERT) [41] model, which uses siamese and triplet network structures to produce sentence embeddings that can be compared using cosine-similarity. We encode the text from each source community into vector representations. Next, we take the videos discussed earlier in this section and encode their comments into a vector representation. Then, we compute the cosine similarity between the encoded representations of YouTube comments and each source community. Finally, we determine the most similar community based on the highest cosine similarity score.

Table 3 shows that, cumulatively, the TF-IDF-based approach outperforms BERT, correctly attributing 790 out of 2,242 videos, while the BERT-based model correctly attributes 756. In particular, on r/The_Donald, it only attributes 97 out of the 985 videos correctly. Therefore, we select TF-IDF due to its simplicity, consistency, lower resource overhead, and better performance.

4.3 Controlling for Lag

To improve accuracy, we opt to control for the lag and the number of comments on the YouTube video. As hypothesized by [23], in the case of an attack, the comments of the source community and the YouTube video are likely to be synchronized; hence, the lag is close to 0. To estimate the lag between the two signals (i.e., the posts on the source community and the comments on the YouTube video), we use cross-correlation and run a grid search to identify the best lag range and the number of comments on the YouTube video.

For /pol/, we get the highest accuracy (71%) with a comment range of [42-72] and a lag of 0. For r/The_Donald and the Incels subreddits, the best settings are, respectively, with a comment range of [65-100] and a lag range of [0-1], and a comment range of [40-100] without any lag restriction, yielding, respectively, 85% and 100% accuracy.

Overall, our analysis shows that once a YouTube link is shared on a polarized community, there is a spike in comments on the video, and the language of these comments is closest to the source community. We use these observations to design our raid attribution system, TUBERAIDER.

5 TUBERAIDER

In this section, we present the three main components of the TUBERAIDER system, namely, 1) language pre-training, 2) peak detection, and 3) attribution. TUBERAIDER’s pipeline is also depicted in Figure 2: TUBERAIDER first learns the language used in all the source communities presented to it; then, for each YouTube video shared in these communities, it detects peaks in their commenting activity. Finally, it uses a machine learning classifier to attribute the comments during the peak to one of the source communities.

TUBERAIDER only requires two sets of timestamped messages, one from the source community (e.g., a /pol/ thread) and one from the target community (e.g., comments on a YouTube video). Therefore, our approach could be easily adapted to other services as well.

5.1 Language Pre-Training

To model the language used by different communities when carrying out aggression attacks, we use the methodology discussed in Section 4.2 and perform a TF-IDF analysis to identify the words that stand out the most from the attacks by a certain community.

We pre-compute TF-IDF scores of all words in the three source communities. We calculate the TF-IDF score for each word by computing TF on the given community and the IDF on the other two communities.

5.2 Peak Detection

Next, we detect peaks in the commenting activity of a video, which we can later attribute to a source community based on linguistic features. From Section 4.1, we know that there is a surge in YouTube video comments once a link to it is posted on a community. Therefore, TUBERAIDER relies on a peak detection module to identify deviations in comment activity that can be tested for potential coordinated attacks.

To identify peaks, we calculate the daily mean number of comments to the video, along with the standard deviation, for all the days we have comments for a specific video. Whenever TUBERAIDER encounters a day or period with more comments than the sum of daily mean and standard deviation, it labels it as *a period with a peak* in the commenting activity of a video, and thus worth to be examined.

5.3 Attribution

Once a peak is identified, we extract all comments during the relevant time range. TUBERAIDER trains a multi-class classifier to attribute a given attack to the source community. Using the TF-IDF scores obtained from Section 5.1, we extract the Top-20 keywords from each source community. The TF-IDF scores for the selected words serve as the features of a given video. In other words, each video has 60 features (i.e., 20 words from each community).

Then, we calculate a TF-IDF score for each word using the comments during the peak of each YouTube video. For words that do not appear in the comments of the video, their

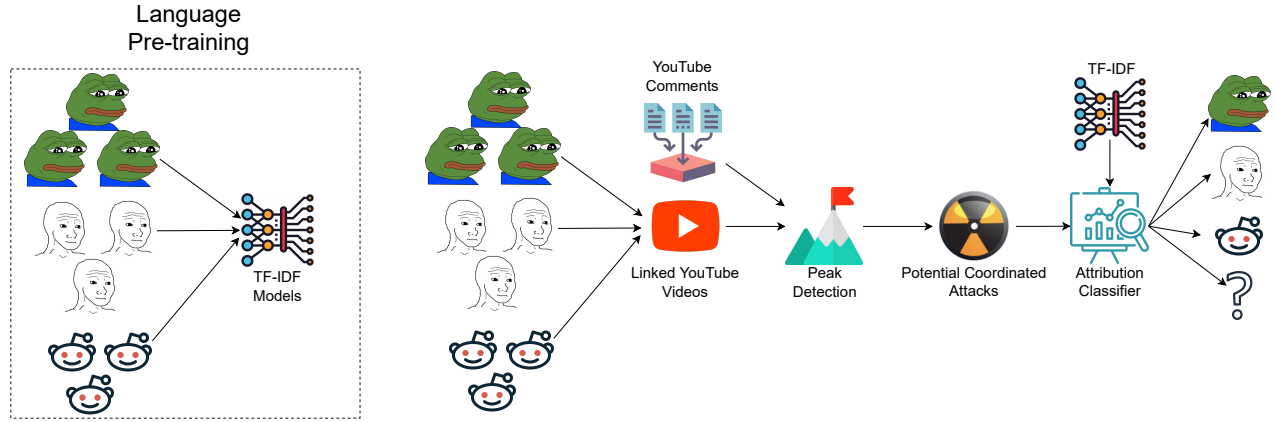


Figure 2: Overview of TUBERAIDER: a set of communities are fed to the system, in this case: (1) 4chan’s /pol/ board, (2) r/The_Donald subreddit, and (3) 16 Incels subreddits. TUBERAIDER learns their language through TF-IDF on the top keywords. To attribute potential attacks, it collects all YouTube comments on videos linked from each source community and identifies peaks in the comment activity of these videos as an indication of a potential coordinated attack. Finally, TUBERAIDER attributes attacks back to a source community using a machine learning classifier based on the TF-IDF scores of top keywords.

	/pol/	T_D	Incels		/pol/	T_D	Incels
1	fuck	trump	incel	11	know	cnn	make
2	nigger	peopl	women	12	kike	fake	thing
3	don	think	men	13	now	need	date
4	white	news	girl	14	faggot	right	life
5	peopl	make	guy	15	us	fuck	chad
6	jew	vote	one	16	time	hillari	realli
7	shit	know	sex	17	right	even	friend
8	make	time	feel	18	good	presid	good
9	want	want	person	19	countri	thing	someone
10	think	content	attract	20	trump	clinton	relationship

Table 4: Top-20 keywords with the highest TF-IDF scores for each source community. (T_D denotes r/The_Donald).

feature value is set to 0. Based on the TF-IDF scores, TUBERAIDER trains a supervised model to identify the community which initiated the attack on the given YouTube video. The model trains on a set of labeled videos, where the label is a source community. On an unseen video, TUBERAIDER outputs one of four labels, corresponding to Incels subreddit, r/The_Donald, /pol/, or that the video cannot be attributed to any of the three source community.

Overall, our attribution classifier allows us to determine with high confidence the community that launched a raid without looking at the context of the text being discussed in the comment thread. We believe this is preferable to simply checking whether or not a link has been shared on a platform; it is also less prone to content drift since, while the topics being discussed might change, we expect the language used by entire communities to change at a slower pace.

6 Evaluation

In this section, we present the results of our experimental evaluation for each component of TUBERAIDER’s pipeline.

6.1 Pre-Training

As discussed, we perform a TF-IDF analysis on all three source communities. We find a total of 1,610,197 unique words on /pol/, 401,738 on r/The_Donald, and 144,252 unique words on the Incels subreddits after removing stop words and using Porter Stemmer [40] to find the word stems. This difference in the number of words used in each community is a relatively strong indication of the difference in the language used in each community.

Table 4 shows the Top-20 keywords for each source community; r/The_Donald includes more political keywords, /pol/ a lot of racist and abusive words, and the Incel subreddits more dating-related keywords.

6.2 Peak Detection

We detect peaks in YouTube videos as per the methodology in Section 5.2. As illustrated in Figure 2, TUBERAIDER feeds the comments from the peak to the attribution module. Depending on the length and comments of the video, a number of peaks may be identified in any given video.

From the 20,325 videos linked from all communities in the first six months of 2019, we find that 562 videos do not have any peaks. The other 19,763 videos have at least a peak—more precisely, 17.5 peaks per video on average. To optimize the attribution, TUBERAIDER only considers as optimal peaks that have a certain number of comments, which we denote as “Minimum Comments Threshold.”

We believe this threshold is best treated as a hyperparameter and thus tune it as part of the experiments presented in Section 6.3, aiming to find the optimal value which produces the best classification results (see Figure 3).

6.3 Attribution

To create our attribution pipeline, we need a set of videos that have not been linked from one of the three source communities. Therefore, we build a dataset of 50 unrelated YouTube videos. We use the YouTube Data API search feature to query

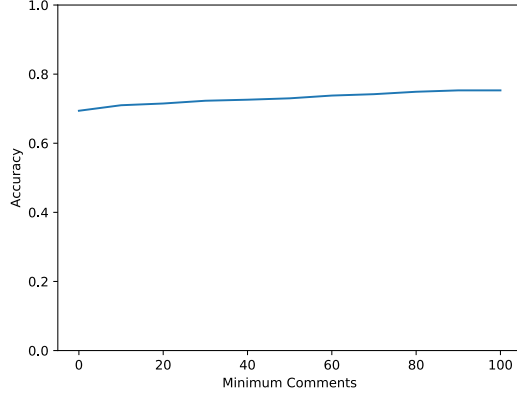


Figure 3: Attribution accuracy for various "Minimum Comments" thresholds.

videos for the following keywords: sports, gaming, and music.

To train our system, we curate a set of 200 total videos, 150 of which are linked from either of the three source communities, selected at random: 1) /pol/, 2) r/The_Donald, and 3) the Incels subreddits. These 150 videos are randomly sampled from our dataset from the first six months of 2019 (before r/The_Donald got quarantined). The other 50 is the set of unrelated YouTube videos, i.e., not linked from our source communities. We perform ten-fold cross-validation and evaluate the performance of four classifiers based on accuracy, precision, recall, and F1 score; please see Table 5. We find that Random Forest yields the best performance, with 75.3% accuracy in correctly attributing a video to its source community.

As mentioned in Section 6.2, we treat the number of comments required to be considered a peak as a hyper-parameter. We experiment with various thresholds by linearly increasing the minimum number of comments by ten during training and analyzing its impact on attribution accuracy. The resulting accuracy for each minimum comment threshold is reported in Figure 3: there is a steady increase in accuracy until 90 comments, and then it levels off. As a result, we set the minimum comments threshold to 90, which gives us a total of 221 peaks for 200 videos.

Comment Threshold Analysis. To ensure that enforcing a minimum comment threshold does not prevent TUBERAIDER from attributing real-world aggression attacks, we manually analyzed a set of videos with under 90 comments to attribute potential aggression attacks. We picked 33 videos linked from each source community and had two annotators independently assess the comments for hateful, aggressive, toxic, and inflammatory speech markers.

Annotator 1 found two potential aggression attacks, while Annotator 2 found one, in common with Annotator 1. This yields a Cohen’s Kappa score [53] of 0.662 (high agreement). As a result, we estimate an upper-bound error rate of approximately 2% (Annotator 1 found 2/99 videos to be potential aggression attacks). This confirms the validity of consider-

Classifier	Accuracy	Precision	Recall	F1-Score
Linear SVM	34.0%	10.0%	29.2%	19.6%
Decision Tree	62.8%	49.3%	47.1%	48.2%
KNN	63.5%	54.9%	55.5%	55.2%
Random Forest	75.3%	70.5%	67.8%	69.2%

Table 5: Performance of the different classifiers using ten-fold cross validation.

Predicted\Actual	Incels	r/The_Donald	/pol/	Unrelated
Incels	9	0	0	1
r/The_Donald	0	9	1	0
/pol/	1	2	6	1
Unrelated	0	0	1	9

Table 6: Confusion Matrix of our classification model.

ing the threshold as a hyper-parameter that can be fine tuned as necessary.

Next, we compute a confusion matrix to understand how well our system labels videos of each class, in particular, the unrelated videos. Class 1 corresponds to the Incels subreddits, 2 to r/The_Donald, 3 to /pol/, and 4 to the unrelated videos. As shown in Table 6, our system can correctly identify the unrelated videos as well.

6.4 Misclassified Videos

As is common with classification tasks, our attribution module misclassifies a (relatively small) fraction of the videos. Using the confusion matrix in Table 6, we observe that the False Positive Rate (FPR) for the Incels subreddits is 0.033 and 0.067 for the other three classes. The highest percentage of false negatives is on /pol/, with a False Negative Rate (FNR) of 0.4; all the other classes have an FNR of 0.1.

In the rest of this section, we shed light on the misclassifications and present a few relevant examples. To make things easier, we divide the misclassified videos into two categories: 1) Crossover Videos and 2) No Aggression Videos.

Crossover Videos. A crossover video features concepts more relevant to another community than the one it was linked from, e.g., a video discussing President Donald Trump on an Incels subreddit. A very small minority of these videos, especially those showing heavy overlap with another community, are misclassified because the comments dilute the specificity of the language of the original source community. For instance, the video “Mick Mulvaney on Trump’s booming economy” is linked from 4chan, but being related to politics and President Donald Trump, the commenting activity uses a language that is more related to those topics and more similar to r/The_Donald’ lingo.

This is arguably a limitation inherent to language modeling. TUBERAIDER uses TF-IDF values as features for attribution; future work could investigate the use of different language model tools to be less susceptible to this phenomenon.

No Aggression Videos. The second kind of misclassified videos are those not attacked by the source community. A

Perspective	Attr.	NonAttr.	Baseline	Ks 1	P-Value	Ks 2	P-Value
Toxicity	0.246	0.218	0.147	0.069	<0.0002	0.199	<0.0002
Severe Toxicity	0.035	0.032	0.015	0.040	<0.0002	0.171	<0.0002
Identity Attack	0.071	0.059	0.034	0.071	<0.0002	0.231	<0.0002
Insult	0.166	0.133	0.092	0.085	<0.0002	0.215	<0.0002
Profanity	0.142	0.136	0.087	0.026	<0.0002	0.131	<0.0002
Threat	0.036	0.036	0.021	0.021	<0.0002	0.117	<0.0002
Sexually Explicit	0.067	0.064	0.044	0.065	<0.0002	0.079	<0.0002
Flirtation	0.374	0.377	0.349	0.020	<0.0002	0.093	<0.0002
Attack on Author	0.148	0.149	0.144	0.042	<0.0002	0.062	<0.0002
Att. on Commenter	0.267	0.270	0.222	0.016	<0.0002	0.085	<0.0002
Incoherent	0.514	0.535	0.648	0.044	<0.0002	0.248	<0.0002
Inflammatory	0.391	0.341	0.270	0.088	<0.0002	0.225	<0.0002
Likely to Reject	0.674	0.660	0.606	0.022	<0.0002	0.121	<0.0002
Obscene	0.281	0.272	0.233	0.014	<0.0002	0.082	<0.0002
Spam	0.111	0.136	0.207	0.085	<0.0002	0.254	<0.0002
Unsubstantial	0.640	0.665	0.711	0.053	<0.0002	0.108	<0.0002
Rewire	Attr.	NonAttr.	Baseline	Ks 1	P-Value	Ks 2	P-Value
Abuse	0.293	0.241	0.226	0.061	<0.0002	0.082	<0.0002
Hate	0.157	0.132	0.098	0.086	<0.0002	0.138	<0.0002
Sexually Explicit	0.032	0.037	0.052	0.023	<0.0002	0.078	<0.0002
Profanity	0.111	0.106	0.071	0.006	0.301	0.040	0.001
Positive	0.168	0.205	0.198	0.075	<0.0002	0.133	<0.0002
Violent	0.039	0.034	0.020	0.025	<0.0002	0.105	<0.0002

Table 7: Toxicity scores for each video category.

large majority of them are music videos with many comments and views, linked from platforms like 4chan and Reddit simply for entertainment purposes, often not receiving any traction in the community. A relevant example is “Earth, Wind & Fire – Let’s Groove,” which has 49.4K comments on YouTube and 231M views as of October 4, 2022. Although it has been linked from 4chan, TUBERAIDER labels it as “unrelated” or, in other words, not attacked by any community, which means that the language in the detected peak(s) does not match the source community. Hence, we postulate that the peaks in videos shared for non-attack purposes are largely observed due to other factors, e.g., artists sharing the music video on their social media, a paid traffic boost, or promotions. However, just because it is coincidentally linked from one of the platforms does not imply that an actual attack is taking place. As a side note, in this case, the original thread on 4chan only received one comment. Therefore, it is safe to assume that in this case and other similar cases, there was no actual aggression attack.

6.5 Detection in the Wild

Finally, we test TUBERAIDER in the wild to attribute attacks to a given source community. We use the data collected in Section 3.2 and extract 3,333 videos per community, obtaining a total of 9,999 videos. We use this data as we have ground truth; i.e., we know which community a given video was linked from. TUBERAIDER attributes 700 videos out of the 9,999 to the labeled source community, while 8,644 videos did not pass our minimum peak threshold of 90 comments and were discarded.

We manually review the attributed videos to understand the attacks; in Section 7, we discuss a few examples of what these aggression attacks look like, from the original thread to

the YouTube video comments.

6.6 Raid vs Commenting Activity

One intuition behind TUBERAIDER is that *peaks in commenting activity* on YouTube videos, combined with the *use of language similar to that of hateful online communities*, strongly suggests the occurrence of raids. Both factors play a crucial role in this determination. To further validate this intuition, it is important to differentiate between benign commenting activity that can appear as coordinated and an actual raid. The latter involves real, targeted hate; the former means that the video is merely receiving attention of some kind (e.g., when a video link is shared across platforms without any malicious intent). Thus, we set out to analyze the commenting activity in videos attributed by TUBERAIDER and compare them to non-raided videos using various metrics that are indicative of raids (e.g., toxicity, abuse, hate, and inflammatory content).

We divide the videos into: 1) attributed videos or raids identified by TUBERAIDER in Section 6.5, 2) non-attributed videos from Section 6.5 that had peaks and passed the minimum comment threshold but were not classified as raids, and 3) 50 generic videos of various categories (e.g., music and gaming) from Section 6.3 that serve as our baseline. For each video, we randomly pick 50 comments from their peaks and feed them to Google’s Perspective API [20] and the Rewire API [42].³ Our intuition is that if the commenting peaks attributed by TUBERAIDER as raids show significantly higher values for those metrics, then this is a strong indication that they are raids.

The average scores for each metric from both APIs and the related video categories are reported in Table 7. Our baseline sample reveals that even generic videos contain some degree of hate speech and negativity. Due to the controversial nature of videos published on platforms like 4chan, we observe a noticeable increase in inflammatory markers for both attributed and unattributed videos. However, raided videos display the highest levels of toxicity, hate, and abuse, even though TUBERAIDER does not consider toxicity scores when performing attribution.

Significance Testing. We perform the Kolmogorov–Smirnov [54] test to determine whether the differences in scores are statistically significant. For each metric, we perform two tests: 1) between attributed and non-attributed videos, denoted as Ks-Score 1, and 2) between attributed and baseline videos, denoted as Ks-Score 2. We report these scores in Table 7, along with P-values. As we are testing multiple hypotheses at once, which increases the chance of rejecting the null hypothesis by chance, we use Bonferroni Correction [57], where we adjust the significance level α (α). We set the initial value of α to 0.01 and divide it by the number of hypotheses we are testing. We test 44 hypotheses (22 metrics and 2 hypotheses per metric),

³Perspective API is an open-source tool that assigns [0-1] scores to a given text across several metrics. Rewire is another machine learning-based tool that detects toxicity in text, developed in partnership with Nvidia, also assigning [0-1] confidence scores.

therefore the adjusted value of α is 0.0002. As a result, we only reject the null hypothesis if the P-value is <0.0002 . Overall, we can reject the null hypothesis for all metrics except “profanity,” i.e., the differences in all the other samples are statistically significant.

Implications. Note that Rewire defines abuse as “content that is insulting, aggressive, or threatening.” It defines hate as “abuse targeted at a protected group or at its members for being a part of that group, where protected groups are based on characteristics such as gender, race, or religion” and profanity as “a word or expression that is socially or culturally offensive, usually due to being obscene or explicit.” We postulate that TUBERAIDER is not a simple offensive speech detection tool, but it can also be used to attribute based on targeted abuse and hate, as shown by the difference in scores for the respective metrics and their associated significance.

7 Case Studies

To shed further light on coordinated attacks attributed by TUBERAIDER, we now discuss a few selected case studies. We do so aiming to analyze the modus operandi of the actors participating in the attacks with respect to spreading hate and how they are directed by the communities.

A common thread in these attacks is the aggressors often posting a controversial link that is relevant to the source community to direct hate and challenge the idea presented in the video. The examples we provide next show how posts in all three communities lead to aggression attacks that range from political to social issues.

7.1 r/The_Donald

One relevant example from r/The_Donald relates to the submission “Never Go Full Feinstein: DiFi Plans on Introducing New Anti-2A Bill (banning 18-20 year olds from purchasing any gun, Unconstitutional).” The submission has 1,351 upvotes and 63 comments. Roughly speaking, r/The_Donald sees a lot of activity from the alt-right community, which is usually vocal against any restrictions to gun ownership, and thus against this bill.

Here are some comments from the Reddit thread:

COMMENT 1: I hate that fuckin witch.

COMMENT 2: She was born in the 1920’s and is still alive. She’s probably being given baby blood transfusions or something.

COMMENT 3: So your old enough to lay your life on the line in the military to defend these fucks but not old enough to own a gun for you or your families protection...I see...

The first two are directed to Senator Feinstein, and one of those about her being a “witch” was also later found on the YouTube video as shown in the next snippet. The comments below are a short snippet of comments made on YouTube during the time this thread was active on Reddit. We remove any information that identifies the posters in the YouTube comments.

COMMENT 1: Wicked witch of the west

COMMENT 2: We have to do something, even if it doesn’t work. So people think that we are doing something.

COMMENT 3: She can introduce all the anti-2A bills she wants. I will defy and ignore every single one.

The discussion on Reddit was largely against the bill and Feinstein. We observe the same trend in comments around that time period on the YouTube video. By looking at the comments on the submission and the snippet from the comments on the YouTube video, we can better understand how these communities push their ideas and spread hate speech at the same time. The comments on the YouTube video, like those on Reddit, are also against the bill and Sen. Feinstein.

7.2 Incels

Next, we discuss a post made in the “IncelTears” subreddit. The post includes a YouTube video link to an old music video from the Donna Reed Show. The video is shot in a traditional setting with a woman singing songs and wearing a sundress typical of the 1950s and 60s. The post on Reddit, as shown below, paints the scene in two ways: (1) how women show much more skin in their dressing and are less feminine in today’s day and age, and (2) women date men based on “hypergamy,” in other words, they date men of a higher status than themselves while rejecting lower value males.

It’s over for Incels: Incels love to talk about feminism causing ‘hypergamy,’ but this scene from the pre-second-wave-feminist Donna Reed Show, set in an wholesome incel fantasy world, features a song that is literally about a girl who rejects all other boys because she only wants Chad

The comments on Reddit, a few of which are shown below, carry the same narrative as the post. The whole idea is that women were far prettier and more feminine back in the day, as shown in comments 2 and 3. However, the notion that women’s dating preferences are based on “hypergamy” is highlighted in the first comment, where the user complains about his mom rejecting low-value males the same way all females do.

COMMENT 1: Yeah that’s like the least problematic thing my mom did by far.

COMMENT 2: Women back then were prettier, in my humble opinion.

COMMENT 3: Feminism should have been stopped at the first wave.

The comments on YouTube from the same time as the Reddit post are shown below. While the general comments on the YouTube video are in praise of the singer and the TV show in general. The comments from the time of the Reddit post carry the same narrative of women becoming more masculine. Comment 1 and 2 clearly show the users being enraged that women “belong in the kitchen” and that they used to be feminine back in the day. The last comment goes to show that obscene language is often a characteristic of these aggression attacks.

COMMENT 1: Why would she go to college? You don't need a college education to clean house and cook.

COMMENT 2: Such a beautiful girl, and a voice to match..Gives me goose bumps to think that girls back then were feminin..No ugly tattoos, no fking and blinding, trying to act like men etc..Take me back there..

COMMENT 3: Using blown up condoms for gym decorations...even back then

7.3 /pol/

Finally, we discuss a post on 4chan's /pol/ board, where the poster shares a YouTube video link of a woman performing stand-up comedy. The author claims that women are not funny and cannot be good comedians.

Why the fuck are there no funny women on this planet? Why do they suck so much at comedy? Just look at this cringe

The post received 305 replies and some of the comments are highlighted below. Ostensibly, the posters believe that women are not smart, are incapable of understanding and appreciating humor, and that there is no biological incentive for them to be funny.

COMMENT 1: Women aren't mentally capable of intelligent humour, why would they?

COMMENT 2: Because women really aren't very smart.

COMMENT 3: They have no evolutionary/reproductive incentive for being funny.

The comments below are taken from YouTube at the time that the video was shared on 4chan. The same idea is pushed in the comments that women are inherently not funny.

COMMENT 1: Has there ever been a funny woman?

COMMENT 2: female comedian

COMMENT 3: That classic tactic, when the audience won't laugh at my jokes, passively aggressively bitch at them because clearly it's their fault.

7.4 Remarks

Overall, these case studies illustrate how coordinated attacks are carried out in various ways, depending on the ideology of the attacking community. In certain instances, as the r/The_Donald example demonstrates, a specific person is targeted with hate speech. In others, the targets are broader, for instance, with misogynistic and abusive words directed at women in general.

Moreover, the rhetoric used in these attacks can be exceptionally condescending and degrading, making them particularly harmful because they target actual individuals.

8 Discussion & Conclusion

In this paper, we presented TUBERAIDER, a generalizable system to attribute coordinated aggression attacks to the community that organized and carried them out. Our experimental analysis showed we can attribute attacks on YouTube

videos to a source community with an accuracy of over 75%. We demonstrated that coordinated attacks result in peaks in the activity of YouTube video comments and that the language used in the comments varies depending on the community the video was linked from. These linguistic traits enabled us to identify the communities where an attack originated. TUBERAIDER uses language features to train a machine learning classifier that attributes peaks in YouTube video comments to a source community. We also presented case studies of several aggression attacks identified by TUBERAIDER with overt elements of hate speech and misogyny to emphasize the rising issue.

8.1 Broader Perspective

Positive Implications. Effective and efficient attribution of coordinated hate attacks enables more nuanced and context-dependent moderation strategies. For instance, if the purpose behind an attack is believed to be racism, the platform could, in addition to deleting the hateful content, point victim to resources relevant to dealing racism. Attribution can also be used in soft moderation schemes (e.g., "this comment is very similar to those originating from 4chan"), which leverage the public's increasing understanding of the damage wrought by the darker parts of the Web.

Overall, TUBERAIDER can help platforms devise tailored policies. Previous work on personalized content moderation is limited because approaches either lack the required nuance [6] or solely rely on the manual selection of relevant keywords by content creators [25]. TUBERAIDER allows going beyond that, as the language models built for each source community can be used to automatically generate blocklists of keywords for attacks orchestrated by those communities.

An alternative deployment of TUBERAIDER could be to prioritize moderation of source communities based on how "dangerous" their attacks are. Since human vetting is the bottleneck of the moderation efforts by many online platforms, this prioritization could help quickly get to the attacks with the highest risk of causing damage to their victims.

Potential Negative Outcomes. Moderation techniques like TUBERAIDER are tantalizing in that their positive impact is pretty clear. At the same time, the worst-case scenario for deploying a moderation strategy is, arguably, not that it would just not work; rather, it might make the problem *worse* as deplatformed users move onto more extreme and echo-chamber-y platforms [2]. Thus, we need to consider several avenues that should be considered in more detail before TUBERAIDER should be deployed.

Malevolent actors who wish to coordinate hate attacks might use TUBERAIDER to learn whether they are at "risk" of being attributed, and change their strategy until the system misclassifies them. However, raids are conducted by several independent attackers stemming from the same platform, and the level of tight coordination required to evade TUBERAIDER would be difficult to achieve.

Moreover, false positives from TUBERAIDER are not entirely different from accusing a community of a crime they

did not commit. This has obvious problems from a “justice” point of view, but it also has the potential to stymie reformation processes that might be occurring within the attacking community (whether organic or due to some other mitigation). While it is true that 4chan is the source of many coordinated aggression attacks, claiming that they were responsible for attacks they did take part in only serves to reinforce some of the conspiratorial narratives that motivate the attacks in the first place.

Finally, we make no recommendation as to what specific moderation action(s) should be taken following the attribution of a coordinated hate attack, but at least some choices could open up additional ways to execute attacks. For example, if a deployment were to automatically provide a victim of an attack with resources, a naive implementation might end up allowing attackers to flood the victim with messages linking to those resources, which could itself be triggering. Even though this is a trivial example, and one that we would argue is still an improvement over the status quo, it does highlight that even effective solutions to the problem of coordinated hate attacks (and other content moderation tasks) have rough edges.

Ethics Considerations. Since we do not work with human subjects and only use data available to the public, our work is not categorized as human subjects research by our institution’s Institutional Review Board (IRB). We follow common ethical standards; we do not attempt to de-anonymize users, we remove all personally identifiable information from the case studies we report, etc.

8.2 Limitations and Future Work

Naturally, our system is not free from limitations. TUBERAIDER relies on language modeling (i.e., TF-IDF scores) to characterize communities; as discussed in Section 6.4, in case of high overlap in language across two communities, the system can misattribute a given aggression attack. TUBERAIDER also requires a certain number of comments to be made on a video before it considers it an aggression attack. Therefore, a relatively niche video with a small number of comments might not be considered by our system even though an attack might be taking place.

TUBERAIDER attributes attacks with an accuracy of 75%. Although our experiments show low false positive rates (at most 0.07), we envision TUBERAIDER being used to flag potential attacks that can be assessed manually or processed further. In other words, TUBERAIDER can act as a helpful warning indicator for a potential attack, helping to quickly identify and deter an orchestrated aggression attack.

We believe that TUBERAIDER can also be improved in several ways. First, TUBERAIDER relies on TF-IDF to attribute text to source communities; in the future, more expressive models can be used, e.g., embeddings, to better understand the context in which these words are used. Another interesting area of research could be investigating the characteristic patterns of the accounts that engage in coordinated aggression attacks and creating tools that detect and flag ac-

counts that belong to organized campaigns. Also, if a video gets traffic from a given community, it does not automatically mean an aggression attack is occurring; thus, future models could incorporate additional factors like the presence of hate speech and the toxicity of comments posted on the video as additional markers of a hate attack.

Finally, we select r/The_Donald, 4chan’s /pol/ board, and Incel subreddits for our research, partly due to their toxicity and tendency towards targeted hate attacks. However, future work should also examine online aggression attacks originating from other communities, perhaps from the other side of the “political spectrum,” since the communities we study in this paper are typically associated with the far-right.

Acknowledgments. This work was supported by the NSF through grants CNS-1942610, CNS-2114407, and CNS-2127232 and REPHRAIN: the UK’s National Research Centre on Privacy, Harm Reduction, and Adversarial Influence Online (UKRI grant EP/V011189/1).

References

- [1] Agarwal, S. and Sureka, A. (2014). A focused crawler for mining hate and extremism promoting videos on youtube. In *ACM HyperText*.
- [2] 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 *ACM WebSci*.
- [3] Alshamrani, S., Abuhamad, M., Abusnaina, A., and Mohaisen, D. (2020). Investigating online toxicity in users interactions with the mainstream media channels on youtube. In *CIKM (Workshops)*.
- [4] Bailey, K. (2021). Near, the Programmer Behind the Legendary BSNES Emulator, Has Died. <https://www.ign.com/articles/near-bsnes-remembrance>.
- [5] Blais, M. and Dupuis-Déri, F. (2012). Masculinism and the antifeminist countermovement. *Social Movement Studies*, 11.
- [6] Chandrasekharan, E., Gandhi, C., Mustelier, M. W., and Gilbert, E. (2019). Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction*, 3(CSCW).
- [7] Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., and Vakali, A. (2017). Hate is not binary: Studying abusive behavior of #gamergate on twitter. In *ACM WebSci*.
- [8] Chelms, C. and Yao, M. (2019). Minority report: Cyberbullying prediction on instagram. In *ACM WebSci*.
- [9] Chen, H., McKeever, S., and Delany, S. J. (2019). The use of deep learning distributed representations in the identification of abusive text. *ICWSM*, 13(01):125–133.
- [10] Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). Antisocial behavior in online discussion communities. In *ICWSM*.
- [11] Collins, K. (2020). Reddit bans controversial forum The_Donald as it revises hate speech policies. <https://www.cnet.com/tech/services-and-software/reddit-bans-controversial-forum-the-donald-as-it-revises-hate-speech-policies/>.
- [12] Dadvar, M., Trieschnigg, D., and Jong, F. d. (2014). Experts

- and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian Conference on Artificial Intelligence*.
- [13] Davidson, T., Warmley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.
- [14] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
- [15] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *The Web Conference (WWW)*.
- [16] ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. M. (2018). Peer to peer hate: Hate speech instigators and their targets. *arXiv:1804.04649*.
- [17] Flores-Saviaga, C., Keegan, B. C., and Savage, S. (2018). Mobilizing the trump train: Understanding collective action in a political trolling community. In *ICWSM*.
- [18] Giannakopoulos, T., Pikrakis, A., and Theodoridis, S. (2010). A Multimodal Approach to Violence Detection in Video Sharing Sites. In *ICPR*.
- [19] Ging, D. (2019). Alphas, betas, and incels: Theorizing the masculinities of the manosphere. *Men and Masculinities*, 22(4):638–657.
- [20] Google (2020). Perspective API. <https://www.perspectiveapi.com>.
- [21] Grigg, D. W. (2010). Cyber-Aggression: Definition and Concept of Cyberbullying. *Australian Journal of Guidance and Counselling*, 20(2).
- [22] Hernandez, N., Rahman, M., Recabarren, R., and Carbanar, B. (2018). Fraud de-anonymization for fun and profit. In *ACM CCS*.
- [23] Hine, G. E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., and Blackburn, J. (2017). Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan’s Politically Incorrect Forum and Its Effects on the Web. In *ICWSM*.
- [24] Jaki, S., De Smedt, T., Gwóźdz, M., Panchal, R., Rossa, A., and De Pauw, G. (2019). Online hatred of women in the Incels.me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, 7(2).
- [25] Jhaver, S., Chen, Q. Z., Knauss, D., and Zhang, A. X. (2022). Designing word filter tools for creator-led comment moderation. In *ACM CHI*.
- [26] Kumar, S., Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2018). Community interaction and conflict on the web. In *The Web Conference (WWW)*.
- [27] Kwon, K. and Gruz, A. (2017). Is offensive commenting contagious online? examining public vs. interpersonal swearing in response to donald trump’s youtube campaign videos. *Internet Research*, 27.
- [28] Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. In *NeurIPS*.
- [29] Liu, Y. (2019). Fine-tune BERT for extractive summarization. *arXiv:1903.10318*.
- [30] Marathe, S. and Shirsat, P. (2015). Approaches for mining youtube videos metadata in cyber bullying detection. *International Journal of Engineering Research and Technology*, 4.
- [31] Mariconti, E., Suarez-Tangil, G., Blackburn, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Serrano, J. L., and Stringhini, G. (2019). “You Know What to Do”: Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. In *ACM CSCW*.
- [32] Moor, P. J., Heuvelman, A., and Verleur, R. (2010). Flaming on youtube. *Computers in Human Behavior*, 26(6).
- [33] Nilizadeh, S., Labrèche, F., Sedighian, A., Zand, A., Fernandez, J., Kruegel, C., Stringhini, G., and Vigna, G. (2017). Poised: Spotting twitter spam off the beaten paths. In *ACM CCS*.
- [34] Olteanu, A., Castillo, C., Boy, J., and Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. *arXiv:1804.05704*.
- [35] Ottoni, R., Cunha, E., Magno, G., Bernardina, P., Jr., W. M., and Almeida, V. A. F. (2018). Analyzing right-wing youtube channels: Hate, violence and discrimination. *arXiv:1804.04096*.
- [36] Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B., Flammini, A., and Menczer, F. (2021). Uncovering coordinated networks on social media: Methods and case studies. *ICWSM*, pages 455–466.
- [37] Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., and Sirivianos, M. (2021). “How over is it?” Understanding the Incel Community on YouTube. *Proceedings of the ACM on Human Computer Interaction*, 5(CSCW2).
- [38] Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., and Sirivianos, M. (2022). “it is just a flu”: assessing the effect of watch history on youtube’s pseudoscientific video recommendations. In *ICWSM*, pages 723–734.
- [39] Papasavva, A., Zannettou, S., De Cristofaro, E., Stringhini, G., and Blackburn, J. (2020). Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. In *ICWSM*.
- [40] Porter, M. (2006). The Porter Stemming Algorithm. <https://tartarus.org/martin/PorterStemmer/>.
- [41] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*.
- [42] Rewire (2023). <https://rewire.online/>.
- [43] Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., Greenberg, S., and Zannettou, S. (2021). The evolution of the manosphere across the web. In *ICWSM*.
- [44] Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., and Meira Jr, W. (2018). Characterizing and detecting hateful users on twitter. In *ICWSM*.
- [45] Salminen, J., Almerexhi, H., Milenković, M., Jung, S.-g., An, J., Kwak, H., and Jansen, B. J. (2018). Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *ICWSM*.
- [46] Sharma, K., Zhang, Y., Ferrara, E., and Liu, Y. (2021). Identifying coordinated accounts on social media through hidden influence and group behaviours. In *ACM KDD*.
- [47] Stringhini, G., Kruegel, C., and Vigna, G. (2010). Detecting spammers on social networks. In *ACSAC*.
- [48] Stringhini, G., Mourlanne, P., Jacob, G., Egele, M., Kruegel, C., and Vigna, G. (2015). Evilcohort: Detecting communities of malicious accounts on online services. In *USENIX Security Symposium*.
- [49] Sureka, A., Kumaraguru, P., Goyal, A., and Chhabra, S. (2010). Mining youtube to discover extremist videos, users and

- hidden communities. In *Asia Information Retrieval Symposium*.
- [50] Tahir, R., Ahmed, F., Saeed, H., Ali, S., Zaffar, F., and Wilson, C. (2019). Bringing the kid back into youtube kids: Detecting inappropriate content on video streaming platforms. In *ASONAM*.
 - [51] Tucker-McLaughlin, M. (2013). Youtube’s most-viewed videos: Where the girls aren’t. *Women & Language*, 36(1).
 - [52] Weaver, A. J., Zelenkauskaitė, A., and Samson, L. (2012). The (non)violent world of youtube: Content trends in web video. *Journal of Communication*, 62(6).
 - [53] Wikipedia (2022a). Cohen’s kappa. https://en.wikipedia.org/wiki/Cohens_kappa.
 - [54] Wikipedia (2022b). Kolmogorov-Smirnov test. https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test.
 - [55] Wotanis, L. and McMillan, L. (2014). Performing gender on youtube. *Feminist Media Studies*, 14(6).
 - [56] YouTube Developers (2022). YouTube Data API. <https://developers.google.com/youtube/v3/>.
 - [57] Zach (2021). The Bonferroni Correction Definition & Example. <https://www.statology.org/bonferroni-correction/>.
 - [58] Zannettou, S., Baumgartner, J., Keegan, B., Squire, M., and Blackburn, J. (2020). The Pushshift Reddit Dataset. In *ICWSM*.
 - [59] Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Suarez-Tangil, G. (2018). On the Origins of Memes by Means of Fringe Web Communities. In *ACM IMC*.
 - [60] Zhao, Y., Xie, Y., Yu, F., Ke, Q., Yu, Y., Chen, Y., and Gillum, E. (2009). Botgraph: Large scale spamming botnet detection. In *USENIX Security Symposium*.