YouTubers Not *madeForKids*: Detecting Channels Sharing Inappropriate Videos Targeting Children

Myrsini Gkolemi
FORTH & University of Crete
Greece

Panagiotis Papadopoulos
Telefonica Research
Spain

Evangelos P. Markatos
FORTH & University of Crete
Greece

Nicolas Kourtellis
Telefonica Research
Spain

ABSTRACT

YouTube is one of the most popular social media and online video sharing platforms, and users turn to it for entertainment by consuming music videos, for educational or political purposes, advertising, etc. In the last years, hundreds of new channels have been creating and sharing videos targeting children, with themes related to animation, superhero movies, comics, etc. Unfortunately, many of these videos have been found to be inappropriate for consumption by their target audience, due to disturbing, violent, or sexual scenes.

In this paper, we study YouTube channels that were found to post *suitable* or *disturbing* videos targeting kids in the past. Unfortunately, we identify a clear discrepancy between what YouTube assumes and flags as inappropriate content and channel, vs. what is found to be disturbing content and still available on the platform, targeting kids. In particular, we find that almost 60% of videos that were manually annotated and classified as *disturbing* by an older study in 2019 (a collection bootstrapped with *Elsa* and other keywords related to children videos), are still available on YouTube in mid 2021. In the meantime, 44% of channels that uploaded such *disturbing* videos, have yet to be suspended and their videos to be removed. For the first time in literature, we also study the "madeForKids" flag, a new feature that YouTube introduced in the end of 2019, and compare its application to the channels that shared *disturbing* videos, as flagged from the previous study. Apparently, these channels are less likely to be set as "madeForKids" than those sharing suitable content. In addition, channels posting *disturbing* videos utilize their channel features such as keywords, description, topics, posts, etc., in a way that they appeal to kids (e.g., using game-related keywords). Finally, we use a collection of such channel and content features to train machine learning classifiers that are able to detect, at channel creation time, when a channel will be related to *disturbing* content uploads. These classifiers can help YouTube content moderators reduce such incidences, by pointing to potentially suspicious accounts, without analyzing actual videos, but instead only using channel characteristics.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
WebSci '22, June 26–29, 2022, Barcelona, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9191-7/22/06…$15.00
https://doi.org/10.1145/3501247.3531556

ACM Reference Format:
Myrsini Gkolemi, Panagiotis Papadopoulos, Evangelos P. Markatos, and Nicolas Kourtellis. 2022. YouTubers Not *madeForKids*: Detecting Channels Sharing Inappropriate Videos Targeting Children. In *14th ACM Web Science Conference 2022 (WebSci '22), June 26–29, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3501247.3531556

1 INTRODUCTION

In the last few years, there has been a dramatic increase in the use of social media, and especially platforms for video sharing and consumption such as TikTok and YouTube [15, 16]. In fact, this has been the case during COVID-19 lockdowns [37], with a general increase in daily and monthly active users [10, 13], and YouTube and Facebook-based content being among the most shared [6, 20].

Nonetheless, along with the generation and exposure to valuable and acceptable content, there have been frequent uploads of media which are deemed inappropriate for specific audiences. This is an important issue regarding YouTube videos, which in spite of presenting kid-related topics (e.g., cartoons, animation movies, etc.), they can often be inappropriate for children, as the videos display disturbing, violent, sexual or other improper scenes [27, 38]. This problem has been of particular importance during recent COVID-related restrictions and confinements, since many parents resort to video platforms, such as YouTube and TV programs, to keep their children occupied while schools are closed. Consequently, children end up spending many hours per day watching videos, some of which could be inappropriate [18, 25].

In order to address this ongoing problem, YouTube has proceeded to apply various methods and filtering in the last few years. Among them are: (i) a system of 3 strikes that forces the channel owner to be careful what they upload or make available on their channel, as they could be banned from the platform [8], (ii) a *Trusted Flaggers* program [49] in which individual users, government agencies and NGOs notify YouTube of content that violates the Community Guidelines, (iii) machine learning methods for detecting inappropriate content [50], (iv) a specialized YouTube platform making available content only for kids [47], and (v) a recently introduced flag, "madeForKids" [44], that allows creators to declare whether their content is kid-appropriate or not. This is not only useful for better promoting and recommending content to users searching for kid-related videos, but also accelerates auditing of such videos by YouTube algorithms and moderators [45].

Past research has examined the problem from a video content point of view, and analyzed features available on videos and channels such as comments posted, number of views, thumbnails, and even video snapshots [14, 17, 27, 33]. However, they have not addressed the problem from the perspective of accounts who post such videos, and whether their various characteristics reveal a tendency for posting *suitable* or *disturbing* videos.

In this paper, we make the following contributions:

- We are the first to study the characteristics of YouTube accounts that publish inappropriate videos targeting kids. In particular, we look into how older videos and accounts have been banned by YouTube for violating its policies on content publishing. We find that only 28.5% of channels that have uploaded disturbing content (and have been assessed as such in 2019) have been terminated by YouTube by mid 2021. In fact, almost 60% (or 546) of manually annotated disturbing videos are still accessible through the platform by mid 2021.
- We study the newly added flag from YouTube called "madeForKids" to understand its association to the inappropriate content and accounts publishing it. We discover that 25% of channels with suitable content are set to "madeForKids", while only 3% of channels with inappropriate content are set as such.
- We analyze 27 different characteristics of channels and how these features are associated with the type of channel and the content it publishes (i.e., if it was found to be disturbing or suitable for kids). Among these features are country and channel creation date, statistics like subscriptions and video views, keywords and topics, social media links, polarity and sentiment of description etc.
- Finally, we demonstrate how these features can help YouTube build a basic machine learning classifier to infer if a channel is likely to share inappropriate/disturbing videos or not, with up to $AUC = 0.873$. In fact, we show how this is possible to perform even at channel creation time, by using only features available at that moment and disregarding historical activity features, with up to $AUC = 0.869$.
- We make our data and code available for research reproducibility and extensibility.[1]

2 DATA COLLECTION

2.1 YouTube Crawling & Feature Extraction

The methodology followed in this study is outlined in Figure 1. We begin by studying the data made available from a past study [27] on the topic. The past ground truth dataset which was randomly sampled by a set of 844K videos assembled by /r/Elsagate and /r/fullcartoonsonyoutube [29] subreddits, includes details of 4797 YouTube videos and their four associated labels as provided by human annotators: *disturbing, suitable, restricted* and *irrelevant*. Each video was annotated by two of the authors of [27] and one undergraduate student with the assistance of a platform that includes a clear description of the annotation task, the labels, as well as all the video information needed for the inspection. Since our focus is videos that target children, we ignore the videos with labels *restricted* and *irrelevant*, and analyze the channels that posted 2442 videos with labels *suitable* or *disturbing*. We call this subset the *GT* dataset. Features are divided into three categories according to the crawling method or channel section they belong to. In Table 1, it is clear that most features were collected via YouTube API v3.

[1] https://github.com/Mirtia/Inappropriate-YouTube

Table 1: Data collected from YouTube channels.

Source	Features Collected
YouTube API derived	country, description, keywords, topicCategories, datePublished, madeForKids, viewCount, videoCount, subscriberCount, postCount, subscriptionCount, hiddenSubsribersCount(boolean), linksCount, descriptionCharCount, topicCount, subscriptionsList
Community Tab Post	datePublished, description, tags, hashtags, externalLinks, youtubeLinks, channelLinks, likeCount, thumbnailVideo
About Tab	email, links (text, URL)

YouTube Data API v3: First step in our data crawling process was to revisit these videos with YouTube's Data API v3, and assess their status (i.e., if they are available or not), as well as collect further public information about channels that published these videos. Each channel is distinguished by a unique 24-character identifier. To reach a channel, you "concat" the identifier with the specified (URLs): *https://www.youtube.com/channel/ID, https://www.youtube.com/c/ID*.

In particular, during this crawling, we collected the status and following attributes associated with each channel: "country", "description", "keywords", "publishedAt", "madeForKids", "topicCategories", 'viewCount", "videoCount", "subscriberCount", as well as calculated counts such as "keywordsCount", "topicCount", "subscriptionCount", "descriptionCharCount" and "postCount". For the sake of clarification, "publishedAt" states the date a YouTube channel joined the platform and "topicCategories" is a list of Wikipedia URLs that describe the channel's content. We note that since YouTube Data API v3 did not provide a method to parse the status of each video or channel, we used the *Beautiful Soup Python Library* [39] instead, to scrape the relative messages from the page source. Ethical considerations of our crawling method are addressed in Appendix A.

Community and About Tabs: Apart from these features, we also inspected other publicly available sources of account-centered information, such as the "Community Tab" and "About Tab". The Community Tab contains posts with enriched media uploaded by the account owner. As this is a newly added feature, YouTube Data API v3 does not offer a method to get its information automatically. Therefore, in order to collect these posts, we used *Puppeteer* [28] and Python's *concurrent.futures* [34] for multi-threading, along with Beautiful Soup to scrape the resulting pages at a limited request rate that may not disturb the YouTube platform. We focused on 100 posts of each channel as an indicator of what type of content the channel owner generally posts. Features extracted per post are: "datePublished", "description", "tags", "hashtags", "externalLinks", "youtubeLinks", "channelLinks", "likeCount", and "thumbnailVideo". In particular, "channelLinks" are URLs of other tagged channels or users in the description; "externalLinks" are URLs found in the description and redirect to other pages than YouTube; "thumbnailVideo" is the ID of the video embedded in a post. The About Tab of a channel consists of a description section, details (email for business inquiries, location), stats (date the user joined YouTube, number of views) and links (social media, merchandise, etc.). We used Puppeteer to collect both links and emails.

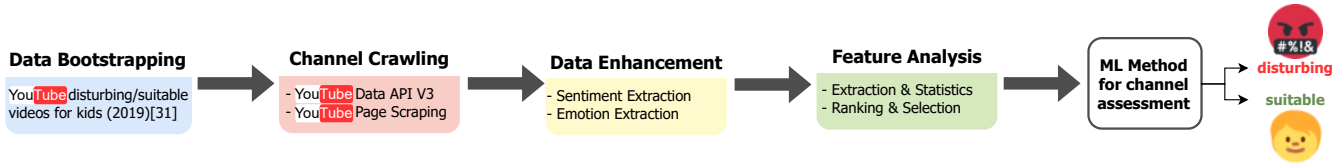

Figure 1: Overview of methodology followed in this study.

Sentiment & Emotion Extraction:: In order to extract features related to *sentiment* and *emotion*, we used the *MeaningCloud Deep Categorization API Emotion Detection* [24] to classify the text description of each channel. In addition to Emotion detection, we calculated polarity of keywords, posts and channel description using the well-known *SentiStrength* [35] library.

2.2 Channel Labeling

As mentioned earlier, the videos were split into four categories: *disturbing, suitable, restricted* and *irrelevant*. We focus on *suitable* and *disturbing*, depending on whether the content shown is appropriate or not for children.

These two labels were introduced in the past study on the subject of detecting disturbing YouTube videos for kids. Any video that is not age-restricted but targets children audience and contains sexual hints, horror scenes, inappropriate language, graphic nudity and child abuse was labeled as *disturbing*. According to YouTube Child safety policy [46], a video would be considered inappropriate(*disturbing*) if it contains misleading family content, cyber-bullying and harassment involving minors. On the other hand, a video is *suitable* when its content is appropriate for children (G-rated [42]) and it is relevant to their typical interests. We consider a channel "potentially disturbing" when they have already uploaded at least one video that was manually annotated as *disturbing* by the previous study. For sake of simplicity, we refer to these channels as *disturbing* for the rest of the study.

Then, we look into the number of *disturbing* videos that each channel posted, from *GT*. Figure 2 plots the CDF of the ratio of disturbing videos to total videos within *GT*, per channel that had at least one disturbing video in the original dataset. Through YouTube v3 API, we confirm that ~5% of accounts with reported disturbing videos have zero "videoCount" because they were probably unlisted, privatized or reported for violation of YouTube Guidelines.

Based on this preliminary result, we make the following assumptions when propagating the video labels to the channels:

- **Suitable Channel:** If it has published only "suitable" videos, based on the videos in *GT*.
- **Disturbing Channel:** If it has published at least one "disturbing" video, based on the videos in *GT*.

Table 2 summarizes the number of videos and channels from our crawls, along with their associated labels which we use in the rest of the study. All crawls on YouTube were performed in mid 2021.

2.3 Examples of Disturbing Channels

Inappropriate content comes into various forms, from a grotesque clickbait thumbnail to horror stories with cartoon characters. For the sake of example, we provide thumbnails of videos that some

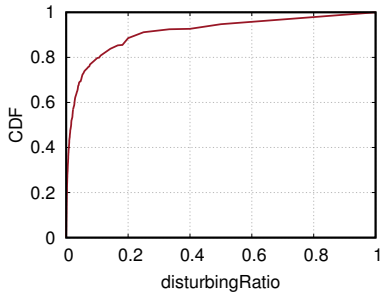

Figure 2: CDF of disturbingRatio, i.e., number of disturbing videos found in an channel over the total number of videos (suitable+disturbing) from that channel, when that channel had at least 1 disturbing video.

Table 2: Number of videos and channels per label. *Total* reflects the number of videos (and consequently channels) that were originally in the *GT* dataset. *Available* reflects the videos and channels that were successfully crawled in 2021 and are studied in this paper.

Category	Channels		Videos	
	Total	Available	Total	Available
suitable	909	779	1513	1505
disturbing	789	559	929	539

channels we labelled as "disturbing" have been hosting in their accounts. Please note that these videos were still available on May 2022, i.e., more than one year after the initial YouTube crawls of our aforementioned dataset, and two years after the initial dataset of inappropriate videos for kids was published [27].

Figure 3 shows various examples (via screenshots) of such inappropriate content targeting kids. To the left side of Figure 3, there is an example of a channel uploading gameplay videos to promote games for children. The thumbnails depict a doll getting tortured with various tools. On the right side of Figure 3, we can see another channel included in the dataset, which uploads implied sexual content of animated characters, mainly Elsa. Other examples, omitted here due to space, include horror parodies of Peppa the Pig and videos with actors role-playing as famous comic characters that engage into explicit acts.

3 CHANNEL FEATURE ANALYSIS

3.1 Why are videos and channels removed?

First, we look into the status of videos annotated by the past study, as well as the accounts that posted them. This is important in order to assess which videos from the *disturbing* set may have been removed by YouTube, and in what extent the reasoning behind the

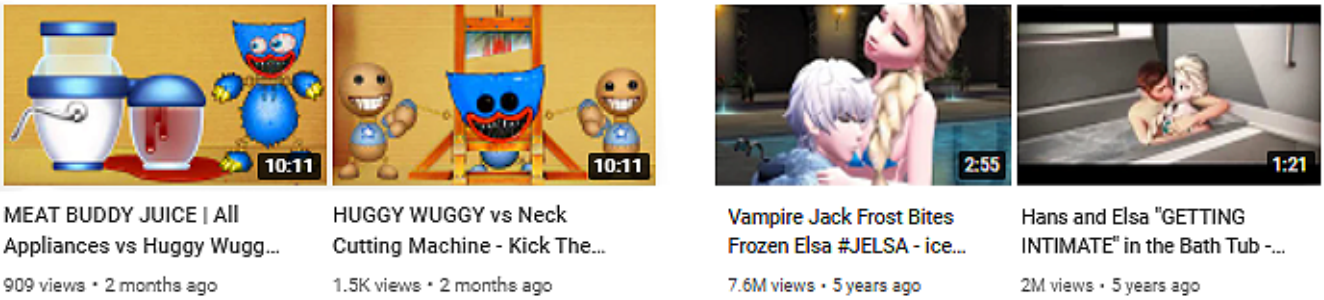

Figure 3: Overview of methodology followed in this study.

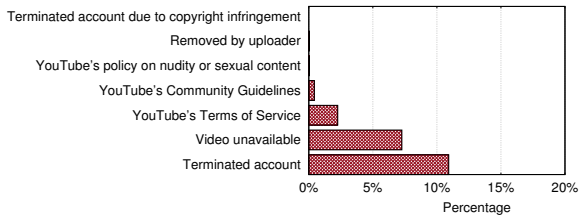

Figure 4: Reasons why YouTube videos labeled as "disturbing" are not currently available on the platform (% total videos in GT).

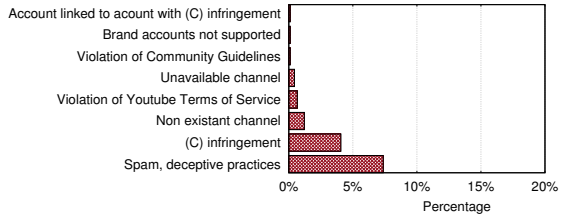

Figure 5: Reasons why YouTube channels labeled as "suitable" are not currently reachable on the platform (% total channels in GT).

removal aligns with the label provided by the past study. Whenever a video is not available in the platform, YouTube displays a characteristic message explaining the reason why the user cannot view the video. Since YouTube API v3 does not include methods to collect error messages on removed videos, we used Beautiful Soup to parse them. In general, YouTube videos may not be reachable because of different reasons: unavailability of the service or network (less likely), if the content was made private by the owner, or if the video was going against the Community guidelines and policies of YouTube and was removed.

We analyze the reasons why videos classified as "disturbing" or "suitable" were removed by YouTube. As shown in Table 2, only 0.1% of *suitable* videos were removed, while more than 40% of *disturbing* videos were taken down, with the dominant reason being account termination. More specifically, and as shown in Figure 4, 10.9% (203) of removed disturbing videos are linked with terminated accounts and 2.2% of such videos are linked with accounts banned because of not respecting YouTube Terms of Service.

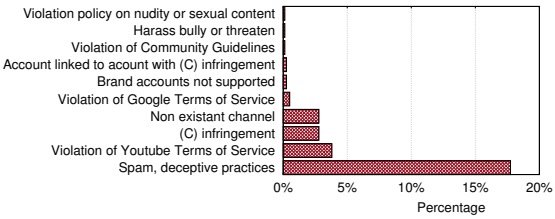

Figure 6: Reasons why YouTube channels labeled as "disturbing" are not currently reachable on the platform (% total channels in GT).

After studying the possible causes of why videos were taken down, we move to examine the status of channels that uploaded these videos. This data collection consists of each channel and their respective videos included in GT. YouTube actions on violating Community Guidelines consist of four levels [8]. In the beginning, the user who owns the account receives a warning, apart from severe abuse cases when the channel is terminated immediately. The second time a user's content is considered improper, they receive a strike. Their actions, such as uploading videos, creating or editing playlists, etc., are restricted for a week. However, the strike remains on the channel for 90 days. In case the user receives a second strike during this period, they become incapable of posting content for two weeks. A third strike during this time interval results in permanent removal of the channel.

As we see in Figure 5, suitable channels were less likely to have been removed during the elapsed time between the past study in our crawls. In fact, 7.37% of suitable channels were terminated due to multiple small or severe violations of YouTube's policy against spam, deceptive practices, and misleading content, or other Terms of Service violations, and 4.07% in consequence of copyright infringement. Instead, in Figure 6, we observe that more than double (17.74%) of disturbing channels were banned from YouTube platform because of spam and deceptive practice policies, as well as for violating YouTube Terms of Service (3.8%), copyright infringement (2.78%) channel absence (2.78%).

Overall, and after our crawls and analysis, while 929 videos were classified in the past study as "disturbing", 58.8% are still reachable in mid 2021. In fact, only 28.5% of the users/channels that have uploaded such disturbing content have been terminated by YouTube, demonstrating a lack of action by the platform.

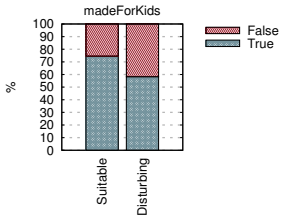

Figure 7: The use of *madeForKids* label by videos on YouTube labeled as suitable or disturbing.

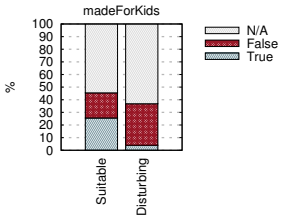

Figure 8: The use of the *madeForKids* label by YouTube channels labeled as suitable or disturbing.

3.2 Are videos and channels *MadeForKids*?

YouTube Creators published a video on the updates of "Complying with COPPA" on 12th of November, 2019 [44] where they introduced the "madeForKids" label for both channels and videos. This feature denotes whether the content of a video or channel is directed at children. More specifically, the content is "madeForKids" if it is child-friendly, and most likely includes child actors, animated characters or cartoon figures, or serves educational purposes.

To comply with the Children's Online Privacy Protection Act (COPPA) [40] and other related laws, YouTube makes certain features of its regular channels unavailable on "made for Kids" content and channels. Regarding videos, these switched-off features include: auto-play on home, cards or end screens, channel branding watermark, comments, donate button, likes/dislikes on YouTube Music, live chat or live chat donations, merchandise and ticketing, notifications, personalized advertising, playback in the Mini-player, Super Chat or Super Stickers, save to playlist and save to watch later. At channel level, the restricted features include Channel Memberships, Notifications, Posts, and Stories. Regarding the aforementioned "madeForKids" flag, a channel can be:

(1) "madeForKids": allowed to only post videos "madeForKids";
(2) not "madeForKids": allowed to only post videos that are not "madeForKids";
(3) not defined: each video is defined if it is "madeForKids" or not on upload time;

However, YouTube is also supported by a machine learning algorithm to detect incorrectly labeled videos and set them according to their content [44].

Figures 7 and 8 summarize the results of the analysis of the "madeForKids" flag, as set by the channel owners. Given that the videos in *GT* are targeting kids audience, it comes as no surprise that, as shown in Figure 7, the majority of videos analyzed are "madeForKids", regardless of category, i.e., if they are disturbing or not. This may be because the creators were aiming to convince the YouTube algorithm that these videos should be recommended to children. It is encouraging that more suitable videos were marked as "madeForKids" than disturbing videos. Also, out of 390 disturbing videos that were removed, only 1.5% were set to "madeForKids". Perhaps surprisingly, and according to Figure 8, most of the channels are not set to "madeForKids", even though they hosted such content, possibly because they did not share only such content. Overall, we find 199 (~25%) suitable channels that are exclusively declared as "madeForKids", while 3% of disturbing channels were so. This may indicate that either the channels posting disturbing videos do not want to draw attention and fast auditing of their videos by YouTube, or their target audience is not kids, and any viewing of their content by kids is accidental. In either case, we believe there is a significant problem at hand, since kids can reach these videos and channels quite easily, with a few clicks, as shown by past research [26, 27].

Table 3: Statistics for YouTube channels annotated as suitable or disturbing.

Features	Suitable		Disturbing	
	Records	Median	Records	Median
videoCount	779	202	559	61
viewCount	779	60M	559	2488k
subscriptionCount	779	0	559	0
subscriberCount	700	348k	524	9.7k
descriptionCharCount	623	287	419	187
keywordsCount	547	12	312	9
topicCount	756	3	524	3.0
postCount	468	2	357	4

3.3 Characteristics of YouTube Channels Hosting Videos For Kids

Next, we analyze the data collected on attributes of each channel, to understand the differences between channels that post only *suitable* videos and those that upload *disturbing* videos.

Channel Date Creation, Country and Email: First, we examine the date (year) channels joined YouTube. As seen in Figure 17, the peak of channel creations for both *disturbing* and *suitable* channels in our dataset is observed in 2016. After that point, there is a steep decrease in count. This is due to several measures taken since 2017. As the term "Elsagate" grew popular, Twitter users drew attention on the topic, and in June 2017, a subreddit r/Elsagate [30] was created to discuss and report the phenomenon. In addition, during the same year, several articles were published about channels featuring inappropriate content and how harmful videos manage to get through the countermeasures of YouTube. To resolve the controversy, YouTube began to take action by deleting accounts and videos and tightening up its Community policies and guidelines [38].

Next, we look into the country of origin which is displayed in the "Details", along with "Email for Business inquires", in case it exists. In Figure 18, we plot the top countries that channel owners featured, as well as "N/A" for channels that did not display this information. As perhaps expected, most of the channels originate from United States, with the top 3 popular channels (ranked based on subscribers) being "Cocomelon" (>100M), "Kids Diana Show" and "Like Nastya", ranging between 70 and 90M, which are classified as "suitable" channels. It should be noted that an important quantity of *suitable* channels have set their location to India, which is not as frequent in the opposing category (*disturbing*). Most popular suitable accounts from India include "ChuChu TV Nursery Rhymes & Kids Songs" (46.2M), "Wow Kidz" (21.9M), and "Green Gold TV - Official Channel" (15.4M).

Channel Statistics and Subscriptions: Next, we perform non-parametric, Kolmogorov-Smirnov (KS) testing to find out whether or not the distributions of the two types of channels are statistically different. To begin with, we study the channel statistics, i.e., viewCount, videoCount, subscriberCount and subscriptionCount. From

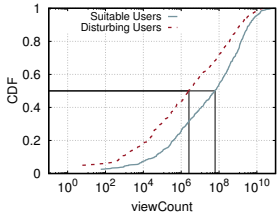

Figure 9: CDF for viewCount (number of total views) per channel for disturbing or suitable users.

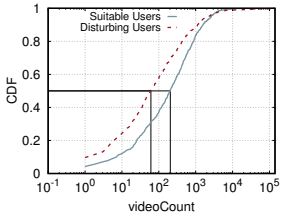

Figure 10: CDF for videoCount (number of current publicly visible videos) per channel for disturbing or suitable users.

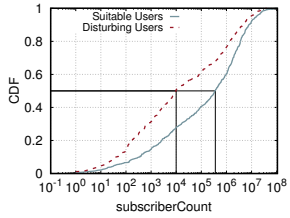

Figure 11: CDF for subscriberCount (can be hidden) per channel for disturbing or suitable users.

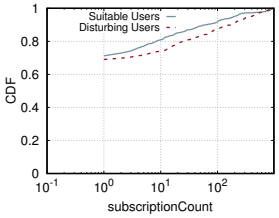

Figure 12: CDF for subscriptionCount (visible subscriptions) per channel for disturbing or suitable users.

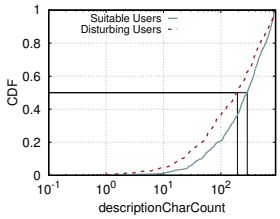

Figure 13: CDF of descriptionCharCount (number of characters in chanel description (no spaces)) per channel for disturbing or suitable users.

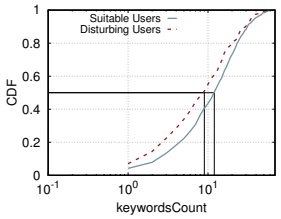

Figure 14: CDF of keywordsCount (number of keywords) per channel for disturbing or suitable users.

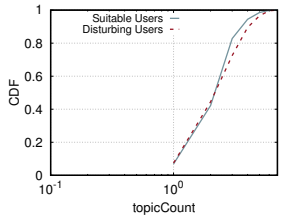

Figure 15: CDF of topicCount (number of topics per channel – can be hidden) per channel for disturbing or suitable users.

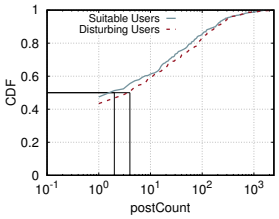

Figure 16: CDF of postCount (number of posts) per channel for disturbing or suitable users.

Figure 9, it is evident that suitable channels have more views, on average, than disturbing channels (~1.7B vs. ~663M). This is also true for number of videos publicly available on each channel (Figure 10), number of subscribers per channel (Figure 11) and number of subscriptions (Figure 12). It should also be pointed out that the average ratio of views per video is three times higher in channels of suitable than disturbing content (4.2M vs. 1.4M) Then, as summarized in Table 3 for the two type of channels, we look closer into the subscriberCount, which indicates how many people have subscribed to a channel to keep up with its newly updated content and support the creator. The public subscriberCount is rounded depending on the number of subscribers. Thus, it is different from the actual subscriber count which is private and exclusively available to the owner of the YouTube channel [7]. We collected public subscribersCount for each channel via YouTube Data v3 API. However, each creator has the option to hide the subscriberCount of their channel. We observe that ~10% of *suitable*, but only ~6% of *disturbing* channels choose to conceal the number of their subscribers. KS test performed on each of these four features allows us to reject the null hypothesis that the two types of channels originate from the same distribution at statistical level $p - value < 0.0328$ or lower (all statistics are summarized in Table 4).

Branding settings, Topic Details and Posts: Next, we examine the attributes that are related to the content description, i.e., descriptionCharCount, keywordsCount, topicCount, and postCount. Again, channels with only *suitable* videos seem to have longer descriptions (Figure 13) and more keywords (Figure 14) used in their configurations. Interestingly, the distribution of number of topics (Figure 15) and number of posts per channel (Figure 16) seem to be similar for the two types of channels. As earlier, we performed KS tests and found that we cannot reject the null hypothesis for the

Table 4: Kolmogorov-Smirnov for count-based channel characteristics.

Feature	p-value	D-statistic
videoCount	2.636e-06	0.21333
viewCount	1.211e-03	0.20359
subscriptionCount	3.288e-02	0.07944
subscriberCount	8.882e-15	0.23482
descriptionCharCount	3.835e-12	0.16439
keywordsCount	2.729e-13	0.13655
topicCount	2.867e-03	0.10285
postCount	6.802e-01	0.05049

postCount feature, and the two types of channels come from the same distribution ($p - value = 0.6802$).

Topic Categories and Keywords: Topic categories and keywords are used to describe and associate a creator's content with specific search results and recommendations. It is of high importance to set up these features properly in order to reach the desired audience and achieve channel growth. Both of these features can be collected via YouTube API v3. In Table 5 we show the top 10 keywords and top 10 topics used, respectively, for the two types of channels. It is evident that, apart from the usual children-associated tags which appear to be prevalent on both types of channels, *disturbing* channels use gaming-related keywords and topics more often than *suitable* channels. This is a result of channels uploading MLG [2] content and heavily moded ROBLOX [43] and Minecraft [41] videos.

3.4 Viewers Interaction & Social Media Presence

Apart from the general features that compose a channel, there are additional capabilities that focus on bridging the connection between a channel and its subscribers. Community Tab, which

Table 5: Ten most used keywords and topicCategories per channel type.

Category	Keywords (frequency)	topicCategories (frequency)
suitable	kids(70), fun(30), toys(47), animation(44), children(41), cartoon(34), funny(30), cartoons(30), for kids(30) nursery rhymes(35)	Entertainment(470), Film(338) Lifestyle_(sociology)(327), Hobby(221) Music(185), Television_program(110) Video_game_culture(87) Action-adventure_game(51) Action_game(50) Role-playing_video_game(44)
disturbing	funny(47), animation(34), comedy(26), gaming(18), cartoon(15), kids(15), cartoons(14), fun(16), minecraft(12), Gaming(11)	Entertainment(343), Film(229) Video_game_culture(135), Music(120) Action-adventure_game(51) Action_game(91) Role-playing_video_game(61) Hobby(61), Pop_music(37)

Table 6: Top social media & websites used or linked in YouTube channels.

Platform	Suitable	Disturbing
facebook	282	129
instagram	217	147
merchandise	16	25
twitch	10	35
twitter	190	160

is one of the latest features offered by YouTube, released its beta version in 2016 [21]. A creator unlocks this feature upon reaching 1000 subscribers, and they can make use of it only if their channel is not set to "madeForKids" [44]. From that point on, they are able to create posts and embed playlists, GIFs, images, videos, polls, etc [9]. Also, viewers get Community post notifications as they get from video uploads, but only in case their notifications are enabled.

Indeed, a large number of suitable channels do not have the Community Tab feature enabled, as, also pointed out in Section 3.2, more than 25% suitable channels are "madeForKids". Thus, even though they have a higher average number of subscribers than disturbing channels (as was shown in Figure 11), a significant portion of these channels cannot use the Community Tab feature. Interestingly, in Figure 16, disturbing channels exhibit more posts per channel on average than suitable channels.

Channel owners can also display their social media and link their channels to other platforms and websites. This is shown in the About Tab, which contains general details about a channel. More specifically, it includes the channel description, statistics such as date of creation and total views, links and e-mail information. For each channel, we collected the social media, external URLs and e-mail associated with the account.

The 10 most frequent social media referenced in the About section are shown in Table 6. As expected, popular networks such as Instagram, Twitter and Facebook are prevalent. The majority of suitable channels display Facebook in their links, while disturbing channels show a preference for Twitter. Moreover, by including their contact info, channel owners encourage communication with their audience and are easily accessible for possible collaborations [48]. However, in Figure 19, we see that less that a half of channels for both types provide their email addresses. Even so, disturbing channels are slightly less likely to add their contact information than suitable channels.

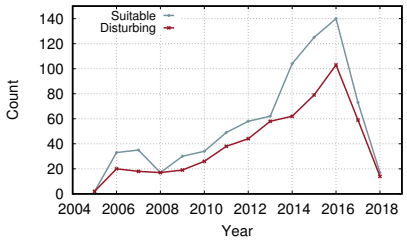

Figure 17: A frequency distribution of the year that YouTube channels were created (channel feature "publishedAt"), and are labeled as "suitable" or "disturbing".

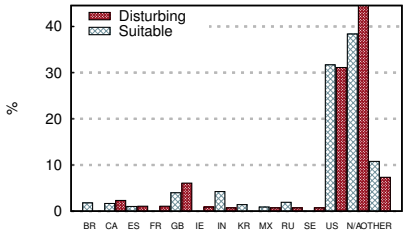

Figure 18: Top 13 countries of channel creation. "N/A" refers to channels that specifically did not define country. "Other" refers to channels in countries beyond the top 13 shown here.

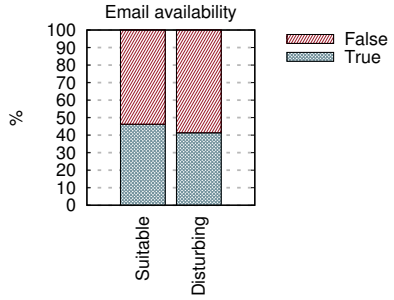

Figure 19: Use of email for business inquiries in YouTube channels labeled as disturbing or suitable.

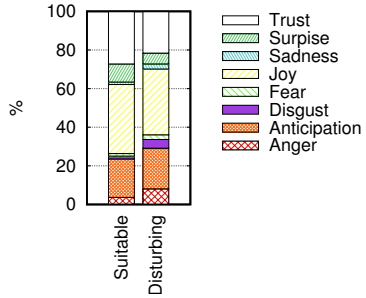

Figure 20: Emotion analysis on YouTube channel description, for channels labeled as suitable or disturbing.

3.5 Sentiment Analysis

Basic Emotions: We present the analysis of sentiment performed on the various data collected per channel that include text, such as the channel keywords and the About and post description. Beginning with channel description, we conducted analysis on the eight

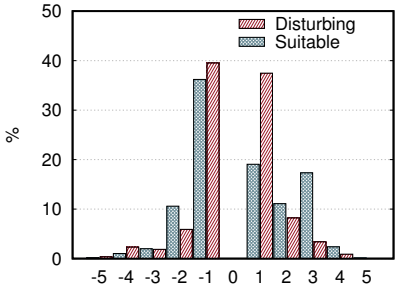

Figure 21: Polarity of description of YouTube channels labeled as suitable or disturbing.

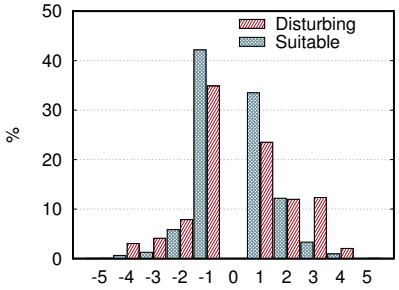

Figure 22: Polarity of keywords of YouTube channels labeled as suitable or disturbing.

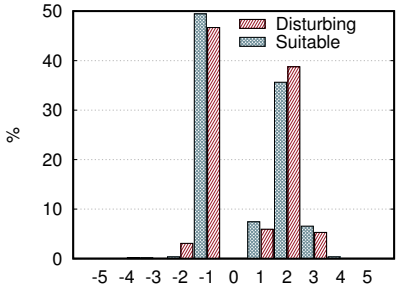

Figure 23: Polarity of posts (mean) of YouTube channels labeled as suitable or disturbing.

Table 7: Analysis of top emoticons used in posts and Community Tab for YouTube channels labeled as suitable or disturbing.

Posts Text						Channel description					
Suitable			Disturbing			Suitable			Disturbing		
Emoji	Count	Score	Emoji	Count	Score	Emoji	Count	Score	Emoji	Count	Score
❤️	19	0.747	😂	7	0.805	❤️	16	0.754	☣️	9	-
👇	10	0.557	❤️	5	0.747	®	15	0.353	❤️	8	0.754
❤️	10	0.754	😍	5	0.765	©	15	0.740	🇬🇧	4	-
😍	10	0.765	❗	4	0.620	❤️	13	0.747	❤️	3	0.747
🐥	10	0.780	♂️	3	0.580	™	11	-	®	3	0.353

basic emotions as in Robert Plutchik's Wheel of Emotions [36]. It covers eight prime emotions: Trust, Surprise, Sadness, Joy, Fear, Disgust, Anticipation and Anger. We use MeaningCloud Emotion Detection Deep Categorization API to extract them. The results presented in Figure 20 show the percentage of emotion detected in text description. Negative emotions such as Anger, Disgust, Sadness and Fear are more prevalent in disturbing channels, while positive emotions such as Trust and Surprise are expressed less frequently. This finding correlates with the nature of disturbing content which is characterized by extreme media content and language. It is observed that suitable channels' descriptions express more Surprise. Also, suitable and disturbing channels show similar percentages of Anticipation and Joy.

Polarity: Then, we look into the positive and negative polarity of the description text, keywords and posts of each channel. In Figures 21, 22 and 23, we show a breakdown of polarity for each of the previous features. Regarding the text in their description, both types of channels are using words that convey slightly negative sentiment (-1). However, disturbing channels' values are higher than suitable channels, in both negative (-1) and positive (+1) sentiment. In fact, for the positive side, the disturbing channels use lightly positive sentiment words (+1) almost twice as much as suitable channels. Overall, disturbing channels use keywords with higher sentiment than suitable channels, both positive and negative. This is probably an attempt to evoke attention, emotion and increase possible engagement with the audience. Similarly, disturbing and suitable channels exhibit a high frequency of lightly negative words (-1) as well as positive words (+2) in their posts.

Emojis: We performed emoji detection in the text of channel description and posts, with the assistance of Python library *emoji* [22]. Table 7 shows the frequency of emojis and their sentiment score according to Emoji Sentiment Ranking v1.0 [23]. Heart emojis such as ❤️ and ❤️ prevail. Suitable channels express ownership in their description by using frequently ©, ® and ™ emojis. The most frequent emoji in disturbing channels' description is ☣️ (bio-hazard emoji), which even if it does not reflect a specific sentiment score, is associated with negative emotion [12].

4 DISTURBING CHANNEL DETECTION WITH MACHINE LEARNING

4.1 Data Preparation & Performance Metrics

We use the aforementioned features (also summarized in Table 8) to train different classifiers for automatic classification of channels into two classes: 1) likely to post only suitable videos (suitable), 2) likely to post at least one disturbing video (disturbing). In order to compute the classification task, we performed basic preprocessing of the features available, such as removing features with very little to zero variability, and applying logarithmic transformation on several numeric features for normality purposes. Table 8 lists the groups of features used in our classification analysis. As mentioned earlier, the "suitable" channels are 779 and "disturbing" channels

Table 8: Groups of features used for machine learning classification of channels as suitable or disturbing.

Group of Attributes	# of features
Channel Details & Activity (count*)	6
Graph-related metrics (subscriptions, etc.)	3
madeforKids Status (ratios, etc)	4
Top media linked	11
Top keywords per channel	10
Emotions in Description	8
Top topics on Description	11
Emoji score Posts/Description	2
Top emojis in Description	10
Top emojis in Posts	10
Polarity Posts/Description/Keywords	6

are 559. We applied 10-fold cross-validation on the available data, and trained and tested various techniques. We measured standard ML performance metrics such as True Positive and False Positive Rates, Precision and Recall, F1 score and Area Under the Receiver Operating Curve (AUC). Where applicable, the scores for these metrics were weighted to take into account individual performance metrics per class.

4.2 Feature Ranking

We also performed an analysis of the available attributes, and ranked them based on contribution to the classification task. In particular, we evaluate the worth of an attribute by measuring the information gain with respect to the class, when each attribute was included or not in the classification task. This effort was performed with a 10-fold cross validation method, and average scores were computed. Our analysis shows that the top feature groups are:

(1) Polarity (keywords or description)
(2) Channel-statistics metrics such as views, subscriber and video counts, country
(3) Top keywords such as nursery rhymes, children, kids, toys
(4) Top topics such as hobby, game-related, lifestyle
(5) Top emotions on description such as trust, surprise, and anger
(6) Emojis and emoji score in text (description, post text, keywords)

This ranking is in line with the results from the previous section, which highlighted that emotions and polarity of channel description have a different profile in disturbing channels than suitable. Also, characteristics of the channels such as activity statistics and keywords or topic categories used are significantly different in disturbing than suitable channels.

4.3 Classifiers Performance

Table 9 presents the results achieved with several different classifiers and meta-classifiers. We find that the typical Random Forest (RF) classifier performs very well across the board, with high True positive and low False positive rates, and higher Precision and Recall than the other classic ML methods. Only another classifier, meta-classifier (Meta:LogitBoost with RF) which uses a regression scheme as the base learner and can handle multi-class problems, performs somewhat better than simple Random Forest, at the expense of higher computation and memory cost. Another meta-classifier consisting of 4 others (Random Forest, Logistic Regression, Naive Bayes and Bagging RF) and applying average probabilistic voting among them performs similarly.

Regarding the neural network classifier, we tried different architectures, including dense layers for normalization, dropout, etc. However, due to the small number of examples available in our dataset (1338 samples), these more complex classifiers did not fare better than the simple architecture reported in the results.

We also attempted to build a RF classifier that uses only the features available at the moment the channel is generated. That is, we dropped features that stem from user and channel activity such as counts (view, video, subscriptions, etc.), posts and their emotion analysis, etc. The results shown in the last row of Table 9 demonstrate that it is in fact possible to predict with good performance which channels are likely to post some disturbing content targeting kids, before they have posted anything in their channel, or had any views or subscribers, etc.

5 RELATED WORK

Previous studies have been conducted regarding disturbing content targeting children in video and streaming platforms. Ishikawa et al. [17] combined raw frames and MPEG motion vectors as a ground dataset to build a classifier detecting Elsagate-related videos. They propose various machine learning models and compare their performances, as well as ways to reach into a mobile compatible solution with 92.6% accuracy. They also mention the ambiguity of "Elsagate" definition, and the danger of false positives of this type of content. Alghowinem [1] used slices of videos accompanied with audio analysis and speech recognition to provide a real-time mechanism for monitoring content on YouTube and detect inappropriate content for kids.

Next study of relevance is KidsTube by Kaushal et al. [19]. Initially, the authors studied three major feature layers: video, user and comment data. Then, they built a classifier on these data, as well as a version that is based on a Convolutional Neural Network that uses video frames. The popularity and network of content uploaders was examined through user statistics such as subscriptions, views, etc. In fact, they found that likes, subscriptions and playlists can form a network of unsafe promoters and video uploaders.

Another user-centered study is by Benevenuto et al. [5] which comments on content pollution in video sharing platforms and provides a classification approach at separating spammers and promoters from appropriate users. Furthermore, Reddy et al. [31] suggested an age detection process for underage YouTube users, supported by performing sentiment analysis on comments. In this way, the authors offer a real time protection mechanism that can be embedded in the current YouTube platform. Continuing with Alshamrani et al. [3] [4], they perform analysis of exposure of YouTube users to comments, and construct a classifier to detect inappropriate comments in children-oriented videos. They find that 11% of comments posted in such videos are toxic.

Lastly, Papadamou et al. [27], collect videos targeting children using various seed keywords from animation movies and popular cartoons. They analyze various types of features available or constructed per YouTube video, and based on these features, the authors build a classifier with 84.3% accuracy which detects inappropriate videos that target children. They also underline the dangers of leaving toddlers to watch YouTube videos unattended, and examine the likelihood of a child browsing the platform and coming across a

Table 9: Performance metrics from ML binary classification of channels. 0: likely to post suitable only content; 1: likely to post at least one disturbing video.

Method	TPRate	FPRate	Precision	Recall	F1	AUC
Random Forest (RF)	0.791	0.225	0.790	0.791	0.790	0.873
Logistic Regression (LR)	0.753	0.256	0.755	0.753	0.754	0.820
Naive Bayes (NB)	0.716	0.321	0.713	0.716	0.712	0.786
Neural Net (38x128x2)	0.761	0.246	0.763	0.761	0.762	0.818
Meta: LogitBoost(RF)	0.796	0.218	0.796	0.796	0.796	0.873
Meta: AvgProb(RF,LR,NB,BRF)	0.782	0.237	0.781	0.782	0.781	0.864
RF with only channel gen. features	0.781	0.222	0.784	0.781	0.782	0.869

disturbing video by chance. Our ground truth dataset originates from this study, from which we use the labels provided per suitable or disturbing video.

Comparison: Our present study goes beyond the aforementioned past works in the following ways:

- We shift the problem of *disturbing* videos into the topic of potentially disturbing users creating this type of content. In fact, we are the first to check the status (i.e., if they are available or not) of the videos and channels after an interval of two years, and examine the reasons why they may have been removed by YouTube and in what extent.
- We are the first to examine the newly introduced "madeForKids" flag for both videos and channels, and how its value associates with the type of channel (suitable or disturbing).
- We extract and analyze Community Tab posts and perform sentiment and polarity analysis on channel description and post texts.
- Furthermore, we use channel public features (e.g., activity and channel related details, posts, keywords, etc.), as well as features not available from the API (e.g., linked media, top emojis topics, polarity, emotions, etc.), to construct a machine learning classifier which detects with good performance channels likely to share disturbing content.

6 DISCUSSION & CONCLUSION

The present study focused on an investigation of YouTube channels with respect to the type of videos they share and if these are classified as *disturbing* or *suitable* for kids.

Findings:

- We looked into whether older videos and accounts have been banned by YouTube for violating its policies on content publishing, and examine the reasons why the channels were removed. Alarmingly, we find that the majority of *disturbing* videos (60%) from a past study (2019), along with their uploaders (channels) (71%) are still available in mid 2021, during the time interval that our data collection was performed.
- We studied the newly added flag from YouTube called "madeForKids" to understand how channels and videos marked as disturbing may be correlated to it. We discovered that 1/4th of channels with suitable content are set to "madeForKids", but only 3% of disturbing channels are set as such, which may stem from efforts to avoid attention from YouTube.

Furthermore, we studied 27 publicly available features and examined how they are linked to the type of YouTube channel (i.e., if it was found to solely share suitable videos for kids, or disturbing as well) and made several observations that differentiate channels hosting disturbing from suitable videos for kids. A list of the most important findings on these features are presented below:

- A large number of channels were created in 2016. After that point, less disturbing channels were created, as "Elsagate" started to gain attention in 2017 leading to shutdown of disturbing channels from YouTube.
- Suitable channels have higher number of views and subscribers than channels with disturbing videos.
- Suitable channels tend to use more keywords and have longer descriptions than disturbing channels.
- Disturbing channels use gaming-related keywords and topics more often than the suitable channels.
- The majority of suitable channels add Facebook in their links; disturbing channels prefer Twitter.
- The majority of channels do not provide their email address. However, disturbing channels are slightly less likely to add their contact information.
- Negative emotions such as Anger, Disgust and Sadness are more prevalent in disturbing channels than suitable channels.
- Disturbing channels use keywords with higher sentiment, negative or positive, in comparison to suitable channels.

Automatic ML Classifier: Finally, based on these studied features, we constructed machine learning (ML) classifiers which detect with adequate performance (up to AUC=0.873) channels likely to share disturbing content. In fact, we show how this classification is possible to be performed even at the time a channel is created, by using only features available at that moment and disregarding their activity history or posting features, with up to $AUC = 0.869$. For reproducibility purposes, we make all our data and code available.

Impact: We believe our analysis of the "madeForkids" flag, the characteristics of the disturbing accounts and the ML-based classifier can be combined with other automated tools readily available by academia and YouTube, to fight against inappropriate content exposure and especially when it is targeting kids. In particular, YouTube could use the results of this study with respect to features differentiating disturbing and suitable accounts, and our suggestion of an ML-based classifier, to create a multi-step process for flagging channels sharing inappropriate content. This process can follow these steps:

Step 1: Extract detailed features per channel, as explained here.

Step 2: Train ML method based on these features to detect accounts posting potentially disturbing videos for kids.
Step 3: Extract detailed features per video posted in such accounts, following methodology of [27].
Step 4: Train ML method based on these features, and use it to detect potentially disturbing videos.
Step 5: Rank said accounts from Step 2 based on appropriate metric of disturbing content severity such as: the probability of said accounts being disturbing (based on the ML classifier of Step 2), the probability of said videos being disturbing (based on the ML classifier of Step 4), the number of disturbing accounts posted by said account, etc.
Step 6: Human moderators can then look into the top ranked disturbing accounts for potential violation of Terms and Conditions and Community Guidelines of YouTube, and consider applying the 3-strike policy.

This process could be used as a safety net when the YouTube for Kids application is not available in the country of residence of the children using YouTube.

Limitations: Last but not least, we shall not forget to mention the limitations of this research. The dataset size is limited as it strictly consists of channels that have uploaded videos from the previous study. There is a selection bias in the sense that the dataset does not cover the whole YouTube platform, but it emerges from child-related content. In addition, from our findings, it is apparent that there is a discrepancy between what YouTube considers inappropriate and worth striking and what humans think of as disturbing. For example, many "disturbing" annotated videos may fall into the category of dark or adult humour which does not necessarily mean that they should be punished by the platform moderators. Consequently, it is difficult to decide whether "disturbing" videos should be removed or there should be better monitoring or categorization of videos to multiple age levels.

Overall, with our present study, we hope to raise awareness about this problem, and encourage YouTube and other similar video sharing platforms to take appropriate measures for protecting children from abusive, disturbing, and generally inappropriate content.

ACKNOWLEDGMENTS

This project received funding from the EU H2020 Research and Innovation programme under grant agreements No 830927 (Concordia), No 830929 (CyberSec4Europe), No 871370 (Pimcity) and No 871793 (Accordion). These results reflect only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

REFERENCES

[1] Sharifa Alghowinem. 2019. A Safer YouTube Kids: An Extra Layer of Content Filtering Using Automated Multimodal Analysis. In *Intelligent Systems and Applications*, Kohei Arai, Supriya Kapoor, and Rahul Bhatia (Eds.). Springer International Publishing, Cham, 294–308.
[2] All Stars Wiki. 2022. MLG Memes. https://allstarsforeveryone.fandom.com/wiki/List$_o$f$_M$LG$_m$emes.
[3] Sultan Alshamrani. 2020. Detecting and Measuring the Exposure of Children and Adolescents to Inappropriate Comments in YouTube. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3213–3216. https://doi.org/10.1145/3340531.3418511
[4] Sultan Alshamrani, Ahmed Abusnaina, Mohammed Abuhamad, Daehun Nyang, and David Mohaisen. 2021. Hate, Obscenity, and Insults: Measuring the Exposure of Children to Inappropriate Comments in YouTube. In *Companion Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 508–515. https://doi.org/10.1145/3442442.3452314
[5] F. Benevenuto, T. Rodrigues, A. Veloso, J. Almeida, M. Goncalves, and V. Almeida. 2012. Practical Detection of Spammers and Content Promoters in Online Video Sharing Systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42, 3 (2012), 688–701. https://doi.org/10.1109/TSMCB.2011.2173799
[6] Timm Böttger, Ghida Ibrahim, and Ben Vallis. 2020. How the Internet Reacted to Covid-19: A Perspective from Facebook's Edge Network. In *Proceedings of the ACM Internet Measurement Conference* (Virtual Event, USA) *(IMC '20)*. 34–41.
[7] YouTube Help Center. 2021. Check your subscriber count. https://support.google.com/youtube/answer/6051134.
[8] YouTube Help Center. 2021. Community Guidelines strike basics. https://support.google.com/youtube/answer/2802032.
[9] YouTube Help Center. 2021. Learn about Community posts. https://support.google.com/youtube/answer/9409631.
[10] Dave Chaffey. 2020. Global social media research summary August 2020. https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/.
[11] David Dittrich and Erin Kenneally. 2012. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. *SSRN Electronic Journal* (08 2012). https://doi.org/10.2139/ssrn.2445102
[12] EmojiGuide.org. 2018-2020. Biohazard Emoji. https://emojiguide.org/biohazard.
[13] Sara Fischer. 2020. Social media use spikes during pandemic. https://www.axios.com/social-media-overuse-spikes-in-coronavirus-pandemic-764b384d-a0ee-4787-bd19-7e7297f6d6ec.html.
[14] Wenlin Han and Madhura Ansingkar. 2020. Discovery of Elsagate: Detection of Sparse Inappropriate Content from Kids Videos. In *2020 Zooming Innovation in Consumer Technologies Conference (ZINC)*. 46–47. https://doi.org/10.1109/ZINC50678.2020.9161808
[15] H.Tankovska. 2021. Social media platforms growth of MAU worldwide 2019-2021. https://www.statista.com/statistics/1219318/social-media-platforms-growth-of-mau-worldwide/ Mar 8, 2021.
[16] Mansoor Iqbal. 2020. YouTube Revenue and Usage Statistics (2020). https://www.businessofapps.com/data/youtube-statistics/.
[17] A. Ishikawa, E. Bollis, and S. Avila. 2019. Combating the Elsagate Phenomenon: Deep Learning Architectures for Disturbing Cartoons. In *2019 7th International Workshop on Biometrics and Forensics (IWBF)*. 1–6. https://doi.org/10.1109/IWBF.2019.8739202
[18] Joseph Johnson. 2021. U.S. kids & teens with 4hrs+ screen time before and during COVID-19 pandemic 2020. https://www.statista.com/statistics/1189204/us-teens-children-screen-time-daily-coronavirus-before-during/ May 6, 2021.
[19] Rishabh Kaushal, Srishty Saha, Payal Bajaj, and Ponnurangam Kumaraguru. 2016. KidsTube: Detection, Characterization and Analysis of Child Unsafe Content & Promoters on YouTube. (08 2016).
[20] Simon Kemp. 2020. Digital 2020: July Global Statshot. https://datareportal.com/reports/digital-2020-july-global-statshot.
[21] Kiley McEvoy. 2016. YouTube Community goes beyond video. https://blog.youtube/news-and-events/youtube-community-goes-beyond-video Sep 13, 2016.
[22] Taehoon Kim and Kevin Wurster. 2022. emoji 1.2.0. https://pypi.org/project/emoji/.
[23] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLoS ONE* 10, 12 (2015), e0144296. http://dx.doi.org/10.1371/journal.pone.0144296
[24] MeaningCloud. 2022. Meaning Cloud, Deep Categorization Models - Emotion. https://www.meaningcloud.com/developer/documentation/predefined/deep-categorization-models/emotion.
[25] Ofcom. 2021. Children and parents: media use and attitudes report 2020/21. https://www.ofcom.org.uk/research-and-data/media-literacy-research/childrens/children-and-parents-media-use-and-attitudes-report-2021 28 April, 2021.
[26] Kostantinos Papadamou. 2021. Characterizing Abhorrent, Misinformative, and Mistargeted Content on YouTube. arXiv:2105.09819 [cs.CY]
[27] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 522–533.
[28] Puppeteer. 2022. Puppeteer. https://pptr.dev/.
[29] Reddit. 2013. rfullcartoonsonyoutube. https://reddit.com/r/fullcartoonsonyoutube/.
[30] Reddit. 2017. rElsaGate. https://www.reddit.com/r/ElsaGate/.
[31] Sanjana Reddy, Nikitha Srikanth, and G. S. Sharvani. 2021. Development of Kid-Friendly YouTube Access Model Using Deep Learning. In *Data Science and Security*, Dharm Singh Jat, Samiksha Shukla, Aynur Unal, and Durgesh Kumar Mishra (Eds.). Springer Singapore, Singapore, 243–250.

[32] Caitlin M. Rivers and Bryan L. Lewis. 2014. Ethical research standards in a world of big data. *F1000Research* 3 (2014), 38. https://doi.org/10.12688/f1000research.3-38.v2
[33] Rashid Tahir, Faizan Ahmed, Hammas Saeed, Shiza Ali, Fareed Zaffar, and Christo Wilson. 2019. Bringing the Kid Back into YouTube Kids: Detecting Inappropriate Content on Video Streaming Platforms. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Vancouver, British Columbia, Canada) *(ASONAM '19)*. Association for Computing Machinery, New York, NY, USA, 464–469. https://doi.org/10.1145/3341161.3342913
[34] The Python Software Foundation. 2019. concurrent.futures module. https://docs.python.org/3/library/concurrent.futures.html.
[35] Thelwall, M., Buckley, K., Paltoglou. 2020. Sentistrength. http://sentistrength.wlv.ac.uk/.
[36] Wikipedia. 2020. Robert Plutchnik. https://en.wikipedia.org/wiki/Robert_Plutchik.
[37] Wikipedia. 2021. COVID-19 lockdowns. https://en.wikipedia.org/wiki/COVID-19_lockdowns.
[38] Wikipedia. 2021. Elsagate. https://en.wikipedia.org/wiki/Elsagate.
[39] Wikipedia. 2022. Beautiful Soup. https://en.wikipedia.org/wiki/Beautiful_Soup_(HTML_parser).
[40] Wikipedia. 2022. Children's Online Privacy Protection Act. https://en.wikipedia.org/wiki/Children%27s_Online_Privacy_Protection_Act.
[41] Wikipedia. 2022. Minecraft. https://en.wikipedia.org/wiki/Minecraft.
[42] Wikipedia. 2022. Motion Picture Association film rating system. https://en.wikipedia.org/wiki/Motion_Picture_Association_film_rating_system.
[43] Wikipedia. 2022. Roblox. https://en.wikipedia.org/wiki/Roblox.
[44] YouTube. 2019. Important Updates for All Creators: Complying with COPPA. https://www.youtube.com/watch?v=-JzXiSkoFKw&ab_channel=YouTubeCreators.
[45] YouTube. 2021. Determining if your content is "made for kids". https://support.google.com/youtube/answer/9528076.
[46] YouTube. 2022. Child Safety Policy. https://support.google.com/youtube/answer/2801999?hl=en.
[47] YouTube. 2022. YouTube Kids. https://www.youtube.com/kids/.
[48] YouTube Help Center. 2021. Business Inquiry Emails. https://support.google.com/youtube/answer/57955.
[49] YouTube Help Center. 2021. YouTube Trusted Flagger program. https://support.google.com/youtube/answer/7554338.
[50] YouTube Team. 2018. More information, faster removals, more people - an update on what we're doing to enforce YouTube's Community Guidelines. https://blog.youtube/news-and-events/more-information-faster-removals-more/ April 23, 2018.

A ETHICAL CONSIDERATIONS

The execution of this work has followed the principles and guidelines of how to perform ethical information research and the use of shared measurement data [11, 32]. In particular, this study paid attention to the following dimensions.

We keep our crawling to a minimum to ensure that we do not slow down or deteriorate the performance of the YouTube service in any way. Whenever possible, we used the recommended YouTube API v3. When the data to be crawled were not available by the API, we crawled the channel page directly. We do not interact with any component in each visited page. In addition to this, our crawler has been implemented to wait for both the page to fully load and an extra period of time before visiting another page. Also, we do not share any data collected by our crawler with any other entity .