Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children

Kostantinos Papadamou,∗ Antonis Papasavva,∗ Savvas Zannettou,∗ Jeremy Blackburn,† Nicolas Kourtellis,‖ Ilias Leontiadis,‖ Gianluca Stringhini,∥ Michael Sirivianos∗

∗Cyprus University of Technology, ∗Max-Planck-Institut für Informatik, †Binghamton University, ‖Telefonica Research, ∥Boston University

{ck.papadamou, as.papasavva}@edu.cut.ac.cy, szannett@mpi-inf.mpg.de, jblackbu@binghamton.edu
{nicholas.kourtellis, ilias.leontiadis}@telefonica.com, gian@bu.edu, michael.sirivianos@cut.ac.cy

Abstract

A large number of the most-subscribed YouTube channels target children of very young age. Hundreds of toddler-oriented channels on YouTube feature inoffensive, well produced, and educational videos. Unfortunately, inappropriate content that targets this demographic is also common. YouTube’s algorithmic recommendation system regrettably suggests inappropriate content because some of it mimics or is derived from otherwise appropriate content. Considering the risk for early childhood development, and an increasing trend in toddler’s consumption of YouTube media, this is a worrisome problem. In this work, we build a classifier able to discern inappropriate content that targets toddlers on YouTube with 84.3% accuracy, and leverage it to perform a large-scale, quantitative characterization that reveals some of the risks of YouTube media consumption by young children. Our analysis reveals that YouTube is still plagued by such disturbing videos and its currently deployed counter-measures are ineffective in terms of detecting them in a timely manner. Alarmingly, using our classifier we show that young children are not only able, but likely to encounter disturbing videos when they randomly browse the platform starting from benign videos.

1 Introduction

YouTube has emerged as an alternative to traditional children’s TV, and a plethora of popular children’s videos can be found on the platform. For example, consider the millions of subscribers that the most popular toddler-oriented YouTube channels have: ChuChu TV is the most-subscribed “child-themed” channel, with 19.9M subscribers (Statista 2018) as of September 2018. While most toddler-oriented content is inoffensive, and is actually entertaining or educational, recent reports have highlighted the trend of inappropriate videos for toddlers, disturbing videos still appear, even in YouTube Kids (Weston 2018), due to the difficulty of identifying them. An explanation for this is that YouTube relies heavily on users reporting videos they consider disturbing2, and then YouTube employees manually inspecting them. However, since the process involves manual labor, the whole mechanism does not easily scale to the amount of videos that a platform like YouTube serves.

In this paper, we provide the first study of toddler-oriented disturbing content on YouTube. For the purposes of this work, we extend the definition of a toddler to any child aged between 1 and 5 years. Our study comprises three phases. First, we aim to characterize the phenomenon of inappropriate videos geared towards toddlers. To this end, we collect,
manually review, and characterize toddler-oriented videos (both Elsagate-related and other child-related videos), as well as random and popular videos. For a more detailed analysis of the problem, we label these videos as one of four categories: 1) suitable; 2) disturbing; 3) restricted (equivalent to MPAA’s NC-17 and R categories); and 4) irrelevant videos (see Section 2). Our characterization confirms that unscrupulous and potentially profit-driven uploaders create disturbing videos with similar characteristics as benign toddler-oriented videos in an attempt to make them show up as recommendations to toddlers browsing the platform.

Second, we develop a deep learning classifier to automatically detect disturbing videos. Even though this classifier performs better than baseline models, it still has a lower than desired performance. In fact, this low performance reflects the high degree of similarity between disturbing and suitable videos or restricted videos that do not target toddlers. It also reflects the subjectivity in deciding how to label these controversial videos, as confirmed by our trained annotators’ experience. For the sake of our analysis in the next steps, we collapse the initially defined labels into two categories and develop a more accurate classifier that is able to discern inappropriate from appropriate videos. Our experimental evaluation shows that the developed binary classifier outperforms several baselines with an accuracy of 84.3%.

In the last phase, we leverage the developed classifier to understand how prominent the problem at hand is. From our analysis on different subsets of the collected videos, we find that 1.1% of the 233,337 Elsagate-related, and 0.5% of the 154,957 other children-related collected videos are inappropriate for toddlers, which indicates that the problem is not negligible. To further assess how safe YouTube is for toddlers, we run a live simulation in which we mimic a toddler randomly clicking on YouTube’s suggested videos. We find that there is a 3.5% chance that a toddler following YouTube’s recommendations will encounter an inappropriate video within ten hops if she starts from a video that appears among the top ten results of a toddler-appropriate keyword search (e.g., Peppa Pig).

Last, our assessment on YouTube’s current mitigations shows that the platform struggles to keep up with the problem: only 20.5% and 2.5% of our manually reviewed disturbing and restricted videos, respectively, have been removed by YouTube.

Contributions. In summary, our contribution is threefold:

1. We undertake a large-scale analysis of the disturbing videos problem that is currently plaguing YouTube.
2. We propose a reasonably accurate classifier that can be used to discern disturbing videos which target toddlers.
3. We make publicly available the classifier, the manually reviewed ground truth dataset that consists of 4,797 videos, and the metadata of all the collected and examined videos so that the research community can build on our results to further investigate the problem.

2 Methodology

Data Collection

For our data collection, we use the YouTube Data API, which provides metadata of videos uploaded on YouTube. Unfortunately, YouTube does not provide an API for retrieving videos from YouTube Kids. We collect a set of seed videos using four different approaches. First, we use information from /r/ElsaGate, a subreddit dedicated to raising awareness about disturbing videos problem (Reddit 2017). Second, we use information from /r/fullcartoonsonyoutube, a subreddit dedicated to listing cartoon videos available on YouTube. The other two approaches focus on obtaining a set of random and popular videos.

Specifically: 1) we create a list of 64 keywords by extracting n-grams from the title of videos posted on /r/ElsaGate. Subsequently, for each keyword, we obtain the first 30 videos as returned by YouTube’s Data API search functionality. This approach resulted in the acquisition of 893 seed videos. Additionally, we create a list of 33 channels, which are mentioned by users on /r/ElsaGate because of publishing disturbing videos (Brandom 2017; Reddit 2017). Then, for each channel we collect all their videos, hence acquiring a set of 181 additional seed videos; 2) we create a list of 83 keywords by extracting n-grams from the title of videos posted on /r/fullcartoonsonyoutube. Similar to the previous approach, for each keyword, we obtain the first 30 videos as returned by YouTube’s Data API search functionality, hence acquiring another 2,342 seed videos; 3) to obtain a random sample of videos, we use a REST API that provides random YouTube video identifiers which we then download using the YouTube Data API. This approach resulted in the acquisition of 8,391 seed random videos; and 4) we also collect the most popular videos in the USA, the UK, Russia, India, and Canada, between November 18 and November 21, 2018, hence acquiring another 500 seed videos.

Using these approaches, we collect 12,097 unique seed videos. However, this dataset is not big enough to study the idiosyncrasies of this problem. Therefore, to expand our dataset, for each seed video we iteratively collect the top 10 recommended videos associated with it, as returned by the YouTube Data API, for up to three hops within YouTube’s recommendation graph. We note that for each approach we use API keys generated from different accounts. Table 1 summarizes the collected dataset. In total, our dataset comprises 12K seed videos and 844K videos that are recommended from the seed videos. Note, that there is a small overlap between the videos collected across the approaches, hence the number of total videos is slightly smaller than the sum of all videos for the four approaches.

For each video in our dataset, we collect the following data descriptors: 1) title and description; 2) thumbnail; 3) tags; and 4) video statistics like number of views, likes,
dislikes, etc. We chose to use these four data collection approaches for three reasons: 1) to get more breadth into children’s content on YouTube, instead of only collecting Elsagate-related videos; 2) to examine and analyze different types of videos while also assessing the degree of the disturbing videos problem in these types of videos; and 3) to train a classifier for detecting disturbing videos able to generalize to the different types of videos available on YouTube.

Ethics. For this study we only collect publicly available data, while making no attempt to de-anonymize users. In addition, all the manual annotators are informed adults.

Manual Annotation Process
To get labeled data, we manually review a 5K videos subset of the collected dataset by inspecting their video content, title, thumbnail, and tags. Each video is presented to three annotators that inspect its content and metadata to assign one of the following labels:

Suitable: A video is suitable when its content is appropriate for toddlers (aged 1-5 years) and it is relevant to their typical interests. Some examples include normal cartoon videos, children’s songs, and educational videos (e.g., learning colors). In other words, any video that can be classified as G by the MPAA and its target audience is toddlers.

Disturbing: A video is disturbing when it targets toddlers but it contains sexual hints, sexually explicit or abusive/inappropriate language, graphic nudity, child abuse (e.g., children hitting each other), scream and horror sound effects, scary scenes or characters (e.g., injections, attacks by insects, etc.). In general, any video targeted at toddlers that should be classified as PG, PG-13, NC-17, or R by MPAA is considered disturbing.

Restricted: We consider a video restricted when it does not target toddlers and it contains content that is inappropriate for individuals under the age of 17 (rated as R or NC-17 according to MPAA). Such videos usually contain sexually explicit language, graphic nudity, pornography, violence (e.g., gaming videos featuring violence, or life-like violence, etc.), abusive/inappropriate language, online gambling, drug use, alcohol, or upsetting situations and activities.

Irrelevant: We consider a video irrelevant when it contains appropriate content that is not relevant to a toddler’s interests. That is, videos that are not disturbing or restricted but are only suitable for school-aged children (aged 6-11 years), adolescents (aged 12-17 years) and adults, like gaming videos (e.g., Minecraft or music videos (e.g., a video clip of John Legend’s song) reside in this category. In general, G, PG and PG-13 videos that do not target toddlers are considered irrelevant.

We elect to use these labels for our annotation process instead of adopting the five MPAA ratings for two reasons. First, our scope is videos that would normally be rated as PG, PG-13, R, and NC-17 but target very young audiences. We consider such targeting a malevolent activity that needs to be treated separately. At the same time, we have observed that a significant portion of videos that would normally be rated as R or NC-17 are already classified by YouTube as “age-restricted” and target either adolescents or adults. Second, YouTube does not use MPAA ratings to flag videos, thus, a ground truth dataset with such labels is not available.

Sampling Process. Our aim is to create a ground truth dataset that enables us to: 1) understand the main characteristics of disturbing toddler-oriented videos compared to suitable children videos on YouTube; and 2) train a deep learning model that will detect disturbing videos with an acceptable performance while being able to generalize to the various types of videos available on the platform. To this end, we use the following videos for the annotation process. 1) We randomly select 1,000 of the 2,342 seed child-related videos aiming to get suitable videos. 2) Since the Elsagate-related collected videos are likely to include disturbing videos, we select all the seed Elsagate-related videos (1,074), as well as a small, randomly selected set (1,171) of their recommended videos. 3) To get a sample of restricted videos, we randomly select 500 of the 2,597 age-restricted videos in our dataset. 4) To ensure that we include irrelevant videos, we select all the seed popular videos (500) as well as a small set (1,000) of the 8,391 randomly collected videos.

Manual Annotation. The annotation process is carried out by two of the authors of this study and 76 undergraduate students (aged 20-24 years), both male and female. Each video is annotated by the two authors and one of the undergraduate students. The students come from different backgrounds and receive no specific training with regard to our study. To ease the annotation process, we develop a platform that includes a clear description of the annotation task, our labels, as well as all the video information that an annotator needs in order to inspect and correctly annotate a video.

After obtaining all the annotations, we compute the Fleiss agreement score ($\kappa$) (Fleiss 1971) across all annotators: we find $\kappa = 0.60$, which is considered “moderate” agreement. We also assess the level of agreement between the two authors, as we consider them experienced annotators, finding $\kappa = 0.69$, which is considered “substantial” agreement. Finally, for each video we assign one of the labels according to the majority agreement of all the annotators, except a small percentage (4%) where all annotators disagreed, which we

<table>
<thead>
<tr>
<th>Crawling Strategy</th>
<th>#Seed</th>
<th>#Recommended</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elsagate-related</td>
<td>1,074</td>
<td>232,263</td>
</tr>
<tr>
<td>Other Child-related</td>
<td>2,342</td>
<td>152,615</td>
</tr>
<tr>
<td>Random</td>
<td>8,391</td>
<td>473,516</td>
</tr>
<tr>
<td>Popular</td>
<td>500</td>
<td>10,474</td>
</tr>
<tr>
<td>Total</td>
<td>12,097</td>
<td>844,702</td>
</tr>
</tbody>
</table>

Table 1: Overview of the collected data: number of seed videos and number of their recommended videos acquired using each crawling strategy.

<table>
<thead>
<tr>
<th></th>
<th>#Suitable</th>
<th>#Disturbing</th>
<th>#Restricted</th>
<th>#Irrelevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elsagate-related</td>
<td>805</td>
<td>857</td>
<td>324</td>
<td>394</td>
</tr>
<tr>
<td>Other</td>
<td>650</td>
<td>47</td>
<td>21</td>
<td>243</td>
</tr>
<tr>
<td>Child-related</td>
<td>805</td>
<td>857</td>
<td>324</td>
<td>394</td>
</tr>
<tr>
<td>Random</td>
<td>27</td>
<td>5</td>
<td>67</td>
<td>867</td>
</tr>
<tr>
<td>Popular</td>
<td>31</td>
<td>20</td>
<td>2</td>
<td>432</td>
</tr>
<tr>
<td>Total</td>
<td>1,513</td>
<td>929</td>
<td>419</td>
<td>1,936</td>
</tr>
</tbody>
</table>

Table 2: Summary of our final ground truth dataset.
ular cartoons like Peppa Pig, Mickey and Minnie mouse,

of videos in our ground truth dataset. Unsurprisingly, the portion for each class of videos for all the different sets shows the top 15 stems found in titles along with their proportion using the Porter Stemmer algorithm. Fig. 2

ing or restricted videos on YouTube. First, we pre-process
tics and terms that are usually used by uploaders of disturb-

the titles in our ground truth dataset to understand the tac-
disturbing videos in the Education and Music categories, we

also exclude from our ground truth dataset. Table 2 summa-
rizes our ground truth dataset, which includes 1,513 suitable, 929 disturbing, 419 restricted, and 1,936 irrelevant videos.

<table>
<thead>
<tr>
<th>Category</th>
<th>Suitable</th>
<th>Disturbing</th>
<th>Restricted</th>
<th>Irrelevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elsagate-related</td>
<td>EN 353 (44%)</td>
<td>208 (24%)</td>
<td>99 (31%)</td>
<td>98 (25%)</td>
</tr>
<tr>
<td></td>
<td>F&amp;A 130 (16%)</td>
<td>190 (22%)</td>
<td>39 (12%)</td>
<td>33 (8%)</td>
</tr>
<tr>
<td></td>
<td>ED 126 (16%)</td>
<td>21 (3%)</td>
<td>17 (5%)</td>
<td>16 (4%)</td>
</tr>
<tr>
<td></td>
<td>P&amp;B 109 (13%)</td>
<td>239 (28%)</td>
<td>71 (22%)</td>
<td>73 (19%)</td>
</tr>
<tr>
<td></td>
<td>M 21 (3%)</td>
<td>15 (2%)</td>
<td>8 (3%)</td>
<td>45 (11%)</td>
</tr>
</tbody>
</table>

| Other              | EN 131 (20%) | 7 (1%)  | 9 (43%) | 51 (21%)  |
|                    | F&A 317 (49%) | 27 (58%) | 3 (14%) | 26 (11%)  |
|                    | ED 27 (4%) | 1 (2%)  | 2 (10%) | 34 (14%)  |
|                    | P&B 130 (20%) | 4 (8%)  | 29 (9%) | 35 (14%)  |
|                    | M 5 (1%) | 2 (4%) | 0 (0%)  | 26 (11%)  |

| Random             | EN 4 (15%) | 1 (20%) | 3 (5%) | 68 (8%) |
|                    | F&A 1 (4%) | 1 (20%) | 1 (2%) | 18 (2%) |
|                    | ED 1 (4%) | 0 (0%) | 0 (0%) | 31 (4%)  |
|                    | P&B 13 (48%) | 3 (60%) | 21 (31%) | 354 (41%) |
|                    | M 1 (3%) | 0 (0%) | 0 (0%) | 79 (9%)  |

| Popular            | EN 12 (39%) | 9 (45%) | 2 (29%) | 168 (39%) |
|                    | F&A 9 (29%) | 7 (35%) | 0 (0%) | 13 (3%)  |
|                    | ED 2 (7%) | 0 (0%) | 0 (0%) | 11 (3%)   |
|                    | P&B 2 (6%) | 0 (0%) | 0 (0%) | 32 (7%)  |
|                    | M 0 (0%) | 1 (5%) | 0 (0%) | 63 (15%)  |

Table 3: Number of videos in each category per class for each subset of videos in our ground truth dataset. EN: Entertainment, F&A: Film & Animation, ED: Education, P&B: People & Blogs, M: Music.

also from the thumbnails of videos in our ground truth. In

also exclude from our ground truth dataset. Table 2 summa-
rizes our ground truth dataset, which includes 1,513 suitable, 929 disturbing, 419 restricted, and 1,936 irrelevant videos.

Ground Truth Dataset Analysis

Category. First, we look at the categories of the videos in our ground truth dataset. Table 3 reports the top five categories, for each subset of videos. Most of the disturbing and restricted in the Elsagate-related videos are in Entertainment (24% and 31%), People & Blogs (28% and 22%), and Film & Animation (22% and 12%). These results are similar with previous work (Chaudhary and Sureka 2013). A similar trend is also observed in all the other sets of videos. In addition, in the Elsagate-related videos we find a non-negligible percentage of disturbing videos in seemingly innocent categories like Education (3%) and Music (2%). This is alarming since it indicates that disturbing videos “infiltrate” categories of videos that are likely to be selected by the toddler’s parents. Unsurprisingly, after manually inspecting all the disturbing videos in the Education and Music categories, we find that the majority of them are nursery rhymes, “wrong heads”, and “peppa pig” videos with disturbing content.

Title. The title of a video is an important factor that affects whether a particular video will be recommended when viewing other toddler-oriented videos. Consequently, we study the titles in our ground truth dataset to understand the tactics and terms that are usually used by uploaders of disturbing or restricted videos on YouTube. First, we pre-process the title by tokenizing the text into words and then we perform stemming using the Porter Stemmer algorithm. Fig. 2 shows the top 15 stems found in titles along with their proportion for each class of videos for all the different sets of videos in our ground truth dataset. Unsurprisingly, the top 15 stems of the Elsagate-related videos refer to popular cartoons like Peppa Pig, Mickey and Minnie mouse, Elsa, and Spiderman (see Fig. 2(a)). When looking at the results, we observe that a substantial percentage of the videos that include these terms in their title are actually disturbing. For example, from the videos that contain the terms “spiderman” and “mous”, 82.6% and 80.4%, respectively, are disturbing. Similar trends are observed with other terms like “peppa” (78.6%), “superhero” (76.7%), “pig” (76.4%), “frozen” (63.5%), and “elsa” (62.5%). Also, we observe a small percentage of the other child-related videos that contain the terms “felix” (7.1%), “cat” (4.2%), and “cartoon” (3.8%) are also disturbing (see Fig. 2(b)). These results reveal that disturbing videos on YouTube refer to seemingly “innocent” cartoons in their title, but in reality the content of the video is likely to be either restricted or disturbing. Note that we find these terms in suitable videos too. This demonstrates that it is quite hard to distinguish suitable from disturbing videos by only inspecting their titles.

Tags. Tags are words that uploaders define when posting a video on YouTube. To study the effect of tags in this problem, we plot in Fig. 3 the top 15 stems from tags found in each subset of videos in our ground truth dataset. We make several observations: first, in the Elsagate-related and other child-related videos there is a substantial overlap between the stems found in the tags and title (cf. Fig. 2 and Fig. 3). Second, in the Elsagate-related videos we find that suitable and disturbing classes have a considerable percentage for each tag, hence highlighting that Elsagate-related disturbing videos use the same tags as suitable videos. Inspecting these results, we find that the tags “funni” (47.8%), “elsa” (58.7%), “frozen” (57.8%), “cartoon” (48.8%), and “anim” (44.5%) appear mostly in disturbing videos. Also, “spiderman” (74.4%) and “mous” (73.0%) appear to have a higher portion of disturbing videos than the other tags (see Fig. 3(a)). Third, we observe that the tags “mous” (73.0%), “anim” (44.5%), “cartoon” (48.8%), “video” (31.5%), “disney” (36.5%), and “kid” (34.2%) that appear in a considerable number of disturbing Elsagate-related videos also appear in a high portion of suitable other child-related videos (cf. Fig. 3(a) and Fig. 3(b)). The main take-away from this analysis is that it is hard to detect disturbing content just by looking at the tags, and that popular tags are shared among disturbing and suitable videos.

Thumbnails. To study the thumbnails of the videos in our ground truth dataset, we make use of the Google Cloud Vision API11, which is a RESTful API that derives useful insights from images using pre-trained machine learning models. Using this API we are able to: 1) extract descriptive labels for all the thumbnails in our ground truth; and 2) check whether a modification was made to a thumbnail, and whether a thumbnail contains adult, medical-related, violent, and/or racy content. Fig. 4 depicts the top 15 labels derived from the thumbnails of videos in our ground truth. In the Elsagate-related case, we observe that the thumbnails of disturbing videos contain similar entities as the thumbnails of both the Elsagate-related and other child-related suitable videos (cartoons, fictional characters, etc.).

Fig. 5 shows the proportion of each class for videos that

11https://cloud.google.com/vision/
contain spoofed, adult, medical-related, violent, and/or racy content. As expected, most of the Elsagate-related videos whose thumbnails contain adult (54.9%) and medical content (53.5%) are restricted. However, this is not the case for videos whose thumbnails contain spoofed (47.4%), violent (60.0%) or racy (34.8%) content, where we observe a high number of disturbing videos (cf. Fig. 5(a)). Surprisingly, we notice that 100.0% of the other child-related videos whose thumbnail contains violent content are suitable. Nonetheless, after manually inspecting some of those thumbnails we notice that they depict mild cartoon violence (i.e., tom hitting jerry), which we consider as suitable. In general, we observe that Elsagate-related videos whose thumbnail is modified with violent, racy, medical, and/or adult content are more likely to be restricted or disturbing, while this is not the case for the other child-related videos.

Statistics. Next, we examine statistics that pertain to the videos in our ground truth dataset. Fig. 6 shows the CDF of the number of views of all the videos in each distinct subset of videos in our ground truth. We observe that Elsagate-related suitable videos have more views than disturbing videos while this is not the case for all the other types of videos. Fig. 7 shows the CDF of the fraction of likes of all the videos in each subset. Interestingly, we observe that in
Figure 5: Per class proportion of videos that their thumbnail contains spoofed, adult, medical, violent, and/or racy content for (a) Elsagate-related; (b) other child-related; (c) random; and (d) popular videos.

Figure 6: CDF of the number of views per class for (a) Elsagate-related (b) other child-related, (c) random, and (d) popular videos.

Figure 7: CDF of the fraction of likes to dislikes per class for (a) Elsagate-related (b) other child-related, (c) random, and (d) popular videos.

Figure 8: CDF of the number of comments/views per class for (a) Elsagate-related (b) other child-related, (c) random, and (d) popular videos.
all cases disturbing and restricted videos have a higher fraction of likes compared to suitable videos, which, particularly in the case of disturbing videos, indicates manipulation to boost their ranking. Lastly, Fig. 8 shows the CDF of the fraction of comments to views. Although, for the Elsegate-related videos the suitable and disturbing videos have a similar ratio of comments, the situation shifts when it comes to all the other types of videos where we observe a higher ratio of comments for disturbing and restricted videos compared to suitable videos.

A general take away from this ground truth analysis is that none of the videos’ metadata can clearly indicate that a video is disturbing or not, thus, in most cases one (e.g., a guardian) has to carefully inspect all the available video metadata, and potentially the actual video, to accurately determine if it is safe for a toddler to watch.

Assessing YouTube’s Counter-measures. To assess how fast YouTube detects and removes inappropriate videos, we leverage the YouTube Data API to count the number of offline videos (either removed by YouTube due to a Terms of Service violation or deleted by the uploader) in our manually reviewed ground truth dataset. We note that we do not consider the videos that were already marked as age-restricted, since YouTube took the appropriate measures.

As of May 10, 2019 only 9.65% of the suitable, 20.5% of the disturbing, 2.5% of the restricted, and 2.4% of the irrelevant videos were removed, while from those that were still available, 0.0%, 6.9%, 1.3%, and 0.1%, respectively, were marked as age-restricted. Alarmingly, the amount of the deleted disturbing and restricted videos, is considerably low. The same stands for the amount of disturbing and restricted videos marked as age-restricted. A potential issue here is that the videos on our dataset were recently uploaded and YouTube simply did not have time to detect them. To test this hypothesis, we calculate the mean number of days from publication up to May, 2019: we find this hypothesis does not hold. The mean number of days since being uploaded for the suitable, disturbing, restricted, and irrelevant videos is 866, 942, 1091, and 991, respectively, with a mean of 947 days across the entire manually reviewed ground truth dataset. This indicates that YouTube’s deployed counter-measures eliminated some of the disturbing videos, but they are unable to tackle the problem in a timely manner.

3 Detection of Disturbing Videos

Dataset and Feature Description

To train and test our proposed deep learning model we use our ground truth dataset of 4,797 videos, summarized in Table 2. For each video in our ground truth our model processes the following:

Title. Our model considers the text of the title by training an embedding layer, which encodes each word in the text in an N-dimensional vector space. The maximum number of words found in a title of videos in our ground truth is 21, while the size of the vocabulary is 12,023.

Tags. Similarly to the title, we encode the video tags into an N-dimensional vector space by training a separate embedding layer. The maximum number of tags found in a video is 78, while the size of the word vocabulary is 40,096.

Thumbnail. We scale down the thumbnail images to 299x299 while preserving all three color channels.

Statistics. We consider all available statistical metadata for videos (number of views, likes, dislikes, and comments).

Style Features. We consider some style features from the actual video (e.g., duration), the title (e.g., number of bad words), the video description (e.g., description length), and the tags (e.g., number of tags). For this we use features proposed in (Kaushal et al. 2016) that help the model to better differentiate between the videos of each class. Table 4 summarizes the style features that we use.

Model Architecture

Fig. 9 depicts the architecture of our classifier, which combines the above mentioned features. Initially, the classifier consists of four different branches, where each branch processes a distinct feature type: title, tags, thumbnail, and statistics and style features. Then the outputs of all the branches are concatenated to form a two-layer, fully connected neural network that merges their output and drives the final classification.

The title feature is fed to a trainable embedding layer that outputs a 32-dimensional vector for each word in the title text. Then, the output of the embedding layer is fed to a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) Recurrent Neural Network (RNN) that captures the relationships between the words in the title. For the tags, we use an architecturally identical branch trained separately from the title branch.

For thumbnails, due to the limited number of training examples in our dataset, we use transfer learning (Oquab et al. 2014) and the pre-trained Inception-v3 Convolutional Neural Network (CNN) (Szegedy et al. 2015), which is built from the large-scale ImageNet dataset.12 We use the pre-trained CNN to extract a meaningful feature representation (2,048-dimensional vector) of each thumbnail. Last, the statistics together with the style features are fed to a fully-connected dense neural network comprising 25 units.

The second part of our classifier is essentially a two-layer, fully-connected dense neural network. At the first layer,

Table 4: List of the style features extracted from the available metadata of a video.

<table>
<thead>
<tr>
<th>Type</th>
<th>Style Features Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video-related</td>
<td>video category, video duration</td>
</tr>
<tr>
<td>Statistics-related</td>
<td>ratio of # of likes to dislikes</td>
</tr>
<tr>
<td>Title-related &amp;</td>
<td>length of title, length of description,</td>
</tr>
<tr>
<td>description-related</td>
<td>ratio of description to title,</td>
</tr>
<tr>
<td></td>
<td>jaccard sim. of title &amp; description,</td>
</tr>
<tr>
<td></td>
<td># &quot;!&quot; and &quot;?&quot; in title &amp; description,</td>
</tr>
<tr>
<td>Tags-related</td>
<td># emoticons in title &amp; description,</td>
</tr>
<tr>
<td></td>
<td># bad words in title &amp; description,</td>
</tr>
<tr>
<td></td>
<td># child-related words in title &amp; description</td>
</tr>
<tr>
<td></td>
<td># tags, # bad words in tags,</td>
</tr>
<tr>
<td></td>
<td># child-related words in tags,</td>
</tr>
<tr>
<td></td>
<td>jaccard sim. of tags &amp; title</td>
</tr>
</tbody>
</table>

12http://image-net.org/
Adam algorithm with an initial learning rate of \(10^{-5}\) to oversample the train set at each fold. Over-sampling technique (SMOTE) (Chawla et al. 2002) to deal with the data imbalance problem we use the Synthetic Minority model using all the aforementioned features. To deal with imbalanced classes, we use five-fold stratified cross-validation (ArLOT, Celisse, and others 2010) and we train and test our model using Keras (Chollet 2015) with TensorFlow as the backend (Abadi et al. 2016). To train our model, we use the proposed deep learning model.

### Experimental Evaluation

We implement our model using Keras (Chollet 2015) with TensorFlow as the backend (Abadi et al. 2016). To train our model we use five-fold stratified cross-validation (ArLOT, Celisse, and others 2010) and we train and test our model using all the aforementioned features. To deal with the data imbalance problem we use the Synthetic Minority Over-sampling technique (SMOTE) (Chawla et al. 2002) to oversample the train set at each fold.

For the stochastic optimization of our model, we use the Adam algorithm with an initial learning rate of \(10^{-5}\), and \(\epsilon = 10^{-8}\). To evaluate our model, we compare it in terms of accuracy, precision, recall, F1 score, and area under the ROC curve (AUC) against the following five baselines: 1) a Support Vector Machine (SVM) with parameters \(\gamma = \text{auto}\) and \(C = 10.0\); 2) a K-Nearest Neighbors classifier with \(n = 8\) neighbors and leaf size equal to 10; 3) a Bernoulli Naïve Bayes classifier with \(n = 1.0\); 4) a Decision Tree classifier with an entropy criterion; and 5) a Random Forest classifier with an entropy criterion and number of trees equal to 100. To further evaluate the performance of our model, we also compare it with two deep neural networks: 1) a simple double dense layer network (DDNN); and 2) a CNN combined with a double dense layer network (CNN-DDNN). For hyper-parameter tuning of all the baselines we use the grid search strategy, while for the deep neural networks we use the same hyper-parameters as with the proposed model. For a fair comparison, we note that all the evaluated models use all the available input features. Table 5 reports the performance of the proposed model as well as the 7 base-lines, while Fig. 10 shows their ROC curves. Although the proposed model outperforms all the baselines in all performance metrics, it still has poor performance.

#### Ablation Study

In an attempt to understand which of the input feature types contribute the most to the classification of disturbing videos we perform an ablation study. That is, we systematically remove each of the four input feature types (as well as their associated branch in the proposed model’s architecture), while also training models with all the possible combinations of the four input feature types. Again, to train and test these models we use five-fold cross validation and the oversampling technique to deal with data imbalance. Table 6 shows the performance metrics of all the models for each possible combination of inputs. We observe that the thumbnail, is more important than the other input feature types for good classification performance.
First, we assess the prevalence of inappropriate videos in appropriate videos on YouTube using our binary classifier. In this section, we study the interplay of appropriate and inappropriate videos to understand their impact on children’s exposure. We train and evaluate the proposed model for binary classification and compare its performance with several baseline models.

Table 7 shows the performance of our proposed model and other evaluated baselines. Our model outperforms all baseline models across all performance metrics, especially for accuracy, precision, recall, F1 score, and AUC. Specifically, our model substantially outperforms the CNN-DDNN model, which has the best overall performance from all the evaluated baselines, on accuracy, precision, recall, F1 score, and AUC by 12.3%, 13.3%, 16.6%, 13.9%, 11.0% respectively.

**Table 7: Performance of the evaluated baselines trained for binary classification and of our proposed binary classifier.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Nearest</td>
<td>0.610</td>
<td>0.343</td>
<td>0.424</td>
<td>0.380</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.678</td>
<td>0.442</td>
<td>0.562</td>
<td>0.495</td>
</tr>
<tr>
<td>SVM</td>
<td>0.718</td>
<td>0.471</td>
<td>0.029</td>
<td>0.054</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.728</td>
<td>0.522</td>
<td>0.363</td>
<td>0.428</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.804</td>
<td>0.740</td>
<td>0.464</td>
<td>0.569</td>
</tr>
<tr>
<td>DDNN</td>
<td>0.734</td>
<td>0.662</td>
<td>0.629</td>
<td>0.637</td>
</tr>
<tr>
<td>CNN-DDNN</td>
<td>0.720</td>
<td>0.688</td>
<td>0.724</td>
<td>0.690</td>
</tr>
<tr>
<td>Proposed Model</td>
<td>0.843</td>
<td>0.821</td>
<td>0.890</td>
<td>0.829</td>
</tr>
</tbody>
</table>

**Figure 11: ROC Curves of all the baselines and of the proposed model trained for binary classification.**

**Binary Classification.** To perform a more representative analysis of the inappropriate videos on YouTube, we need a more accurate classifier. Thus, for the sake of our analysis in the next steps, we collapse our four labels into two general categories, by combining the suitable with the irrelevant videos into one “appropriate” category (3,499 videos) and the disturbing and restricted videos into a second “inappropriate” category (1,348 videos). We call the first category “appropriate” despite including PG and PG-13 videos because those videos are not aimed at toddlers (irrelevant). On the other hand, videos rated as PG or PG-13 that target toddlers (aged 1 to 5) are disturbing and fall under the inappropriate category. When such videos appear on the video recommendation list of toddlers, it is likely to expose their child to inappropriate videos.

We train and evaluate the proposed model for binary classification on our reshaped ground truth dataset following the same approach as the one presented above. Table 7 reports the performance of our model as well as the baselines, while Fig. 11 shows their ROC curves. We observe that our deep learning model outperforms all baseline models across all performance metrics. Specifically, our model substantially outperforms the CNN-DDNN model, which has the best overall performance from all the evaluated baselines, on accuracy, precision, recall, F1 score and AUC by 12.3%, 13.3%, 16.6%, 13.9%, 11.0% respectively.

**4 Analysis**

In this section, we study the interplay of appropriate and inappropriate videos on YouTube using our binary classifier. First, we assess the prevalence of inappropriate videos in each subset of videos in our dataset and investigate how likely it is for YouTube to recommend an inappropriate video. Second, we perform live random walks on YouTube’s recommendation graph to simulate the behavior of a toddler that selects videos based on the recommendations.

**Recommendation Graph Analysis**

First, we investigate the prevalence of inappropriate videos in each subset of videos in our dataset by running our binary classifier on the whole dataset, which allows us to find which videos are inappropriate or appropriate. Table 8 shows the number of appropriate and inappropriate videos found in each subset. For the Elsagate-related videos, we find 231K (98.9%) appropriate videos and 2.5K (1.1%) inappropriate videos, while the proportion of inappropriate videos is a bit lower in the set of other child-related videos (0.4% inappropriate and 99.5% appropriate). These findings highlight the gravity of the problem: a parent searching on YouTube with simple toddler-oriented keywords and casually selecting from the recommended videos, is likely to expose their child to inappropriate videos.

But what is the interplay between the inappropriate and appropriate videos in each subset? To shed light to this question, we create a directed graph for each subset of videos, where nodes are videos, and edges are recommended videos (up to 10 videos due to our data collection methodology). For instance, if video2 is recommended via video1 then we add an edge from video1 to video2. Then, for each video in each graph, we calculate the out-degree in terms of appropriate and inappropriate labeled nodes. From here, we can count the number of transitions the graph makes between differently labeled nodes. Table 9 summarizes the percentages of transitions between the two classes of videos in each subset. Unsurprisingly, we find that most of the transitions in each subset (98%-99%), are between appropriate videos, which is mainly because of the large number of appropriate videos in each set. We also find that when a toddler watches an Elsagate-related benign video, if she randomly follows one of the top ten recommended videos, there is a 0.6% probability that she will end up at a disturbing or restricted video. Taken altogether, these findings show that the problem of toddler-oriented inappropriate videos on YouTube is notable, especially when considering YouTube’s massive scale and the large number of views of toddler-oriented videos. That is, there is a non-negligible chance that a toddler will be recommended an inappropriate video when watching an appropriate video.

**Table 8: Number of appropriate and inappropriate videos found in each subset of videos in our dataset.**

<table>
<thead>
<tr>
<th>Videos subset</th>
<th>Appropriate (%)</th>
<th>Inappropriate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elsagate-related</td>
<td>230,890 (98.95%)</td>
<td>2,447 (1.05%)</td>
</tr>
<tr>
<td>Other Child-related</td>
<td>154,262 (99.55%)</td>
<td>695 (0.45%)</td>
</tr>
<tr>
<td>Random</td>
<td>478,420 (99.28%)</td>
<td>3,487 (0.72%)</td>
</tr>
<tr>
<td>Popular</td>
<td>10,947 (99.75%)</td>
<td>27 (0.25%)</td>
</tr>
</tbody>
</table>

**Table 9: Number of transitions the graph makes between differently labeled nodes.**

<table>
<thead>
<tr>
<th>Videos subset</th>
<th>Appropriate (%)</th>
<th>Inappropriate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elsagate-related</td>
<td>230,890 (98.95%)</td>
<td>2,447 (1.05%)</td>
</tr>
<tr>
<td>Other Child-related</td>
<td>154,262 (99.55%)</td>
<td>695 (0.45%)</td>
</tr>
<tr>
<td>Random</td>
<td>478,420 (99.28%)</td>
<td>3,487 (0.72%)</td>
</tr>
<tr>
<td>Popular</td>
<td>10,947 (99.75%)</td>
<td>27 (0.25%)</td>
</tr>
</tbody>
</table>
used to seed them. Fig. 12(a) shows the cumulative per-
dom walks for each seed keyword, while at the same time
single random walk. We repeat this operation for 100 ran-
cess until we reach ten hops, which constitutes the end of a
video and select one randomly. We iterate with the same pro-
top ten results. Then, we obtain the recommendations of the
query on YouTube and randomly select one video from the
are not biasing the search results with any sensitive words.
appropriate words13. We do this to assess the degree of the prob-
the Elsagate-related keywords using a dictionary of inappro-
construct by stripping all the inappropriate words from all
seed keywords used for constructing our dataset, as well as
we use the lists of Elsagate-related and other child-related
eral videos according to the recommendations. To do this,
searches the platform for a video and then he watches sev-
allow us to simulate the behavior of a “random toddler” who
random walks on YouTube’s recommendation graph. This
prominent the problem is on a larger scale, we perform live
Y ouTube’s recommendation graph. Therefore, to assess how
encountered at each hop for: (a) Elsagate-related, sanitized
Elsagate-related, and other child-related seed keywords; and
(b) clusters of seed keywords.

How likely is it for a toddler to come across
inappropriate videos?

In the previous section, we showed that the problem of
toddler-oriented videos is prevalent enough to be cause for
concern. However, it is unclear whether the previous re-
sults generalize to YouTube at large since our dataset is
based on a snowball sampling up to three hops from a set of
seed videos. In reality though, YouTube comprises billions
of videos, which are recommended over many hops within
YouTube’s recommendation graph. Therefore, to assess how
prominent the problem is on a larger scale, we perform live
random walks on YouTube’s recommendation graph. This
allow us to simulate the behavior of a “random toddler” who
searches the platform for a video and then he watches sev-
eral videos according to the recommendations. To do this,
we use the lists of Elsagate-related and other child-related
seed keywords used for constructing our dataset, as well as
a list of sanitized Elsagate-related seed keywords which we
construct by stripping all the inappropriate words from all
the Elsagate-related keywords using a dictionary of inappro-
priate words11. We do this to assess the degree of the prob-
lem around Elsagate-related videos while ensuring that we
are not biasing the search results with any sensitive words.

For each seed keyword, we initially perform a search
query on YouTube and randomly select one video from the
top ten results. Then, we obtain the recommendations of the
video and select one randomly. We iterate with the same pro-
cess until we reach ten hops, which constitutes the end of a
single random walk. We repeat this operation for 100 ran-
dom walks for each seed keyword, while at the same time
classifying each video we visit, using our binary classifier.

First, we group the random walks based on the keywords
used to seed them. Fig. 12(a) shows the cumulative per-
centage of inappropriate videos encountered at each hop of
the random walks for Elsagate-related, sanitized Elsagate-
related, and other child-related search keywords. We observe
that, when using sanitized Elsagate-related keywords, we
find at least one inappropriate video in 3.5% of the walks,
while for the other child-related keywords we find at least
one inappropriate video in 1.3% of the walks. We also ob-
serve that most of the inappropriate videos are found early
in our random walks (i.e., at the first hop) and this num-
ber decreases as the number of hops increases. These find-
ings highlight that the problem of inappropriate videos on
YouTube emerges quite early when users are browsing the
platform starting from benign toddler-oriented search terms.

Next, to assess whether our results change according to
the content of the videos we use the k-means clustering algo-
ithm (Hartigan and Wong 1979) to create clusters from all
the seed keywords. Then, we manually inspect the clusters
and associate a label to each cluster. We create five clusters:
1) “Elsa and Spiderman” (24 keywords); 2) “Peppa Pig” (9
keywords); 3) “Finger Family” (6 keywords); 4) “Minnie
and Mickey mouse” (5 keywords); and 5) “Other Cartoon”
(11 keywords). Then, based on the clusters we group the ran-
dom walks. Fig. 12(b) shows the cumulative percentage of
inappropriate videos encountered at each hop for each clus-
tor. We observe interesting differences across the clusters:
specifically, we observe the higher percentages in the “Elsa
and Spiderman” (3.2%), and “Peppa Pig” (2.9%) cluster,
whereas for the clusters “finger family” (1.0%) and “other
cartoon” (0.5%) we observe a lower percentage of walks
with inappropriate videos. Also, we find that most of the in-
appropriate videos are found in the beginning of the random
walks in particular for the clusters “Peppa pig” (2.3%) and
“Elsa and Spiderman” (1.3%) (see first hop in Fig. 12(b)).

Note that, by merging the two classes in the binary clas-
sifier while seeking out disturbing videos with short ran-
dom walks from suitable videos, we correct for the mis-
classification of disturbing videos as restricted. That is,
an NC-17 video in the proximity of benign toddler-oriented
videos could be erroneously classified as restricted by the
multi-class classifier (because of similarities with the NC-
17 videos that do not target toddlers). However, due to this
proximity, this is probably an NC-17 video that targets tod-
dlers and should have therefore been classified as disturb-
ing. Thus, the vast majority of inappropriate videos detected
during the random walks are expected to be disturbing. In
fact, 84.6% of the detected inappropriate videos are disturb-
ing (obtained by inspecting all the 338 detected inappropriate
videos). On the other hand, videos that would be classi-
cified as irrelevant by the multi-class classifier, fall under
the appropriate category of the binary classifier. However,
the training set for the appropriate category includes irrele-

<table>
<thead>
<tr>
<th>Source</th>
<th>Elsagate-related (%)</th>
<th>Other Child-related (%)</th>
<th>Random (%)</th>
<th>Popular (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appropriate</td>
<td>917,319 (97.80%)</td>
<td>648,406 (99.49%)</td>
<td>1,319,518 (98.82%)</td>
<td>34,764 (99.12%)</td>
</tr>
<tr>
<td>Appropriate</td>
<td>5,951 (0.64%)</td>
<td>1,681 (0.26%)</td>
<td>7,014 (0.53%)</td>
<td>64 (0.18%)</td>
</tr>
<tr>
<td>Inappropriate</td>
<td>14,202 (1.51%)</td>
<td>1,542 (0.24%)</td>
<td>7,946 (0.59%)</td>
<td>246 (0.70%)</td>
</tr>
<tr>
<td>Inappropriate</td>
<td>478 (0.05%)</td>
<td>72 (0.01%)</td>
<td>831 (0.06%)</td>
<td>0 (0.00%)</td>
</tr>
</tbody>
</table>

Table 9: Number of transitions between appropriate and inappropriate videos for each subset of videos in our dataset.

1https://tinyurl.com/yxb4kmxg
vant videos, which include PG and PG-13 videos that do not target toddlers. Therefore, the binary classifier may classify such a video that is in the proximity of suitable videos as appropriate. However, a PG and PG-13 video in the proximity of suitable videos is likely to actually be disturbing, thus inappropriate. This negatively affects the accuracy of the binary classifier. Yet, only 1.6% of the videos encountered during the random walks and classified as appropriate were in fact disturbing (obtained by sampling 300 of the 176,619 detected appropriate videos).

5 Related Work

Prior work studied YouTube videos with inappropriate content for children, as well as spam, hate or malicious activity. Inappropriate Content for Children. Several studies focused on understanding videos that target young children, and how they interact with them and the platform. (Buzzi 2011) suggests the addition of extra parental controls on YouTube to prevent children from accessing inappropriate content. (Araújo et al. 2017) study the audience profiles and comments posted on YouTube videos in popular children-oriented channels, and conclude that children under the age of 13 use YouTube and are exposed to advertising, inappropriate content, and privacy issues. (Eickhoff and de Vries 2010) propose a binary classifier, based on video metadata, for identifying suitable YouTube videos for children. (Kaushal et al. 2016) focus on the characterization and detection of unsafe content for children and its promoters on YouTube. They propose a machine learning classifier that considers a set of video-, user-, and comment-level features for the detection of users that promote unsafe content. (Ishikawa, Bollis, and Avila 2019) studied the Elsagate phenomenon and they propose a deep learning model for detecting Elsagate content on YouTube trained on an unannotated dataset of 3K videos. (Singh et al. 2019) focus on the detection of child unsafe content. They propose an LSTM-based deep neural network for the detection of unsafe content. Deep neural network based architectures have also been proposed for large-scale video recommendation on YouTube (Covington, Adams, and Sargin 2016).

Spam, Hate and other Malicious Activity. A large body of previous work focused on the detection of malicious activity on YouTube. (Sureka et al. 2010) use social network analysis techniques to discover hate and extremist YouTube videos, as well as hidden communities in the ecosystem. (Agarwal and Sureka 2014) develop a binary classifier trained with user and video features for detecting YouTube videos that promote hate and extremism. (Giannakopoulos, Pikrakis, and Theodoridis 2010) use video, audio, and textual features to train classifier for detecting violent YouTube videos. (Ottoni et al. 2018) perform an in-depth analysis on video comments posted by alt-right channels on YouTube. They conclude that the comments of a video are a better indicator for detecting alt-right videos when compared to the video’s title. (Aggarwal, Agrawal, and Sureka 2014) use video features for detecting videos violating privacy or promoting harassment. With regard to spam detection, (Chowdury et al. 2013) explore video attributes that may enable the detection of spam videos on YouTube. (Sureka 2011) focuses on both user features and comment activity logs to propose formulas/rules that can accurately detect spamming YouTube users. Using similar features, (Bulakh, Dunn, and Gupta 2014) characterize and identify fraudulently promoted YouTube videos. (Chaudhary and Sureka 2013) use only video features, and propose a one-class classifier approach for detecting spam videos.

(Benevenuto et al. 2012) propose two supervised classification algorithms to detect spammers, promoters, and legitimate YouTube users. Also, in an effort to improve the performance of spam filtering on the platform, (Alberto, Lochter, and Almeida 2015) test numerous approaches and propose a tool, based on Naïve Bayes, that filters spam comments on YouTube. Finally, (Zannettou et al. 2018) propose a classifier for identifying videos that use manipulative techniques to increase their views (i.e., clickbait).

In contrast to all the above inappropriate content and malicious activity studies, we are the first to focus on the characterization and detection of disturbing videos, i.e., inappropriate videos that explicitly target toddlers. We collect thousands of YouTube videos and manually annotate them according to four relevant categories. We develop a deep learning classifier that can detect inappropriate videos with an accuracy of 84.3%. By classifying and analyzing these videos, we shed light on the prevalence of the problem on YouTube, and how likely it is for an inappropriate video to be served to a toddler who casually browses the platform.

6 Conclusion and Discussion

An increasing number of young children are shifting from broadcast to streaming video consumption, with YouTube providing an endless array of content tailored toward young viewers. While much of this content is age-appropriate, there is also an alarming amount of inappropriate material available. In this paper, we present the first characterization of inappropriate or disturbing videos targeted at toddlers. From a ground truth labeled dataset, we develop a deep learning classifier that achieves an accuracy of 84.3%. We leverage this classifier to perform a large-scale study of toddler-oriented content on YouTube, finding 1.05% of the 233,337 Elsagate-related videos in our dataset to be inappropriate. Even worse, we discover a 3.5% chance of a toddler who starts by watching appropriate videos to be recommended inappropriate ones within ten recommendations.

Although scientific debate (and public opinion) on the risks associated with “screen time” for young children is still on going, based on our findings, we believe a more pressing concern to be the dangers of crowd-sourced, uncurated content combined with engagement oriented, gameable recommendation systems. Considering the advent of algorithmic content creation (e.g., “deep fakes”) and the monetization opportunities on sites like YouTube, there is no reason to believe there will be an organic end to this problem. Our classifier, and the insights gained from our analysis, can be used as a starting point to gain a deeper understanding and begin mitigating this issue.

Note that in this work, we collect and analyze a large number of Elsagate-related, other child-related, as well as random and popular videos available on YouTube. Although
not representative of the entirety of YouTube, we believe that the set of seed keywords (Elsagate-related and other child-related keywords) cover a wide range of child-related content available on the platform. With regard to our sampling process, we believe that by including a wide range of child-related content as well as other types of videos in our ground truth dataset, we aid the proposed model to generalize to different types of videos that are available on YouTube.

Limitations. Finally, we discuss some of the limitations of this work. First, we collect and analyze videos only from YouTube and not YouTube Kids. This is because YouTube does not provide an open API for collecting videos that appear on YouTube Kids. However, according to YouTube, only videos marked as age-restricted are excluded from YouTube Kids unless specific settings are set by the parent. Second, we acknowledge that the performance of our classifier is highly affected by the small training size: we were unable to provide a larger annotated dataset mainly due to lack of resources for the annotation process.

Acknowledgments. This project has received funding from the European Union’s Horizon 2020 Research and Innovation program under the Marie Skłodowska-Curie ENCASE project (Grant Agreement No. 691025) and from the National Science Foundation under grant CNS-1942610.

References
Bulakh, V.; Dunn, C. W.; and Gupta, M. 2014. Identifying Fraudulently Promoted Online Videos. In WWW.
Chaudhary, V., and Sureka, A. 2013. Contextual feature based one-class classifier approach for detecting video response spam on youtube. In IEEE PST.

14https://tinyurl.com/youtube-kids-restricted

Sureka, A.; Kumaraguru, P.; Goyal, A.; and Chhabra, S. 2010. Mining Youtube to Discover Extremist Videos, Users and Hidden Communities. In AIRS.