Samba: Identifying Inappropriate Videos for Young Children on YouTube

Binh M. Le
bmle@g.skku.edu
Sungkyunkwan University
Suwon, South Korea

Rajat Tandon
rajattan@usc.edu
University of Southern California
USC Information Sciences Institute
Los Angeles, CA, USA

Chingis Oinar
chingisoinar@gmail.com
Sungkyunkwan University
Suwon, South Korea

Jeffrey Liu
Uma Durairaj
jliu5021@usc.edu
uduraira@usc.edu
University of Southern California
Los Angeles, CA, USA

Jiani Guo
Spencer Zahabizadeh
jennyguo@usc.edu
szahabiz@usc.edu
University of Southern California
Los Angeles, CA, USA

Sanjana Ilango
Jeremy Tang
ilango@usc.edu
tangjere@usc.edu
University of Southern California
Los Angeles, CA, USA

Fred Morstatter
fredmors@isi.edu
University of Southern California
USC Information Sciences Institute
Los Angeles, CA, USA

Simon S. Woo
swoo@g.skku.edu
Sungkyunkwan University
Suwon, South Korea

Jelena Mirkovic
mirkovic@isi.edu
University of Southern California
USC Information Sciences Institute
Los Angeles, CA, USA

ABSTRACT

YouTube videos are one of the most effective platforms for disseminating creative material and ideas, and they appeal to a diverse audience. Along with adults and older children, young children are avid consumers of YouTube materials. Children often lack means to evaluate if a given content is appropriate for their age, and parents have very limited options to enforce content restrictions on YouTube. Young children can thus become exposed to inappropriate content, such as violent, scary or disturbing videos on YouTube. Previous studies demonstrated that YouTube videos can be classified into appropriate or inappropriate for young viewers using video metadata, such as video thumbnails, title, comments, etc. Metadata-based approaches achieve high accuracy, but still have significant misclassifications, due to the reliability of input features. In this paper, we propose a fusion model, called Samba, which uses both metadata and video subtitles for content classification. Using subtitles in the model helps better infer the true nature of a video improving classification accuracy. On a large-scale, comprehensive dataset of 70K videos, we show that Samba achieves 95% accuracy, outperforming other state-of-the-art classifiers by at least 7%. We also publicly release our dataset.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557442

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**; • **Computing methodologies → Machine learning**.

KEYWORDS

Inappropriate Video Detection, Video Dataset, Young Children, YouTube Video

ACM Reference Format:
Binh M. Le, Rajat Tandon, Chingis Oinar, Jeffrey Liu, Uma Durairaj, Jiani Guo, Spencer Zahabizadeh, Sanjana Ilango, Jeremy Tang, Fred Morstatter, Simon S. Woo, and Jelena Mirkovic. 2022. Samba: Identifying Inappropriate Videos for Young Children on YouTube. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557442

1 INTRODUCTION

The emergence of social platforms with thousands of videos published every second, such as YouTube, TikTok, Facebook Watch, etc. has attracted users worldwide. Among them, YouTube is the most popular platform with an estimated 2.1 billion users, and more than a billion hours of content uploaded worldwide as of September 2021 [12].

When uploading a video, publishers are asked to self-categorize their content, e.g., whether it is appropriate for children. This is both a broad criteria (e.g., it could be appropriate for teenagers, but not for preschoolers) and a subjective one (publishers can lie). To make matters worse, much interaction with YouTube content happens through browsing of related or recommended videos, not through direct search. Users have no explicit control over content that will be offered to them by algorithms, and are thus exposed

to very diverse and potentially inappropriate or unwanted content, such as videos of real violence (e.g., murder, rape), fake information, discriminatory speech, emotional triggers for special audiences (e.g., suicide videos). While inappropriate content is a big problem for every user population, in this paper we focus on one specific group – children 5 years and younger. Such children are too young to make their own decisions about which content is appropriate for them, most of them cannot yet read, and are also too young to consistently follow parental rules. Thus, automated solutions are needed to identify and filter inappropriate content for such young audiences.

Platforms themselves (e.g. YouTube) may have some proprietary algorithms to classify video content for various audiences, but this is not publicized and we could find no information that such algorithms exist. Even as of May 2022, YouTube Kids shows videos promoting drug culture and firearms to toddlers [42, 44]. We were also able to discover hundreds of videos on YouTube Kids that are inappropriate for young children to watch, within a few minutes [15].

At TikTok, a feature called "Family Pairing" allows parents to link their child's account to their own where they can control direct messages, set screen time limits, and turn on/off restricted content directly from their phone. However, inappropriate material is not flagged properly [27]. With a quick hashtag search, one's child can access mature content. Because of the increase and promotion of inappropriate content, TikTok has been banned in several countries.

There is a wide range of studies on detecting videos inappropriate for young audiences [1, 10, 34]. However, all of them classify videos based on their metadata, such as comments and thumbnails, title and description, etc. Video metadata may not be very well aligned with the actual content of the video, hence it can lead to a low classification accuracy. Prior work only achieves classification accuracy of 60 − 84% on their datasets, and 88% accuracy on our dataset.

In this paper, we develop a novel content-based classifier that can efficiently distinguish inappropriate videos from appropriate ones, using both metadata and video subtitles. Use of subtitles along with metadata improves classification accuracy compared to metadata-only classifiers from 88% to 95%. Our contributions are summarized as follows:

A novel classification model for inappropriate videos for young children. We propose a novel machine learning model, Samba, which includes two stages of training and embeds videos' subtitles and metadata, as shown in Figure 1. The first stage trains a word embedding model to represent semantically meaningful segments of subtitles and to accommodate subtitles of variable length. The second stage trains an end-to-end detection model with our proposed recurrent fusion module, including the output of the first stage and metadata encodings. The recurrent fusion module that we propose can selectively aggregate discriminative features across different input variants. Our model can be further generalized for other types of input (e.g., visual or audio). We publicly release our code[1].

We evaluate Samba and several competing models on a manually collected, large and diverse dataset of appropriate and inappropriate videos for young children. We demonstrate that Samba outperforms metadata-based approaches by a large margin (of 7%), achieving 95% classification accuracy. We also show that other ML-based classification approaches (e.g., decision trees, random forests) that use subtitles, metadata or both achieve much lower accuracy than Samba.

Collection and release of a comprehensive, labeled dataset of YouTube videos that are appropriate and inappropriate for young children. We created this dataset systematically, by identifying various content categories that may be appropriate or inappropriate, based on publicly available sources for content categorization. We then sampled YouTube channels in these categories, labeled them manually and mined their content for our dataset. We ended up with 142 K videos, out of which we randomly sampled a balanced evaluation dataset of 70 K videos, with equal proportion of appropriate and inappropriate videos. We publicly release our dataset, so others can reproduce our results and further investigate video classification[2].

Our work brings focus on social and ethical issues around video hosting platforms. Our work raises important social and ethical concerns around indiscriminately sharing content with wide audiences, and the need for developing technical solutions to prevent such problems. Young viewers are not the only ones at risk. Any vulnerable population (e.g., sexual violence survivors, people at risk for suicide, anorexia survivors, etc.) can be affected by content targeting their vulnerability. While hosting platforms could handle these concerns, their business model may not prioritize this. Our proposed approach can help develop client-based solutions to filter inappropriate content of any kind, given good-quality datasets for training.

Analysis and Recommendations. We discuss how our model can be used to improve safety of young audiences on YouTube, and chart some directions for future work, such as: (1) automatically identifying and tagging targeted demographics, kids vs. adults, during video uploads, (2) restricting videos on YouTube as a default setting and (3) extending our work to other platforms, or AI assistant devices, such as Alexa.

2 PROBLEM

YouTube is a video sharing service where users can watch, like, share, comment and upload their own videos. The videos also include additional information generated by their publishers, consumers and YouTube, known as metadata. Each YouTube video includes the following metadata:

- **Tags** are keywords that publishers can select for their own videos. The tags can help users find that content easily.
- **Title** of YouTube video, a required field on the platform.
- **Thumbnails** let viewers view a quick snapshot of YouTube videos. When a video gets uploaded, the publisher can choose a thumbnail from the three options that YouTube automatically generates using the video content, or upload another representative image.
- **Likes/ Dislikes** let a video creator know how viewers liked the content.
- **Views** represent the number of times a video is watched.

[1] https://github.com/leminhbinh0209/samba

[2] https://sites.google.com/view/samba-kids/

- **Comments** are posted by YouTube users, and visible publicly.
- **Duration** is the playing length of a YouTube video.
- **Video description** is the text shown below a video. It helps viewers find content and decide whether to watch a video or not. The publisher supplies this information.

The publishers can upload YouTube videos that may be targeted to specific demographics or audiences. However, this rating is subjective and can be manipulated by the publisher. Thus, many viewers get exposed to content that is not suitable for them to watch such as smoking fetish videos [28], videos about teenage suicide [11], body shaming [22], underage access to alcohol [3], etc. Young children is one audience being exposed to inappropriate content [4, 19, 24]. In 2015, Google released a children-oriented platform called YouTube Kids, which should allegedly host only child-appropriate videos, and parents can select a broad age-based category (e.g., preschooler) for their kid. Nevertheless, YouTube Kids still displays inappropriate content in many cases. A recent study from *Common Sense Media* on YouTube Kids [9] found that 27% of videos that were watched by kids 8 and under were intended for older target audiences, with violence being the most likely negative content type. Tech Transparency Project [42] found videos on YouTube Kids that talk positively about cocaine and crystal meth, give instructions on concealing a gun, encourage skin bleaching, and introduce diet culture to children.

Other platforms such as TikTok, and Facebook, are unsafe for kids too [43, 47]. Therefore, intelligent content filtering solutions are needed for users to prevent inappropriate videos being recommended to them or their children in the future. Such solutions could be deployed at client side, and customized to their viewing preferences.

3 RELATED WORK

Much research has focused on detecting inappropriate content on YouTube by analyzing the video metadata or video material. Carsten Eickhof and Arjen DeVries [16] investigate the metadata of YouTube videos including 18 features, such as number of views and likes, and introduce a binary classifier to distinguish suitable videos for children. Kaushal *et al.* [26] propose a machine learning classifier that considers three levels of input data types, including video, user, and comment-level features, for detecting the users that intentionally promote disturbing videos.

By inspecting profiles and comments of audience in popular children-oriented channels on YouTube, Araújo *et al.* [2] come to the conclusion that children under the age of 13 are easily exposed to unsuitable information and advertising. Using videos' frames as the input, authors in [37] propose a recurrent deep neural network for identifying unsafe content videos. By exploring the Elsagate phenomenon in the literature, Ishikawa *et al.* [23] develop a static- and motion-based deep learning model for detecting disturbing cartoon videos. To identify inappropriate videos, Tahir *et al.* [40] use audio and visual elements, video frames, embedded audio, and character motions. Papadamou et al. build a classifier, using metadata, to identify inappropriate content that targets toddlers on YouTube [34]. Unlike previous works, we not only rely on metadata as features,

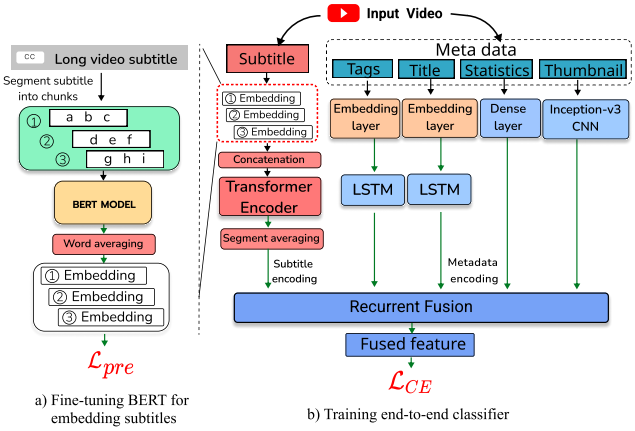


Figure 1: Overview of our two-stage training classification approach. In the former stage (a), BERT model is trained for enriching subtitle's representation vectors. The $\mathcal{L}_{pre}$ can be either contrastive loss or binary cross-entropy loss. In the classification training stage, (b) We apply Gated Recurrent Unit module for aggregating features from different input types of one video. Green arrows indicate flows of gradient.

but also utilize subtitles from YouTube, which helps us achieve better accuracy.

4 SAMBA: USING SUBTITLES AND METADATA FOR CLASSIFICATION

Prior research [34] achieves solid classification results for metadata-based classifier, 84% accuracy on their dataset and 87% accuracy on ours. But metadata can be easily manipulated by publishers. Publishers may want to misrepresent their content as suitable for kids to enlarge their audience, since larger audience leads to larger monetization. Some publishers may also be malicious and intentionally target children with inappropriate content [19]. Thumbnails, tags, and titles can be modified deceptively, comments can be locked, deleted or manipulated by paying users to leave specific, positive comments. Thus metadata-only classification is not sufficient to reliably identify inappropriate content.

Our work focuses on use of subtitles, in addition to metadata, for video classification. While there are some videos that only contain music, the majority of videos on YouTube has some spoken language, which gives the video its meaning. Thus, if publishers wanted to manipulate the spoken language in a video to misguide our classification, they would have to sacrifice the meaning of the videos, and thus lose a portion of their main audience. For example, videos showcasing horror games could modify the speech to be classified as appropriate for younger audiences, but they would then not appeal to teenagers and young adults, which is their main audience.

One approach to video classification could use only subtitles and not metadata. In Section 6, we show that such approach achieves substantial improvements over metadata-only classification, but still does worse than a combined subtitle-metadata approach, which we propose.

4.1 Samba Architecture

Figure 1 shows the architecture diagram of our proposed approach, Samba, including two stages of training. First, we fine-tune BERT [13] on segmented fixed-length chunks of a given subtitle. This is done to address input length limitation issues faced with very long texts. When chunk embeddings are produced by our fine-tuned BERT we feed all embeddings into Transformer Encoder [45]. We then average all the chunk encodings to obtain one fixed-length encoding for a given subtitle.

In addition to subtitles, our proposed architecture also ingests embeddings of meta-data, namely tags, title, statistics (duration, the number of views, likes, dislikes, and comments) and thumbnail. We use approach from Papadamou et al. [34] to embed metadata. Finally a feature fusion is performed to combine the embeddings into a single embedding responsible for the classification. We use GRU [8] for feature fusion, so our model can selectively learn which features match or mismatch to promote or ignore.

4.2 Subtitle Embedding

We aim to produce a representation of video subtitles as a single vector. Ideally, our representation would capture not just presence of certain words in subtitles, but their semantic relationship with the rest of the subtitle text. To capture this relationship we use BERT [13] to fine-tune it on our subtitle data. We use the English-language $\texttt{BERT}_{\texttt{base}}$ uncased model pre-trained on extremely large corpora for years, and focus on videos with English subtitles.

Due to input-text length restrictions with using BERT, and due to requirements of many machine-learning approaches for fixed-length inputs, we want to transform subtitles into a fixed-length representation. We split subtitles into smaller segments and fine-tune the pre-trained BERT model to extract useful embeddings for each segment.

Specifically, we average all the word embeddings in each segment applying the technique from [36]. This means that we pool segment embeddings by computing the mean of all output word embeddings, which does not include CLS-token. This gives us one embedding per segment, which has a dimension of D. D is a customizable parameter. We use 768 as a default from BERT model.

Unsupervised contrastive learning of segment embeddings. We hypothesize that two subsequent segments within a given subtitle usually convey a similar context. Thus, we use contrastive learning on BERT segment embeddings to help our model learn similarities between subsequent segments in the same subtitle. This kind of learning needs at least two segments from a given subtitle. We split our subtitles into 100-word segments. If a segment is too short we pad it with PAD-token. However, If a segment is shorter than 60 words we will drop it instead of pad it, because it may lack sufficient information. We drop subtitles that are too short (less than 160 words)

During contrastive learning, we aim to minimize the distance between embeddings of two consecutive segments of a given subtitle, while maximizing the distance between segments from different subtitles or between non-consecutive segments within the same subtitle. For this particular task, we adopt Momentum Contrast (MoCo) for unsupervised representation learning [21]. Unlike traditional contrastive learning, MoCo uses two encoders, an encoder and a momentum encoder, where the latter is updated in step with the main encoder. Additionally, MoCo introduces a dynamic dictionary, which is updated with latent vectors obtained from the momentum encoder from the current mini-batch every iteration. A mini-batch contains encodings of segments from multiple videos.

Subtitle embeddings during end-to-end classifier training. We concatenate each segment embedding, produced by our fine-tuned BERT, for a given subtitle to obtain $N \times D$ matrix, where N is the number of segments and D is embedding dimension respectively. Since subtitles may have different number of segments, we pad the subtitles with zero vectors to match the length of the longest subtitle within a mini-batch. Finally, the input is passed into a simple Transformer Encoder architecture [45]. Transformer encoder outputs one encoding for each segment embedding. We average these encodings to obtain the final encoding for a given subtitle. We do not include padded vectors while averaging.

We have tried other metric learning methods (see Section 6.3), but the embeddings obtained using MoCo give us superior performance.

4.3 Metadata embedding

We follow the same approach as [34]'s to pre-process the metadata. In particular, each video's thumbnail photo is embedded by a pre-trained Inception-v3 [39] to form a representation vector of 2048 dimensions. For the video's title, it is encoded to one-hot vector, using the vocabulary of all titles in the training set. The videos' tags are handled in the same manner as the titles. The remaining information about a video (i.e., duration, video category, the number of views, likes, dislikes, and comments) is all embedded into one vector, with each value taking a separate sub-vector. The title and tags of a video are embedded into an N-dimensional vector space using a long-short term memory (LSTM) network. Statistics features and thumbnail image's representation are fed forward through multi-layer perceptrons to produce similar-size embedding vectors.

4.4 Feature Fusion and Downstream Classifier

In this stage we combine subtitle embedding with metadata embedding into an end-to-end training model. The model learns how different features and their combinations influence the final classification (appropriate/inappropriate) using a recurrent neural network (RNN) in recurrent fusion step.

Recurrent fusion. To combine subtitle and metadata embeddings we find inspiration in the work of Kar et al. [25], which uses the Gated Recurrent Unit (GRU) [7, 8] to fuse multiple unordered features of one object under different views to form one representation feature. We propose to use GRU to aggregate embedding features from different video's input features and build up our end-to-end training model, dubbed Samba (***S**ubtitle **a**nd **m**eta-data **b**ased ensemble **a**rchitecture*). Intuitively, the GRU's hidden states selectively update the discriminative information by running through all the input features. Therefore, GRU module can learn which features it should pay attention to and when to ignore mismatching features that might affect the model's prediction. In our experiment, we empirically observe that the order of feature sequence has a relatively low effect on our model's performance.

The trained recurrent fusion model becomes our classifier. In evaluation, we freeze the transfomer encoder and recurrent fusion

model, encode subtitles and metadata in the same way as during training and use the outputs of recurrent fusion model as classification outputs.

5 DATASET

While Papadamou et al. [34] released a dataset of 4,797 videos that they used in their research, we found that this dataset was too small to evaluate our model. Deep learning-based classifiers, such as BERT, achieve better classification accuracy when trained on large datasets, including hundreds of thousands of samples [6]. Also, 30% of videos, which were available at the time Papadimou et al. [34] performed their research, have since been removed from YouTube. Among the 70% videos that remain available, only 30% have subtitles and hence, are not sufficient to do a large scale evaluation of our approach. We thus decided to collect our own dataset, and release it publicly at [41].

5.1 Channel Selection

We needed hundreds of thousands of videos, along with their ground truth labels (appropriate or inappropriate) for our evaluation. It would have been difficult to manually classify each video in such a large dataset. Instead, we decided to identify YouTube channels that may host a certain kind of appropriate or inappropriate content for young audiences, classify the channel manually and apply the same ground-truth label to each video on the channel.

To acquire a diverse set of appropriate and inapproppriate content, we used the definitions of what is appropriate from YouTube guidelines [49] and FTC's Children's Online Privacy Protection Act (COPPA) [17]. This led us to define the following kinds of channels that may be inappropriate: classic cartoons edited to have inappropriate text or visuals, gaming content for adults, adult cartoons, toy destruction videos, deceptive channels targeting children, family channels demonstrating child abuse, frightening or violent videos, age-inappropriate how-to videos, television or movie scenes with adult content, supernatural phenomena, pranks, and music videos with adult content. In addition to these categories, we also consider videos that contain irrelevant content for young children as inappropriate, such as news videos, late night shows, etc. Our dataset includes 80 inappropriate channels, some of which are shown in Table 1 as examples.

We also identified categories of content that can be considered appropriate for children. These include: nursery rhymes, video game plays without any inappropriate content, kids' toy demonstrations, toy ratings, children's music or dance performances, unedited cartoons and animations for young children, educational kids' videos, and animal videos. These categories are selected based on the use of animated characters, the age of characters or models, child-oriented activities and incentives, and simple language or content appropriate for a general audience. Our dataset includes 80 appropriate channels, and some example channels are provided in Table 1. To identify specific channels we included in our dataset, five paper authors individually looked for channels in each of our appropriate or inappropriate categories, using simple manual YouTube searches. Each person watched roughly 5 to 10 videos of a selected channel to ensure that they all fit in a given category.

Table 1: Categories and examples of appropriate and inappropriate channels that are annotated by four of our co-authors.

Category	Sample channel
Appropriate	
Nursery rhymes & Kids Songs	Little Baby Bum - Nursery Rhymes
Classic cartoons / animations	Peppa Pig Toy Videos
Educational videos	Happy Learning English
Kids toys	Genevieve's Playhouse - Learning Videos for Kids
Simple gaming videos	Games-BnB
Animal videos	Anaimal Videos
Kids Bop	The Kiboomers - Kids Music Channel
Kids show	Kids Roma Show
Inappropriate	
Gaming content for adults	Super
Frightening / Violent Situations	MindSeed TV
Edited classic cartoons	Leo Koutakis
Adult cartoons	Happy Tree Friends HD
Toy Destruction	SovietYurii
Supernatural phenomena	ParanormalCollection
Pranks	Ownage Pranks
Scary Music Videos	Top Music Video
"Family Channels" with Kids Being Hurt / Abused	The ACE Family
Inappropriate Movie / TV Scenes	Binge Society - The Greatest Movie Scenes
Advertisements intended for adults	UkraineArsenal
Irrelevant	Global News

Table 2: Number of videos with and without subtitle per class.

Statistic	*Appropriate*	*Inappropriate*	Total
# Channels	80	80	160
# Videos	61K	146K	207K
# Videos with subtitles	35K	107K	142K
# Videos without subtitles	26K	39K	65K

5.2 Labeling Appropriate and Inappropriate Content

To label the data, we manually review each channel by inspecting their channel content, video content, channel titles, video titles, thumbnails and tags in YouTube. Each channel is presented to four annotators who inspect the channel content and assign either of the following labels. **Appropriate.** A video is labeled as appropriate when its content is appropriate for toddlers and preschoolers (aged 1-5 years) and it is relevant to their typical interests. **Inappropriate.** A video is labeled as inappropriate when it contains inappropriate visual content, language or both, or it contains content irrelevant/uninteresting to young viewers. Snippets of a few inappropriate videos are provided in Figure 2.

Four annotators (authors of this paper), independently label each channel that they did not originally contribute to the dataset. Statistics about the number of channels that were labeled, and videos with and without subtitle per class are shown in Table 2. Discrepancies were resolved by the author who recommended the channel as fifth vote.

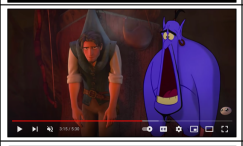
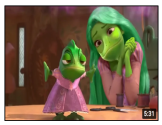
Eg.	Snippet	Title	Subtitles	Tags	Thumbnail	Likes
1		Mokey's show - 9/11	the house?!?!? the house!!!! ooh noo!!! bitch!! dying	Mokey, dilan, groovy, mickey, donald, sr pelo, September 11 Attacks (Event)		372K
2		3 TRUE GRINCH HORROR STORIES ANIMATED	so he'd bleed through the snow he was buried in the toboggan he was offering us	animated horror stories, horror stories, animated, horror, stories, mcdonald's horror story		9.5K
3		Tangled Craziness !	don't they see if you watch that video you die yeah that's a lot of baloney	Animation, Craziness, Leo, Koutakis, cartoon, disney, big, hero, parody, ytp, crack, kids, comedy		86K
4		Peek-A-Boo Clown | Spirit Halloween 2020	i just love that game particularly with crying little babies i'm ready to play again	video, sharing, camera phone, video phone, free, upload		85

Figure 2: Example snippets, subtitles and some metadata information of inappropriate videos.

Table 3: Number of videos in each category of training and test set in our dataset.

Dataset	# *Appropriate* Videos	# *Inappropriate* Videos	# Total Videos
Training	24.6K	24.6K	49.2K
Test	10.4K	10.4K	20.8K

Inter-annotator agreement. We compute the agreement across the raters using Bennett et al.'s *S score* [20, 46]. The S score value that we get is 0.85, which indicates a strong agreement between raters. While Cohen's Kappa [30] measures the overall agreement between two raters, Bennett et al.'s *S score* is one of the common techniques used for calculating inter-annotator agreement for more than two raters, as in our case. It accommodates the percentage of rater agreement that might be expected by chance, instead of just the simple agreement between raters, as with Cohen's Kappa [30].

Data crawling. Each YouTube video has one unique identification code (ID). We use this ID to download metadata for processing on Google Cloud. We use Google API`googleapiclient` library in Python to send a request with a video's ID to Google Cloud to download the videos, and then we extract the video's metadata: channel titles, video titles, thumbnail photos, tags, duration, video category, the number of views, likes, dislikes, and comments (statistics). Separately, we also obtain the video's subtitles using YouTube APIs.

Ethics. We collect only the data that is publicly available on the Web and do not (1) interact with online users, nor (2) imitate any logged-in activity on YouTube or other platforms. Therefore, the IRB approval was not required for this work.

5.3 Balanced Subset and Metadata

Using the dataset that the annotators labeled (see statistics in Table 2), we create a well-balanced subset for training and evaluation purposes, shown in Table 3. The subset is created by randomly selecting a large number of videos with subtitles, with an equal number of appropriate and inappropriate videos. To avoid any bias, we ensure that not only the videos should be different between the training and testing set (Table 3), but the channels should also not be repeated between the training and testing set. We present some meaningful statistics about the metadata of the balanced subset that we use for training and testing as follows.

Observation 1: Appropriate videos have more views than inappropriate videos. Figure 3a shows the logarithm of number of views for the appropriate and the inappropriate videos of the balanced subset, with appropriate videos having significantly more views.

Observation 2: Appropriate videos have more likes than inappropriate videos. Figure 3b shows the logarithm of number of likes for the appropriate and the inappropriate videos in the balanced subset, with appropriate videos having many more likes.

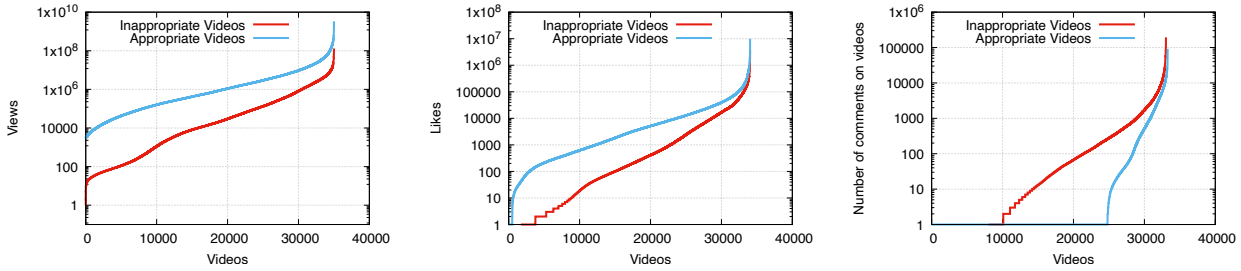
Observation 3: Inappropriate videos receive more comments than appropriate videos. Figure 3c shows the logarithm of number of comments for the appropriate and the inappropriate videos in the balanced subset, with inappropriate videos receiving many more comments.

6 EVALUATION

We evaluate Samba along with several competing approaches using our balanced subset (Section 5.2). It contains 70 K videos (50% appropriate, 50% inappropriate) and we use 49 K for training (70%) and 21 K for testing (30%).

6.1 Evaluation Settings

We evaluate Samba on our dataset, and compare its performance with several competing approaches. To highlight the benefits of Samba over classical machine learning approaches we evaluate:



(a) Views (b) Likes (c) Comments

Figure 3: Views, likes, and number of comments on appropriate and inappropriate videos of the balanced dataset that we use for training and testing.

Naïve Bayes, K-nearest neighbor, decision tree, support vector machines (SVM), and random forest. We apply these approaches separately to metadata and to subtitles, and then to their combination, to evaluate both the ability of unsophisticated models to detect inappropriate content, and how much each feature type contributes to classification accuracy. When applying classifiers to subtitle input, we use `TweetTokenizer` module from `nltk` [5] library to tokenize subtitles. Next, we employ `word2vec` module from `gensim` library [35] to embed the tokens into N-dimensional vectors. Classifiers are trained and evaluated using these vectors.

Metadata-based models. In addition to baselines, we also compare to closest related work that uses metadata for classification – [34]. We reimplement their approach and train and evaluate it on our dataset.

Subtitle-based models. To evaluate how well a subtitle-only model performs (without metadata), we evaluate BERT [14] and three recent improvements – RoBERTa [32], XLNet [48] and SimCSE-BERT [18]. We train these models on fixed-size segments of subtitles, but do not combine the final embeddings. Instead, in evaluation, we classify each segment of the subtitles separately. We then assign to the entire video a majority label (> 50%) of its segments.

6.1.1 Pre-training stage. In our pre-training stage for Samba, we adopt `AdamW` [33] as our optimizer and train the model for 3 epochs with the learning rate of $2e^{-5}$. For the contrastive approach, the queue size is set to 512, whereas the momentum and temperature parameters are set to 0.9 and 0.07, respectively.

6.1.2 Classification stage. The settings of all classifiers are as follows:

- Naïve Bayes. We use Bernoulli Naïve Bayes with smoothing parameter of 1.0.
- K-nearest neighbor. The number of neighbors and leaf size are set to 8 and 10, respectively.
- Decision Tree Classifier. We employ entropy as a function to measure the quality of a split.
- Support Vector Machine (SVM). We use the linear kernel and set the regularization parameter to 1, and size of the kernel cache to 200.
- Random Forest Classifier. The number of trees in the forest is set to 100, and we use entropy for measuring the information gain.
- Model proposed by Papadamou *et al.*[34]. We follow recommendations by the authors. At the penultimate layer, we concatenate all embedded vectors horizontally to form a general representation vector of one video. Finally, it is forwarded to an output layer and the whole model is then trained by `Adam` [29] optimizer in 64 epochs.
- Transformer-based models for subtitles data. First, we partition subtitles into small chunks of length 100 each. Next, we fine-tune pre-trained BERT [14, 38], RoBERTa [32], XLNet [48], and SimCSE-BERT [18], respectively, to embed sentences and optimized them with binary cross entropy loss at the output. In the inference stage, we perform conservative voting from all chunks for one subtitles and deciding the final class. BERT and XLNet achieve their best performance at the threshold of 0.6, while it of SimCSE and RoBERTa is 0.5.
- For our proposed Samba model, we use the learning rate of $1e^{-3}$, excepting $1e^{-4}$ for the subtitle model. The best model is obtained upon the best validation accuracy at every epoch.

All models in our experiments are trained on a single GeForce RTX 3090 24GB GPU with Intel Xeon Gold 6230R CPU @ 2.10GHz, and their hyper-parameters are empirically fine-tuned by grid search.

6.2 Comparison with Baselines

Table 4 summarizes the performance of all methods for accuracy (percentage of accurate classifications), precision (percentage of inappropriate classifications that are correct), recall (percentage of inappropriate videos that are correctly classified) and F1 score (geometric mean of precision and recall).

Baseline methods achieve accuracy of 55–82% on metadata, 56–79% on subtitles, and 62–79% on both metadata and subtitles. These classifiers rely only on immediate features and their combinations, and do not mine deeper relationships between metadata or subtitle segments, which leads to their low accuracy. Subtitle-based classification performs comparable to metadata-based classification with each baseline classifier. BERT and its improvements do better on subtitle-only datasets, when compared to classic machine learning models, and achieve 85–88% accuracy. Papadamou et al. [34]'s model achieves the best performance out of competing approaches, using only metadata, with an accuracy of 88%. It outperforms other metadata-only approaches by at least 9%. Samba outperforms all metadata-based and subtitle-based approaches. Our proposed

method Samba with contrastive learning achieves superior accuracy, precision and recall of 95%, 94% and 96%, respectively. This is because our model captures the context and semantic meaning of an entire subtitle by learning relationships between pairs of segments via contrastive learning, and by learning how metadata and subtitle embedding relate to the final classification of the video with GRU-based fusion module.

If the training set and testing set include the same channels but different videos, Samba achieves superior accuracy (up to 0.99). However, we present a more realistic and conservative evaluation design in the paper, because in real deployment the classifier cannot predict which channels a user will watch.

6.3 Ablation Study

Next, we describe our ablation study related to the alternatives of GRU-based feature aggregation, subtitles pre-training methods, and our experiment regarding videos without subtitles.

6.3.1 Aggregation methods. Given embedding vectors of *five* input data types, *i.e.,* thumbnail photo, headline, statistics, tags, and subtitles, we have tested several approaches for aggregating them into one representation vector for a video. We evaluate three common approaches – averaging, concatenating, and attention method (Lin et al. [31]) – against our approach using GRU. Classification results are shown in Table 5. Averaging does slightly better (accuracy 88%) than concatenation or attention method (accuracies 84% and 82% respectively), but GRU outperforms them all by at least 7% (accuracy 95%).

6.3.2 Pre-training methods of subtitles embedding. Next, we also provide a study on subtitle embeddings in Table 6. We evaluate and compare various pre-training methods with the method we use in Samba, which is constrastive learning [21].

Binary pre-training. We customize the BERT model to output appropriate/inappropriate labels for each segment. Specifically, given the embedding vector output from BERT, we perform a simple classification, by passing it through another feed forward Artificial Neural Network (ANN) layer, and train our network with cross-entropy (CE) Loss. Unlike the unsupervised approach, this approach promotes segment embeddings to correspond to one of the target classes (appropriate or inappropriate). **SimCSE**. Gao *et al.* introduced SimCSE [18] as a simple contrastive learning framework that advances the state-of-the-art sentence embeddings. We include SimCSE pre-trained on a combination of MNLI and SNLI datasets, provided by the authors, along with SimCSE pre-trained on our dataset directly, denoted with asterisk.

Results are shown in Table 6. Contrastive learning methods outperform the supervised binary pre-training by at least 4% in accuracy. The method we use in Samba, MoCo [21], achieves the best performance over all four metrics.

6.3.3 Videos without subtitles. Some videos on YouTube have no subtitles, which means that we would have to rely only on metadata for classification. We simulate this situation by masking the all subtitles' embedding by zeros and validate our model on the test set. Samba achieves 93% accuracy, which is still quite high compared to [34], even though they both only leverage metadata in this experiment. Samba's superior performance occurs because of our proposed recurrent fusion module combines feature vectors in a more sophisticated manner (using GRU [8]) than Papadamou's approach of concatenating all feature vectors.

7 RECOMMENDATIONS

In this section, we discuss how our work could be used to improve experience of YouTube consumers.

Automatically identifying targeted demographics, kids vs. adults, during video upload: YouTube can use classifiers to automatically identify the targeted demographics during video uploads, instead of allowing publishers to self-label content categories. This will aid in applying automatic filters for kids during video searches and recommendations, e.g., on YouTube Kids.

Default settings for restricted mode on YouTube: By default, YouTube's restricted mode is off. Turning this setting on by default will enforce a safer YouTube experience.

Extending our work to other platforms: Inappropriate content for kids is pervasive at other video-sharing platforms too. Even more broadly, other online platforms, such as Facebook, Instagram, Google search, which serve a mixture of textual, graphical and video content should consider how to categorize this multimodal content and enforce protections for young audiences. Our approach could help here, because we already integrate subtitles and metadata in our model, and could extend this to other types of content. Online content classification is challenging not only because content categorization is difficult, but also because a client device used to access content can be shared between users of different age. Future work is thus needed on both content classification and on ongoing user identification.

Identifying inappropriate content for different audiences. While our current paper focuses on identifying appropriate content for young audiences, similar classifiers could be developed for any vulnerable audience. For example, domestic violence survivors could benefit from filters that identify and block violent content and teenagers could benefit from filters that identify and block content that promotes violence, drug and alcohol use, unsafe sex, suicide, bullying, cyberbullying and eating disorders. Such classifiers could be developed in the same manner as our classifier, one would just need to collect sufficiently a large, labeled dataset.

8 LIMITATIONS AND FUTURE WORKS

Our work focuses on using metadata and subtitles to classify videos as appropriate or inappropriate. Some videos may have no subtitles, which may lower our classification accuracy. In Section 6, we show that Samba's accuracy decreases when subtitles are not present, but remains higher than accuracy of other, competing approaches.

Our dataset and our study only uses videos with English subtitles, which is a limitation. Further study and a larger, more diverse dataset are needed to evaluate if same accuracy trends apply to videos with subtitles in other languages.

In some cases, inappropriate content in a video may be due to visual representations (e.g., blood/gore superimposed on regular cartoons) or non-speech audio, which may not be reflected in subtitles or metadata. In those cases, visual and audio information would need to be mined from videos and included in classification. Our Samba offers a promising aggregation approach, and we plan to

Table 4: Performance of different methods using different input data. Samba outperforms all other approaches.

Method	Data		Metrics			
	Subtitles	Metadata	Accuracy	Precision	Recall	F1 score
Naïve Bayes	✓		0.59	0.56	0.91	0.69
		✓	0.60	0.62	0.60	0.59
	✓	✓	0.62	0.66	0.52	0.58
K-Nearest	✓		0.74	0.70	0.85	0.77
		✓	0.66	0.66	0.66	0.66
	✓	✓	0.65	0.69	0.55	0.61
Decision Tree	✓		0.70	0.70	0.69	0.70
		✓	0.82	0.82	0.82	0.82
	✓	✓	0.79	0.75	0.86	0.80
SVM	✓		0.56	0.58	0.43	0.49
		✓	0.68	0.69	0.68	0.68
	✓	✓	0.66	0.69	0.59	0.64
Random Forest	✓		0.79	0.78	0.80	0.79
		✓	0.55	0.56	0.55	0.55
	✓	✓	0.62	0.59	0.76	0.66
BERT [14, 38]	✓		0.86	0.85	0.86	0.86
RoBERTa [32]	✓		0.85	0.82	0.89	0.86
XLNet [48] (NeurIPS'19)	✓		0.85	0.80	0.93	0.86
SimCSE-BERT [18] (EMNLP'21)	✓		0.87	0.86	0.88	0.87
Papadamou *et al.* [34] (AAAI'20)		✓	0.88	0.86	0.92	0.89
Samba ***(ours)***	✓	✓	**0.95**	**0.94**	**0.96**	**0.95**

Table 5: Ablation study on aggregation methods for representing one video's features.

Method	Accuracy	Precision	Recall	F1-Score
Averaging	0.88	0.89	0.88	0.88
Concatenating	0.84	0.84	0.84	0.84
Attention	0.82	0.82	0.82	0.81
GRU	**0.95**	**0.94**	**0.96**	**0.95**

Table 6: Ablation study on Samba with various pre-training methods of sentence embeddings. We also trained SimCSE on our dataset, indicated by (*).

Method	Accuracy	Precision	Recall	F1-Score
Supervised	0.88	0.89	0.88	0.88
SimCSE [18] (EMNLP'21)	0.93	0.93	0.93	0.93
SimCSE*	0.92	0.92	0.92	0.92
MoCo	**0.95**	**0.94**	**0.96**	**0.95**

extend it with additional input types, such as visual features from a video.

Our evaluation uses a balanced dataset, while in real deployment the data would likely be imbalanced, with many more appropriate than inappropriate videos. Thus even a small false positive rate may lead to many videos being blocked. We believe this cost is acceptable, given the benefit of protecting children from inappropriate content. If Sambawere deployed at platform side, publishers could always report inaccurate classifications to the platform, which could trigger retraining and improve classification accuracy.

9 CONCLUSION

Rapid growth of popular video-based entertainment platforms, which allow anybody to freely broadcast their creative works, has resulted in a large number of videos that are inappropriate for different audiences. In this paper, we developed a large-scale, comprehensive dataset that contains both metadata and subtitle data characterized for appropriate and inappropriate videos for young children. Furthermore, we proposed a novel machine learning model that can effectively detect inappropriate videos by aggregating their subtitle representation and metadata features. Adding subtitles to metadata improves classification accuracy. Our model is also robust in cases when subtitle data is missing or is present in a different language (and thus effectively not part of our model).

ACKNOWLEDGMENTS

This work was partially supported by the Basic Science Research Program through National Research Foundation of Korea (NRF) grant funded by the Korean Ministry of Science and ICT (MSIT) under No. 2020R1C1C1006004 and Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean MSIT: (No. 2022-0-01199, Graduate School of Convergence Security at Sungkyunkwan University), (No. 2022-0-01045, Self-directed Multi-Modal Intelligence for solving unknown, open domain problems), (No. 2022-0-00688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes), (No. 2021-0-02068, Artificial Intelligence Innovation Hub), (No. 2019-0-00421, AI Graduate School Support Program at Sungkyunkwan University), and (No. 2021-0-02309, Object Detection Research under Low Quality Video Condition).

REFERENCES

[1] 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*. 3213–3216.

[2] Camila Souza Araújo, Gabriel Magno, Wagner Meira, Virgilio Almeida, Pedro Hartung, and Danilo Doneda. 2017. Characterizing videos, audience and advertising in Youtube channels for kids. In *International Conference on Social Informatics*. Springer, 341–359.

[3] Adam E Barry, Emily Johnson, Alexander Rabre, Gabrielle Darville, Kristin M Donovan, and Orisatalabi Efunbumi. 2015. Underage access to online alcohol marketing content: a YouTube case study. *Alcohol and alcoholism* 50, 1 (2015), 89–94.

[4] BBC. 2017. The disturbing YouTube videos that are tricking children. https://www.bbc.com/news/blogs-trending-39381889. [Online; Accessed 01-December-2021].

[5] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[6] Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. 2014. Project adam: Building an efficient and scalable deep learning training system. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*. 571–582.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[9] Common-sense. 2021. Parents' Ultimate Guide to YouTube Kids. https://tinyurl.com/2t49syfp.

[10] Alexandre Ashade Lassance Cunha, Melissa Carvalho Costa, and Marco Aurélio C Pacheco. 2019. Sentiment analysis of youtube video comments using deep neural networks. In *International Conference on Artificial Intelligence and Soft Computing*. Springer, 561–570.

[11] Anjali Dagar and Tatiana Falcone. 2020. High viewership of videos about teenage suicide on YouTube. (2020).

[12] Brian Dean. 2021. How Many People Use YouTube in 2021. https://backlinko.com/youtube-users. [Online; Accessed December,2021].

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 4171–4186. https://doi.org/10.18653/v1/N19-1423 arXiv:1810.04805v1

[15] Dropbox Link. 2022. Inappropriate videos for young children on YouTube Kids. https://tinyurl.com/5352spef.

[16] Carsten Eickhoff and Arjen P de Vries. 2010. Identifying suitable YouTube videos for children. *3rd Networked and electronic media summit (NEM)* (2010).

[17] FTC. 2021. YouTube channel owners: Is your content directed to children? https://tinyurl.com/56e7e5hp. [Online; Accessed 01-December-2021].

[18] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[19] The Guardian. 2018. How Peppa Pig became a nightmare video for children. https://www.theguardian.com/technology/2018/jun/17/peppa-pig-youtube-weird-algorithms-automated-content. [Online; Accessed 01-December-2021].

[20] Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Advanced Analytics, LLC.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.

[22] Mallory Hussin, Savannah Frazier, and J Kevin Thompson. 2011. Fat stigmatization on YouTube: A content analysis. *Body image* 8, 1 (2011), 90–92.

[23] Akari Ishikawa, Edson Bollis, and Sandra Avila. 2019. Combating the elsagate phenomenon: Deep learning architectures for disturbing cartoons. In *2019 7th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 1–6.

[24] Julia Jacobo. 2019. YouTube Kids video featuring suicide instructions removed after reports from parenting blog. https://tinyurl.com/2p95rrsd. [Online; Accessed 01-December-2021].

[25] Abhishek Kar, Christian Häne, and Jitendra Malik. 2017. Learning a Multi-View Stereo Machine. In *Advances in neural information processing systems*.

[26] Rishabh Kaushal, Srishty Saha, Payal Bajaj, and Ponnurangam Kumaraguru. 2016. KidsTube: Detection, characterization and analysis of child unsafe content & promoters on YouTube. In *2016 14th Annual Conference on Privacy, Security and Trust (PST)*. IEEE, 157–164.

[27] Kids Matters Counselling. 2022. 5 Things Parents Need to Know about Tik-Tok. https://tinyurl.com/yk8nu9xx.

[28] Kyongseok Kim, Hye-Jin Paek, and Jordan Lynn. 2010. A content analysis of smoking fetish videos on YouTube. *Health communication* 25, 2 (2010), 97–106.

[29] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[30] Tarald O Kvålseth. 1989. Note on Cohen's kappa. *Psychological reports* 65, 1 (1989), 223–226.

[31] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[33] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[34] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Ilias Blackburn, 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.

[35] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).

[36] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[37] Shubham Singh, Rishabh Kaushal, Arun Balaji Buduru, and Ponnurangam Kumaraguru. 2019. KidsGUARD: Fine grained approach for child unsafe video representation and detection. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. 2104–2111.

[38] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification?. In *China National Conference on Chinese Computational Linguistics*. Springer, 194–206.

[39] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[40] Rashid Tahir, Faizan Ahmed, Hammas Saeed, Shiza Ali, Fareed Zaffar, and Christo Wilson. 2019. Bringing the kid back into youtube kids. In *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE, 464–469.

[41] Rajat Tandon. 2022. https://sites.google.com/view/samba-kids/.

[42] Tech Transparency Project. 2022. Guns, Drugs, and Skin Bleaching: YouTube Kids Still Poses Risks to Children. https://tinyurl.com/3yefk73h.

[43] TechHQ. 2022. Meta under fire from UK watchdog over child safety in VR. https://techhq.com/2022/01/meta-under-fire-from-uk-watchdog-over-child-safety-in-vr/. [Online; Accessed 01-December-2021].

[44] The Guardian. 2022. YouTube Kids shows videos promoting drug culture and firearms to toddlers. https://tinyurl.com/bdf8jfcf.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[46] Matthijs J Warrens. 2012. The effect of combining categories on Bennett, Alpert and Goldstein's S. *Statistical Methodology* 9, 3 (2012), 341–352.

[47] Wavy. 2021. TikTok school threats prompt call for parents to talk to kids. https://tinyurl.com/mtnft4wv. [Online; Accessed 01-December-2021].

[48] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[49] YouTube. 2021. YouTube channel owners: Is your content directed to children? https://tinyurl.com/4r7sxcxy. [Online; Accessed 01-December-2021].