Hate, Obscenity, and Insults: Measuring the Exposure of Children to Inappropriate Comments in YouTube

Sultan Alshamrani
salshamrani@knights.ucf.edu
University of Central Florida
Orlando, Florida, USA

Ahmed Abusnaina
ahmed.abusnaina@knights.ucf.edu
University of Central Florida
Orlando, Florida, USA

Mohammed Abuhamad
mabuhamad@luc.edu
Loyola University Chicago
Chicago, Illinois, USA

Daehun Nyang
nyang@ewha.ac.kr
Ewha Womans University
Seoul, South Korea

David Mohaisen
mohaisen@ucf.edu
University of Central Florida
Orlando, Florida, USA

ABSTRACT

Social media has become an essential part of the daily routines of children and adolescents. Moreover, enormous efforts have been made to ensure the psychological and emotional well-being of young users as well as their safety when interacting with various social media platforms. In this paper, we investigate the exposure of those users to inappropriate comments posted on YouTube videos targeting this demographic. We collected a large-scale dataset of approximately four million records and studied the presence of five age-inappropriate categories and the amount of exposure to each category. Using natural language processing and machine learning techniques, we constructed ensemble classifiers that achieved high accuracy in detecting inappropriate comments. Our results show a large percentage of worrisome comments with inappropriate content: we found 11% of the comments on children's videos to be toxic, highlighting the importance of monitoring comments, particularly on children's platforms.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

YouTube Comments, Online Behavior Analysis, NLP.

ACM Reference Format:
Sultan Alshamrani, Ahmed Abusnaina, Mohammed Abuhamad, Daehun Nyang, and David Mohaisen. 2021. Hate, Obscenity, and Insults: Measuring the Exposure of Children to Inappropriate Comments in YouTube. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3442442.3452314

1 INTRODUCTION

The influence of social media on the intellectual and emotional well-being of children and adolescents has been the focus of many studies in recent years, with social media being a central daily activity in children and adolescents' lives alike [11]. Among the various platforms, YouTube is the most popular video-sharing platform and is commonly used by children as an alternative to traditional TV, and as a source of entertainment and educational materials alike. A recent study by [22] reported that 81% of U.S. parents allow their children to use YouTube as an entertainment activity. Moreover, another study shows that children under the age of eight spend 65% of their time on the Internet using YouTube [8]. Therefore, researchers have spent enormous efforts understanding the age-appropriate experience of children and adolescents when using YouTube, and have shown that inappropriate contents—such as contents with sexual hints, abusive language, graphic nudity, child abuse, horror sounds, and scary scenes—are common, with promoters for such contents targeting this demographic [12, 18, 23].

Parents and custodians trust children-oriented YouTube channels, such as Nick Jr., Disney Jr., and PBS Kids, to present educational and entertaining material for their children even with no supervision. However, children can be exposed to inappropriate and disturbing videos, suggested by the YouTube recommendation system, as children are tricked to click on innocent-looking thumbnail [18]. To ensure their well-being and safety, it is important to study the exposure of children and adolescents to inappropriate material presented on YouTube, including visual, audio, and written content. Even when watching videos from trusted family-friendly channels, the written content, such as user comments, might contain inappropriate language that could influence the children's offline behavior. The limited work on YouTube textual contents, as opposed to the various efforts on understanding YouTube's video/audio contents, creates the need for comments-based studies.

Our study explores measuring the exposure of children and adolescents to age-inappropriate comments posted on videos of the top-200 children shows [7]. This task is challenging for several reasons. First, studying comments on children's videos requires manually collecting channels and shows targeting this demographic, knowing YouTube categories are not established by age-group but rather by the topic they present. Second, assigning age groups to the collected videos can be daunting in measuring exposure by separate groups. Third, the lack of a ground truth dataset for safe and inappropriate content posted on such videos makes it difficult for machine-learning models to capture the children's exposure on a large scale. Considering the variety of age-inappropriate content for children, building a unified system for detecting such content is challenging.

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452314

To address those challenges, we built a large collection of YouTube comments on children-oriented videos for the top 200 shows categorized by different age groups [5]. We extended the dataset with ground truth data from different sources to establish five age-inappropriate categories; toxic, obscene, insult, and identity hate. The used ground truth dataset compresses annotated data provided by Conversation AI on Wikipedia's comments, and our manually-annotated data from YouTube comments posted on children's videos. We leveraged natural language processing and machine learning techniques to construct an ensemble of models, each of which specializes in detecting a specific inappropriate category. The models are trained and tested on ground truth samples, and separately and collectively achieve remarkable results. Utilizing our ensemble, we uncovered a large number of age-inappropriate comments among those posted on children's YouTube videos. Measuring the exposure by age group, our results show that children between 13 and 17 years old are the most exposed to such contents. For inappropriate categories, toxic-related comments are the most common, with 15.54% out of the total comments, then insult (7.96%) and obscene (6.84%).

Contribution. This work contributes to measuring the exposure of children to inappropriate content present in the kids' YouTube videos comments. We summarize our contribution as follows:

- We collected a large-scale dataset of comments on children's YouTube videos from the top-200 ranked children's shows. The list of shows, retrieved search results, categorization of shows by age group, and other artifacts related to the data collection process are manually vetted.
- We built a manually-annotated ground truth dataset collected from comments posted on children's videos, which includes about 6,000 comments.
- Leveraging natural language processing and deep learning techniques, we designed and implemented an ensemble of classifiers to detect five age-inappropriate contents. Models of the ensemble are trained, fine-tuned, and evaluated using the ground truth dataset.
- Adopting the ensemble classifier on the YouTube comments domain, we detected and measured children's exposure to inappropriate comments.
- We provided an in-depth analysis of children's exposure to inappropriate content in terms of age groups, user interactions, and YouTube video channels.

2 RELATED WORKS

Recently, several studies have been conducted with the aim of exploring the effects of social media on children, since the use of social media has become a significant part of their daily routines. To ensure the safety of kids on YouTube, Alshamrani *et al.* [2] studied the exposure of children to malicious URLs on videos targeting young users. Another study by [16] which has encouraged parents to understand and be aware of the various possible offline and online behaviors of their children, such as cyber-bullying, privacy issues, sexting, and Internet addiction. Among other social media platforms, YouTube has been the subject of many studies since it is considered the most popular social media platform in the United States [21], and the second-largest search engine after Google worldwide [15]. Studying the appropriateness of contents being presented to children on YouTube was first considered, to the best of our knowledge by Kaushal *et al.* [12] who studied kids-unsafe contents and promoters. The authors provided a framework for detecting unsafe contents using measures calculated on the video, user, and comment levels with an accuracy of 85.7%.

Table 1: The distribution of the collected dataset. The collected comments are from two sources: Wikipedia and YouTube. 5,940 YouTube comments are manually labeled for the evaluation of the ensemble models.

Source	Dataset	Count
Wikipedia	Safe Comments	143,000
	Toxic	15,294
	Obscene	8,449
	Insult	7,877
	Threat	478
	Identity hate	1,405
YouTube	Unlabeled	≈ 3,700,000
	Safe Comments	1,832
	Toxic	4,126
	Obscene	2,367
	Insult	1,650
	Threat	550
	Identity hate	788

Another work by Papadamou *et al.* [18] shows that inappropriate toddler-oriented videos are common and likely to be suggested by YouTube's recommendation system. Using manually-annotated videos, the authors investigated the detection of inappropriate content (containing sexual hints, abusive language, graphic nudity, child abuse, horror sounds, and scary scenes) collected from videos targeting kids using deep learning algorithms to achieve an accuracy of 84.3% for this task. More recently, Tahir *et al.* [23] demonstrated that even children-focused apps, such as *YouTube Kids* which is considered a kids-safe platform, are prone to compromise with inappropriate videos.

As part of studying users' comments, Alexandre *et al.* [4] studied and analyzed users' opinions on several aspects, such as the quality of the video, YouTuber presence, and videos' contents. Improving the content enables achieving higher popularity as Figueiredo *et al.* [9] outlined. In particular, the quality and user perception of the contents facilitate popularity on YouTube. Bermingham *et al.* [3] is another related work, in which they provided YouTube comment-based sentiment analysis of topics potentially serving a radicalizing agenda. Recently, several works explored the analysis and detection of hate speech in different social media platforms [1, 13, 24, 25]. This work studies and measures the exposure of children to age-inappropriate content in YouTube comments posted on children's shows.

3 APPROACH AND TECHNIQUES

3.1 Data Collection and Measurements

Our dataset includes YouTube comments and two datasets of ground truth, one from the Conversation AI team and another one annotated by our team for ground truth from the YouTube comments. For the YouTube comments, we collected more than 3.7 million comments

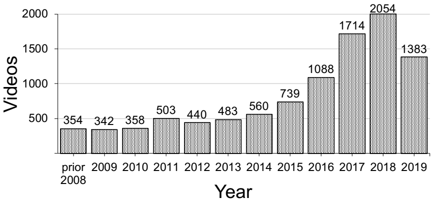


Figure 1: The publish date distribution of the collected YouTube kids' videos.

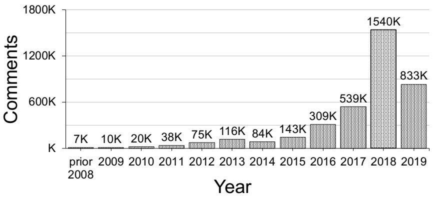


Figure 2: The distribution of YouTube kids' videos comments over past years.

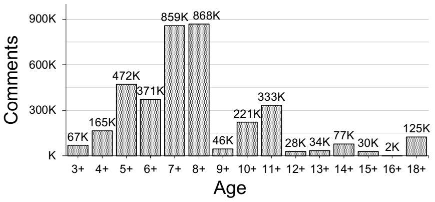


Figure 3: The distribution of YouTube kids' comments over different ages.

posted on roughly 10,000 children's videos, distributed over the period from January 2005 until March 2019.

Children's Shows. We collected comments on videos of the top-200 children's shows based on Ranker [7], a crowdsourced platform that relies on millions of users to rank a variety of media contents such as shows and films. The list of shows was originally made by Ranker TV and received more than 1.2M votes, and has 380 kids' shows. Among them, we selected the top 200 shows. We augmented our list with part of Wikipedia's list of cartoon shows.

Collection Approach. Using YouTube APIs, we extracted the top-50 videos of the search results on every show on our list. Using each retrieved video's ID, we also used the API to obtain video statistics, such as the number of views, likes, dislikes, etc. We used YouTube Comments API to collect all comments from the videos. In total, we collected more than 3.7 million comments from 10,000 videos.

Age-Appropriateness of Children's Shows. We defined age appropriateness as the adequate age group to be the subject of the show. Defining the age appropriateness for children's shows is challenging since most shows do not specify the target age group. Therefore, we used *Common Sense Media* [5], a non-profit organization that provides education and advocacy to families on providing safe media for children, as the main source for defining the age group of the targeted children's shows. Using *Common Sense Media*, we were able to retrieve the appropriate age group for most of the kids' shows on our list. However, a few shows do not appear in *Common Sense Media*, and for those we turned to IMDB [6], an online database of information about different types of media such as films, television programs, home videos, video games, etc., to obtain the age group for those particular shows. Some shows have different versions, each for a certain age group, therefore the age group is assigned based on the most prevalent version in the YouTube search. Some other kids' shows are assigned an age group based on their respective categories, e.g., Loony Tunes (a well-known collection of cartoons for age 7+). We note that we conducted a manual inspection on the age appropriateness for the retrieved top-50 results on each show to define non-kids contents and assigned them to 17+ age group, which is the highest age group in our dataset.

Data Statistics and Measurements. Here we provide general statistics of our data. The collected YouTube comments were posted by more than 2.5 million users on about 10,000 videos from more than 3,000 different channels. These retrieved videos have an average viewers count of roughly 2.4 million views and an average comments count of 8,068 comments per video. Observing the publishing date of the videos in our collection, Figure 1 demonstrates the rapid increase in children's videos over the past few years. The figure shows an increase in popularity of five folds in ten years from 2008 (with 354 videos) to 2018 (with 2,054 videos). This rapid growth in popularity is observed through the first three months of 2019 with 1,383 videos included in our collection (by March of 2019). We note that the collection of YouTube videos is based on their relevance, and not the publishing date nor the view count; this is also the case when retrieving videos from the top-50 search result and when querying the targeted shows. The search results do not always reflect the popularity. However, the top-ranked videos are often characterized by bursts of popularity [10]. Generally, a consistent trend is observed in the year-over-year increasing number of videos included in our collection. Similar patterns are observed with the number of comments from around 7,000 comments on videos prior to 2008 to more than 1.5 million comments on videos from 2018. This growth is steady through the first three months of 2019 as illustrated in Figure 2. We also provided the distribution of comments across the age groups as shown in Figure 3 where most of the collected comments were posted on videos for kids between the age of five and eight (a total of approximately 2.5 million comments).

Age-Appropriateness of Contents. Contents that are regarded as age-appropriate for children and adolescents ideally should not contain toxic words or imply an insult, threat, identity hate, or obscenity. To study the appropriateness of YouTube comments, we collected ground truth datasets to establish a baseline for modeling contents with different labels (i.e., toxic, obscene, insult, threat, and identity hate). The ground truth data includes: (1) labeled comments from Wikipedia that is manually annotated by Conversation AI, a research team started by Jigsaw and Google to provide tools and solutions for improving online conversions; (2) labeled comments posted on YouTube videos targeting children that are manually annotated for the purpose of this study.

(1) Wikipedia Ground Truth Toxic Dataset. We used the manually-annotated dataset provided by Conversation AI, with approximately 160,000 comments from Wikipedia Talk pages of which approximately 143,000 comments are labeled as safe, while the remaining are labeled to have different types of toxicity (i.e., 15,294 toxic, 8,449 obscene, 478 threat, 7,877 insult, and 1,405 identity hate). A summary of the collected data is provided in Table 1.

(2) Manually Annotated Ground Truth. We manually annotated 5,958 YouTube comments posted on YouTube videos for the evaluation of the ensemble. The total number of the manually labeled comments is distributed as follows: safe: 1,832, toxic: 4,126, obscene: 2,367, insult: 1,650, threat: 550, and identity hate: 788.

For our manual labeling, we used several explicit rules. Each comment was labeled as either toxic or safe. A toxic comment may belong to one or more unsafe categories; obscene, threat, insult, or identity hate. A comment is considered obscene when it is morally offensive in a sexual way, or when it has socially offensive words. When such offensive language is used against or to describe other users, video publishers, or anyone else, the comment is considered an insult. When such offensive language is directed to another group of people, by imposing a negative stereotype or prejudices about people based on their race, color, or ethnicity, the comment is considered as identity hate.

The annotation is challenging since identifying identity hate is highly subjective [17] [20]. Some comments did not have any profanity or offensive language, but implied a threat to other users or the video publisher; we labeled such a comment to be a threat. In the annotation process, we encountered comments that are socially unacceptable and are age-inappropriate, however, they do not belong to any of the four unsafe categories, and so we labeled them as *toxic* only. The manual labeling has been done by the same annotator (lead author of this work), upon refining the above ruleset. We avoided using multiple annotators across different folds of the manually-labeled dataset, and rather pursued this slow labeling method, to avoid inconsistency and subjectivity in interpretation against the predetermined labeling rules.

Ground Truth: Safe Dataset. For safe content, we used our labeled safe YouTube comments as well as safe-labeled comments from the Conversation AI team dataset, which include roughly 143,000 comments in total.

3.2 Data Preprocessing

Several preprocessing steps are taken before the final data representation, modeling, and evaluation, and to ensure a clean and proper representation of the collected data. YouTube comments are the focus of this study, which we addressed with the preprocessing steps as follows: (1) We initially removed all *non-English contents* across all datasets, and limit our analysis to English comments. (2) We eliminated unwanted characters and tokens, e.g., punctuation, and other characters that represent or encode emojis.

3.3 Data Representation

Comments Data Representation. In order to perform an analysis of textual data, we first transformed this data into an embedding (i.e., numerical representation) that can be used by machine learning models. Such a representation allows the machine learning models to learn and capture different patterns of the text. We utilized different data representation methods, namely, *Word2Vec* [14] and *Glove* [19].

Pre-trained Word2Vec. Using the pre-trained model for comments representation, we have the following two cases of distinct models. **(1) Gensim:** This technique transforms textual data by examining word statistical co-occurrence patterns within a corpus of the provided textual documents. Examining different configurations for both word embedding and the document vector. We found that the highest accuracy can be achieved using a size of 300 for the word embedding and the document vector size of 50. **(2) Glove:** This technique is an unsupervised learning algorithm used to generate numerical vector representations for words. The training process is done on aggregated global word-word co-occurrence statistics from a corpus. We used Glove to represent the comments; similar to Gensim, we tried different configurations and selected the configuration with the highest accuracy, using a size of 50 for the word embedding and 100 for the document vector.

3.4 Ensemble Classification Models

To understand and measure children's exposure to inappropriate comments on YouTube videos by first identifying them, we adopted an ensemble classifier to build five specialized models for classifying five unsafe categories: *toxic*, *obscene*, *threat*, *insult*, and *identity hate*. The models are trained, in a supervised manner, using the Wikipedia toxic comments dataset and the manually annotated ground truth of YouTube comments. Each model predicts whether an input belongs to a specific category, functioning as a binary classification task. We note that a comment can belong to one or more categories (e.g., toxic, insult, and identity hate simultaneously), thus the output of the ensemble is positive if the comment is labeled as at least one age-inappropriate category.

Our approach adopts an ensemble of classifiers to predict different age-inappropriate categories using DNN models. Based on our experiment, DNN performs very well in terms of identifying different age-inappropriate categories as opposed to CNN and RNN. We found that different pre-trained models for feature representation, such as Glove and Gensim, work better in certain scenarios for identifying certain age-inappropriate comments categories (i.e., Glove with DNN for identifying threat comments). The ensemble uses DNN for identifying five age-inappropriate categories, DNN with gensim Word2Vec for identifying toxic, obscene, insult, and identity hate categories, and DNN with Glove Word2Vec for identifying threat category.

We feed Word2Vec vectors of the comments to the first input layer in the network while the output layer has a single node for binary classification to predict whether the provided comment belongs to a certain class or not. Our model architecture is composed of two dense layers of size 128 units with a ReLU activation function, each followed by a dropout operation with a rate of 20%. The last layer is fully connected to a sigmoid function, which generates real values in the range (0,1) using the function $sigmoid(z) = 1/(1 + e^{-z})$. Since the output $\{y \in \mathbb{R} \mid 0 \leq y \leq 1\}$, determines the probability of assigning an input to the target, a threshold can be defined for target $\bar{y}$ assignment (e.g., a commonly-used threshold is 0.5 where $\bar{y} = 1 \; if \; y \geq 0.5$). We explored different thresholds for each category to optimize the true negative rate and the true positive rate.

Model Training Settings. We used the entire Wikipedia annotated comments to train five models, each of which is specialized in detecting one age-inappropriate category. Then we fine-tuned the trained model using 50% of our manually labeled comments from YouTube, by only retraining the last layer of the model. We then used the other half for the evaluation of the models. The training process is guided by minimizing the binary-cross-entropy as follows:

$$\text{loss}(\theta) = \frac{-1}{N} \sum_{i=1}^{N} [y_i \times \log(p_i) + (1 - y_i) \times \log(1 - p_i)],$$

where p_i is the conditional probability $p(y_i|x_i, \theta)$ for a target y_i given an input x_i and a set of parameters θ, i is the i-th record, and N is the total number of records in the training set. The optimization is

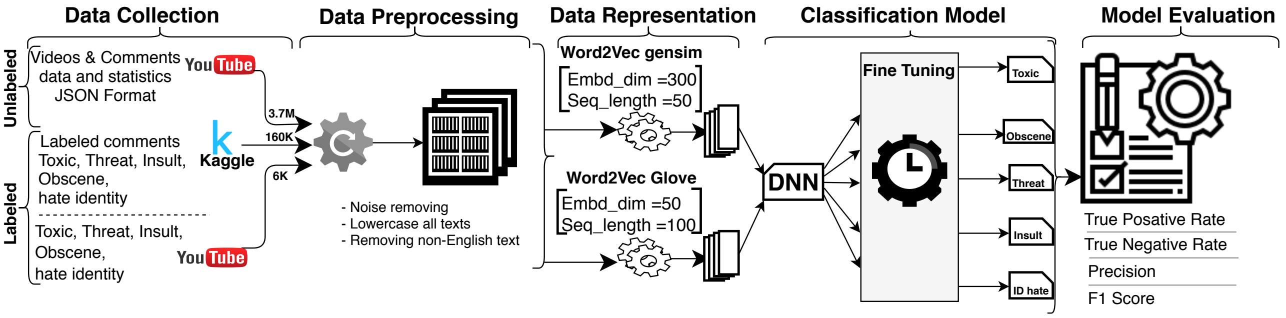


Figure 4: The ensemble pipeline. The system design consists of five stages, starting from data collection and labeling, followed by the preprocessing of the data to generate efficient representation. Then, the ensemble of five classification models is used for comments classification. Further, the models are evaluated using four evaluation metrics.

Table 2: The performance of the ensemble model on Wikipedia comments, and the fine-tuned models across different metrics. Overall, the fine-tuned ensemble achieved a TNR of 86.8% and TPR of 78.5%.

Class	Wikipedia			Fine Tuned		
	Recall	Prec	F1	Recall	Prec	F1
Toxic	92.5	82.5	87.2	93.5	83.1	88.0
Obscene	81.9	82.9	82.4	86.6	83.5	85.0
Threat	64.4	43.7	52.1	71.3	42.3	53.1
Insult	74.5	55.7	63.3	66.7	64.4	65.6
Identity hate	53.9	89.8	67.4	74.8	87.8	80.8
Overall	73.4	70.9	70.4	78.5	72.2	74.5

done using *RMSprop* optimizer, a stochastic optimization algorithm, with a learning rate of 10^{-3} without decaying over time. We used a mini-batch approach with a batch size of 128, and for preventing the overfitting we used dropout regularization with a dropout rate of 0.2. The termination criterion is set to be a specified number of training iterations, which is set to 100 for all models.

Evaluation Metrics. This study uses four evaluation metrics, which are *Precision*, *F1-score*, *True Positive Rate* (TPR), and *True Negative Rate* (TNR). Precision represents the percentage of which a model was correct in predicting the positive class (P = TP/TP+FP). F1-score is the harmonic mean of the precision and recall, and is expressed as (F1-score = 2TP/2TP+FP+FN) where TP, FP, and FN represent True Positive, False Positive, and False Negative, respectively. The TPR is the proportion of the positive predictions, positive labeled-data correctly predicted to be positive, form the total positive-labeled data (TPR = TP/TP+FN). The TNR is the proportion of the negative predictions, negative labeled-data correctly predicted as negative, from the total of negative-labeled data (TNR = TN/TN+FP).

4 RESULTS AND DISCUSSION

In this section, we review the results of the ensemble for classifying five categories of inappropriate contents, including, toxic, obscene, threat, insult and identity hate. Then, we measured children's exposure to inappropriate comments on YouTube using the best-performing models.

4.1 Ensemble Model Performance

The ensemble model performance is reported in Table 2 using three metrics. We reported the performance of the models trained on Wikipedia then evaluated on the annotated YouTube comments as well as the performance of these models after being fine-tuned. The results are based on the specific probability threshold providing the best trade-off between TPR and TNR as shown in Figure 5. An emphasis on high TPR is considered when choosing the threshold to ensure high correctness for positively predicted output (i.e., some positive contents might not be detected but barely mistaken when they are detected). This high performance can be seen with the F1-score, with a high of 86.6% for the toxic and a low of 52.9% for the threat. We also observed the challenge in achieving high TPR for the threat and identity-hate categories due to several reasons, including the limited number of samples for those categories (see Table 1) and the ambiguity caused by the used language.

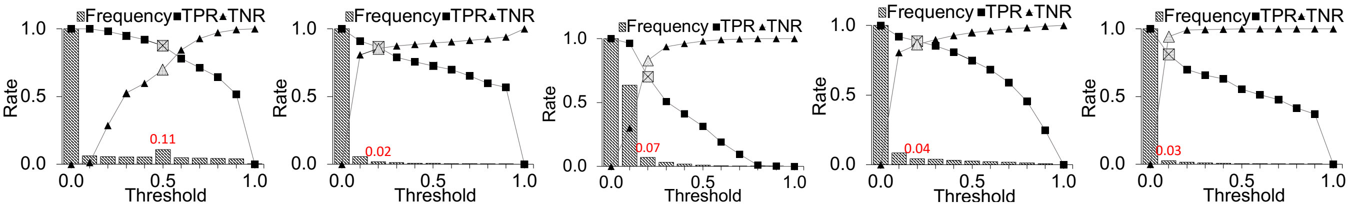
4.2 Ensemble Adoption and Measurement

Using the best TPR-TNR trade-off thresholds, we constructed an ensemble model to evaluate and measure kids' exposure to inappropriate comments. We first show the measurement using the individual models, followed by the overall performance of the ensemble of multiple models for the multi-label classification task.

(1) Toxic Comments. We measured the toxicity of YouTube comments using the toxic comments detection model. Figure 5(a) shows the performance of the model in terms of TPR and TNR using different thresholds, and 0.520 is selected as the threshold with the best trade-off. Applying the model on our dataset, Figure 5(a) shows 11% (405,290 comments) of all comments were classified as toxic.

(2) Threat Comments. Similarly, the model for detecting threat comments achieved a TNR of 86%. We set the threshold for this category to 0.220, providing the best trade-off with a TPR of 85% as shown in Figure 5(b). Adopting the model to detect threat comments, 2% of the comments (63,939 comments) were labeled as a threat.

(3) Insult Comments. The insult comments model provides a TPR of 66% and TNR of 85%. Figure 5(c) shows the results using different thresholds. In our design, we selected 0.210 as a threshold for predicting insult comments. Using this model with the adopted threshold, 7% of the collected comments are detected as an insult (262,934 comments).



(a) Toxic (b) Threat (c) Insult (d) Obscene (e) Identity hate

Figure 5: The evaluation of the ensemble model across categories in terms of TPR and TNR. The x-axis represents the chosen threshold, and y-axis shows the respective TPR, TNR, and percentage of detected YouTube comments.

(4) Obscene Comments. The obscene comments model, operating with a prediction threshold of 0.270, achieves a TPR of 86% and TNR of 88%. Figure 5(d) shows the results of adopting different thresholds, most of which provide high scores. Applying the model on the comments, 4% were detected as obscene (159,823 comments).

(5) Identity Hate Comments. The model for detecting identity hate comments shows a high performance as demonstrated in Figure 5(e). Using a prediction threshold of 0.140, the achieved TPR and TNR are 74% and 98%, respectively. Applying the model to YouTube comments, we found that among the comments, 3% were labeled as identity hate comments or 101,311 comments.

4.3 Inappropriateness Exposure Analysis

Exposure by Age-Group. Applying the ensemble models shows the exposure magnitude of kids to inappropriate content on YouTube comments. Investigating the exposure by different age groups, Figure 8 shows the distributions of the inappropriate comments from each age-inappropriate category over different age groups. For simplicity, we studied the contents of comments posted on YouTube videos targeting different age groups instead of the age in years (distributions of collected comments on videos for a specific age is shown in Figure 3). Applying the ensemble models on the collected comments, we observed that toxic comments are highly common in children's videos and exceed 200,000 comments on videos only targeting the age group of six to eight years old. Insulting comments can be clearly noticed in videos targeting young children; e.g., age group 3-5 has 48,306 comments, which corresponds to 6.83% out of the total comments collected on videos of this age group (707,161 comments). Comments with some sort of toxicity are also present in the collected dataset with 81,303 toxic, 17,384 obscene, 48,306 insult, 9,065 threat, and 21,329 identity hate which were detected in comments posted on videos for the age group of 3-5. These records increase to 241,352 and 36,150 for toxic and insult, respectively, on videos for the age group of 6-8. These patterns of appearance for toxic comments are observed for videos targeting all age groups. The number of comments that contain obscene, threat, and identity hate are noticeably high for all age groups (e.g., they reach 99,165, 36,150 and 53,517, respectively, for the age group 6-8). We note that the reported numbers of detected categories of inappropriate comments in Figure 8 do not reflect their percentage with respect to the total number of comments for a certain age group. We observed that children in the age group of 3-5, which are the youngest audience, are the second most exposed to inappropriate comments, with 7.71%, 2.46%, 6.83%, 1.28%, 3.02% for toxic, obscene, insult, threat and identity hate categories (out of the total), respectively. This age group is only second to the 13-17 age group, which has 15.54%, 6.84%, 7.96%, 2.24%, 4.20%, for the same types.

Exposure and User Interaction. Acquiring YouTube kids videos, where comments were collected and investigated, is done using the top-50 search results from the YouTube Search APIs with measures of relevance and popularity (i.e., it is safe to state that the considered videos are popular). We show statistics of users' interactions with videos that contain different inappropriate content (for the five investigated categories) in Figure 6. Considering the number of videos with age-inappropriate comments, we observed that the highest number of videos (6,037 videos) are reported for those with insulting comments, which has the second most number of comments among other categories (262,934). The videos with threatening comments have an average of 6.4 million views and 18,640 likes per video. More interestingly, videos with threatening comments tend to get higher interaction in terms of the number of likes (18,640) than videos with either obscene or identity hate comments (an average of 17,000 comments). Another observation is that the number of dislikes for the videos is positively proportional to the number of threat and identity hate comments.

We explored user interaction with inappropriate comments in terms of the number of likes and replies. Figure 7 shows the average number of likes and replies for comments that belong to the five inappropriate categories. The more likes and replies a comment gets will increase the likelihood of that comment being shown in the top comments. We have noticed that threatening and insulting comments have the highest average of likes and replies, e.g., around 11 likes and 0.5 replies per comment for the threat category. As opposed to the other age-inappropriate categories, identity hate, and insulting comments have a high number of average replies, with an average reply of 0.53 per comment. Even though the users' interaction with comments from other categories is less than threatening and insulting comments, the interaction can be seen for all categories in Figure 7.

Exposure by YouTube Channel. Investigating the top-10 most comment-contributing YouTube channels to our collected comments, Table 3 shows the distribution of age-inappropriate comments across different channels with respect to the five investigated categories. The table highlights the number of videos of which we collected the comments as well as the number of collected comments enabling the estimation of the percentages of inappropriate comments. The highest number of detected inappropriate comments is reported for *moviemaniacsDE* channel with 43.2% of the total comments classified as inappropriate. Furthermore, there is an alarming number

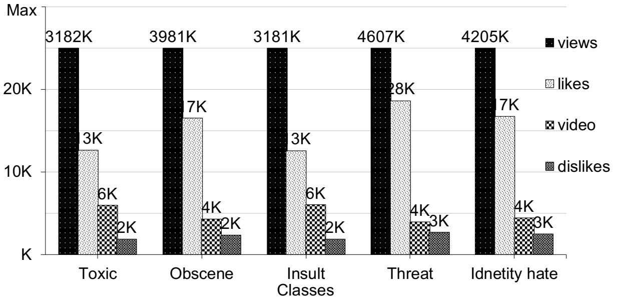


Figure 6: The average number of views and likes on kids' videos containing inappropriate comments. YouTube kids' videos with unsafe comments have a high number of views, likes, and dislikes.

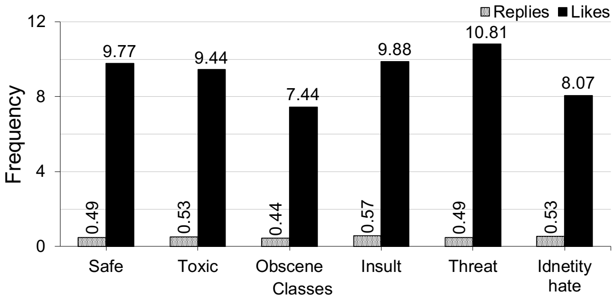


Figure 7: The average number of likes and replies on inappropriate comments. On average, comments associated with threat, insult and identity hate have a higher number of likes and replies.

Table 3: Distribution of the inappropriate comments over different YouTube Channels as well as the average of the inappropriate comments to the overall comments posted on each category.

Channel Name	# Video	# Comments	Safe	Toxic	Obscene	Insult	Threat	Identity hate	Unsafe / video	Unsafe / comment
Warner Bros. Pictures	3	140594	113569	20994	10980	8002	1761	2954	9008	19.2
Cartoon Hangover	52	118352	101385	11013	4191	6625	1926	1664	326	14.3
Talking Tom and Friends	44	99293	88935	5374	460	4491	1077	2272	235	10.4
Cartoon Network	81	89620	80142	4003	137	3320	1956	1763	117	10.6
moviemaniacsDE	3	21052	11948	7012	3795	3631	595	1397	3035	43.2
Flashback FM	16	40833	29301	8788	4202	4376	940	1170	721	28.2
Mickey Mouse	46	56423	50159	2230	106	3411	561	1183	136	11.1
Nickelodeon	38	46097	42730	1222	66	1629	489	595	89	7.3
DEATH BATTLE!	3	45652	39448	3608	810	2242	1094	634	2068	13.6
Official Pink Panther	60	41730	36950	1388	175	1738	353	2197	80	11.5

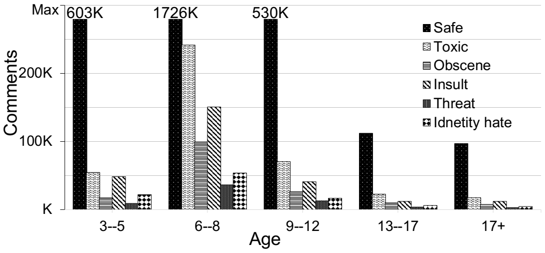


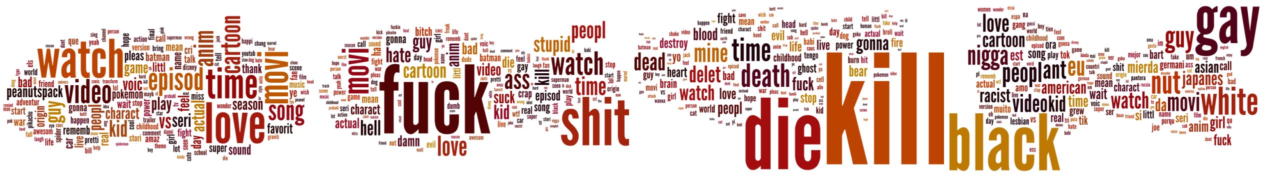
Figure 8: The distribution of inappropriate comments over different age groups.

of unsafe comments posted on the *Warner Bros. Pictures* channel videos, where the average number of inappropriate comments is 9,008 comments per video. In contrast, *Official Pink Panther* has the lowest average number of unsafe comments per video, with only 80 comments per video. We also observed a high number of detected inappropriate comments from the Nickelodeon channel, with 7.3% of the total comments in this channel classified as inappropriate. This percentage was the lowest among other channels, although still an alarming score of exposure to inappropriate comments for impressionable children.

4.4 Discussion

YouTube Platform for Children. Social media has become well-established and a part of most people's daily routine [11]. Many studies have shown that children under the age of 18 spend a substantial amount of time on social media, especially on YouTube. A survey, conducted by the Pew Research Center in 2018, shows that 81% of parents in the United States with children younger than 11 years of age allow their children to watch YouTube videos, and 34% of parents stated that their children watch YouTube videos regularly [22]. The collection of our dataset confirms the rapid growth of popularity for YouTube videos targeting children. More importantly, the results show that posted comments on children's videos contain contents that are inappropriate, this might affect their safety, privacy, intellect, emotion, or/and behavior. We note that YouTube established the *YouTube Kids* mobile app (in February 2015) and website (in August 2019), a safe platform for kids where the comment feature is disabled. However, a large percentage of children; i.e., 80% according to a study by [5], still use YouTube's original website and/or mobile app. Therefore, and based on our study, children who use YouTube unsupervised might encounter inappropriate content in the comments section, highlighting the risks of media platforms, and calling for measures to ensure their safety online.

Awareness of Inappropriate Comments. This study sheds light on the exposure of adolescents to inappropriate comments on YouTube, and shows that visual and audio are not the only media that should be supervised but also the written contents. Figure 9 shows some of the frequently inappropriate words detected to be one of the five age-inappropriate categories investigated in our study from comments



(a) Safe (b) Toxic (obscene, insult) (c) Threat (d) Identity hate

Figure 9: The most frequent words in YouTube comments per category. Since *toxic, obscene, and insult* share similar frequent words, we represented them in one cloud.

posted on children's videos. This study shows that among inappropriate comments, there exists a large number of comments that have toxic, threatening, insulting, or/and identity hate contents which possibly can influence the psychological well-being of children.

5 CONCLUSION

In this work, we studied the exposure of kids to inappropriate and comments posted on kids' YouTube videos. We studied the exposure to five age-inappropriate categories, namely, toxic, obscene, insult, threat, and identity hate. Using an ensemble of specialized models trained on labeled data, we measured the exposure of each category by different age groups to find out that the age group of 13-17 is the most exposed group to the inappropriate comments followed by the 6-8 age group. The results show that toxic comments are common on children's videos with 10.95% of the total comments having toxic language, followed by insults (7%), obscene (4.32%), identity hate (2.74%), and threat (1.73%) comments. We also measured users' interactions (views, likes, and dislikes) with videos having age-inappropriate comments as well as the comments themselves. We found that videos and comments with toxic or threatening comments tend to have higher interaction. Videos with threat comments have a high degree of popularity with an average of 4.6 million views and 28,000 likes per video. Similar popularity is observed for comments promoting identity hate with an average of 4.2 million views and 17,000 likes per video. This research shows that children are exposed to inappropriate comments, and call for increased awareness of such exposure and take measures to ensure children's safety from this exposure while on YouTube.

Acknowledgement. This work was supported in part by NRF grant NRF-2016K1A1A2912757 (GRL program), GPU award by NVIDIA.

REFERENCES

[1] Sultan Alshamrani, Mohammed Abuhamad, Ahmed Abusnaina, and David Mohaisen. 2020. Investigating online toxicity in users interactions with the mainstream media channels on YouTube. In *International Workshop on Mining Actionable Insights from Social Networks*. 1–6.

[2] Sultan Alshamrani, Ahmed Abusnaina, and David Mohaisen. 2020. Hiding in Plain Sight: A Measurement and Analysis of Kids' Exposure to Malicious URLs on YouTube. In *The ACM/IEEE Workshop on Hot Topics on Web of Things*. 1–6.

[3] Adam Bermingham, Maura Conway, Lisa McInerney, Neil O'Hare, and Alan F Smeaton. 2009. Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. In *International Conference on Advances in Social Network Analysis and Mining, ASONAM*. 231–236.

[4] Alexandre Ashade Lassance Cunha, Melissa Carvalho Costa, and Marco Aurélio Cavalcanti Pacheco. 2019. Sentiment Analysis of YouTube Video Comments Using Deep Neural Networks. In *International Conference in Artificial Intelligence and Soft Computing, ICAISC*. 561–570.

[5] Developers. 2021. Common Sense Media. www.commonsensemedia.org. Accessed: 2021-13-01.

[6] Developers. 2021. IMDB. www.imdb.com. Accessed: 2021-13-01.

[7] Developers. 2021. Ranker. https://www.ranker.com/crowdranked-list/my-favorite-cartoons-of-all-time. Accessed: 2021-13-01.

[8] FamilyZone. 2020. FamilyZone. https://www.familyzone.com/au/families/blog/what-kids-did-online-2016

[9] Flavio Figueiredo, Jussara M. Almeida, Fabrício Benevenuto, and Krishna P. Gummadi. 2014. Does content determine information popularity in social media?: a case study of youtube videos' content and their popularity. In *Conference on Human Factors in Computing Systems, CHI*. 979–982.

[10] Flavio Figueiredo, Fabrício Benevenuto, and Jussara M. Almeida. 2011. The tube over time: characterizing popularity growth of youtube videos. In *International Conference on Web Search and Web Data Mining, WSDM*. 745–754.

[11] Urs Gasser, Sandra Cortesi, Momin M Malik, and Ashley Lee. 2012. Youth and digital media: From credibility to information quality. *Berkman Center Research Publication* 2012-1 (2012), 20–40.

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

[13] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *The ACM Human–Computer Interaction Journal* 3, CSCW (2019), 207:1–207:21.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations, ICLR*. 1–12.

[15] MushroomNetworks. 2020. How SD-WAN/Multi-WAN Technology Handles the Data Avalanche from Youtube (Infographic). www.tinyurl.com/ybwzmaxe

[16] Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, et al. 2011. The impact of social media on children, adolescents, and families. *Pediatrics* 127, 4 (2011), 800–804.

[17] Alexandra Olteanu, Kartik Talamadupula, and Kush R. Varshney. 2017. The Limits of Abstract Evaluation Metrics: The Case of Hate Speech Detection. In *Web Science Conference, WebSci*. 405–406.

[18] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2020. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. 14 (2020), 522–533.

[19] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing, EMNLP*. 1532–1543.

[20] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint:1701.08118* (2017), 1–4.

[21] Aaron Smith and Monica Anderson. 2018. Social Media Use in 2018. tinyurl.com/y3htxhlq. Accessed: 2021-13-01.

[22] Smith, Aaron and Toor, Skye and Kessel, Patric Van. 2018. Many Turn to YouTube for Children's Content, News, How-To Lessons. https://www.pewresearch.org/internet/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/. Accessed: 2021-13-01.

[23] Rashid Tahir, Faizan Ahmed, Hammas Saeed, Shiza Ali, Fareed Zaffar, and Christo Wilson. 2019. Bringing the Kid back into YouTube Kids: Detecting Inappropriate Content on Video Streaming Platforms. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM*.

[24] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. 2021. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *IEEE Security and Privacy, S&P*. 1–21.

[25] Savvas Zannettou, Mai ElSherief, Elizabeth M. Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and Characterizing Hate Speech on News Websites. In *Conference on Web, WebSci*. ACM, 125–134.