AUTOMATED INAPPROPRIATE CONTENT DETECTION IN YOUTUBE VIDEOS USING DEEP LEARNING TECHNIQUES

[1]MURALI MOHAN REDDY M, [2]PARSHAVENI UMA MAHESHWARI
[1]Assistant Professor,[2]Student
Department of CSE
Sree Chaitanya College of Engineering, Karimnagar

ABSTRACT

YouTube's video content has grown exponentially, drawing billions of viewers, most of whom are in the younger age range. Additionally, malicious uploaders use this site as a means of disseminating disturbing visual information. For example, they use animated cartoon videos to disseminate stuff that is improper for children. Therefore, it is strongly advised that social media networks have an automated real-time video content screening method. This paper proposes a new architecture based on deep learning for the identification and categorisation of objectionable information in films. A pre-trained convolutional neural network (CNN) model called EfficientNet-B7 is used in the proposed framework to extract video descriptors, which are then fed into a bidirectional long short-term memory (BiLSTM) network to enable multiclass video classification and the learning of effective video representations. In order to apply the attention probability distribution in the network, an attention mechanism is also included after the BiLSTM. A carefully annotated dataset including 111,156 cartoon clips gathered from YouTube videos is used to assess these algorithms. EfficientNet-BiLSTM (accuracy D 95.66%) outperforms the attention mechanism-based EfficientNet-BiLSTM (accuracy D 95.30%) framework, according to experimental data. Second, deep learning classifiers outperform typical machine learning classifiers in terms of performance. All things considered, the EfficientNet and BiLSTM design with 128 hidden units produced state-of-the-art results (f1 score D 0.9267). Additionally, the performance comparison against current cutting-edge methods confirmed that BiLSTM on top of CNN captures better contextual information of video descriptors in network architecture, leading to better results in the detection and classification of inappropriate video content for children.

An essential element of the expanding discipline of data science is machine learning. Several types of algorithms are taught to create predictions or classifications and to unearth important insights in this project by use of statistical methodologies. Subsequently, these insights inform business and application decisions, which ideally influence important growth metrics.

Without being specifically taught to do so, machine learning algorithms create a model based on this project data, sometimes referred to as training data, in order to generate predictions or judgements. In many different datasets, machine learning algorithms are used when it is impractical or impossible to create traditional algorithms to carry out the necessary tasks.

1. INTRODUCTION

Over the last several years, there has been a significant increase in both the production and consumption of films on social media sites. Of all the social media platforms, YouTube is the most popular for sharing videos, offering a vast array of films in many categories. Over 2 billion people have registered on YouTube worldwide, and more than 500 hours of video material are posted every minute, according to YouTube statistics [1]. As a result, viewers of all ages may discover both general and personalised information throughout billions of hours of films [2]. In light of such an extensive crowdsourced database, it is very difficult to monitor and control the supplied material in accordance with platform requirements. Aasia Khanum served as the associate editor in charge of organising the review of this article and authorising it for publication. Because of this, malevolent individuals have more opportunity to engage in spamming activities by deceiving audiences with information (text, audio, or video) that is fraudulently marketed. Malicious users' most disruptive actions include exposing young audiences to upsetting information, especially when it is presented as safe for them. Since children spend the majority of their waking hours online, YouTube has become a popular substitute for conventional screen media, such as television [3], [4]. Less limitations are the cause for this high level of approbation, as the YouTube news release [5] also affirmed the social media site's great appeal among younger viewers as compared to other age groups [6].

Because there are no restrictions on the Internet, children may be exposed to any kind of information, unlike television. Among the several risks to children's online safety is the exposure of them to distressing material (such cyberbullying, cyber predators, hatred etc.) [7]. Frequent exposure to upsetting video material may have a short- or long-term effect on children's behaviour, emotions, and cognitive abilities, according to Bushman and Huesmann [8]. Inappropriate information is increasingly being distributed in children's videos, according to several studies [9]– [12]. The Elsagate controversy, which involved popular childhood cartoon characters (such as Disney characters, superheroes, etc.) being portrayed in disturbing scenes such as stealing, performing mild violence, engaging in nudity or sexual activities, and drinking alcohol, caught the attention of the mainstream media and brought attention to this trend [13], [14]. Laws such as the Children's Online Privacy Protection Act (COPPA) compel websites to include safety measures for children under the age of thirteen in an effort to offer a secure online environment. Additionally, YouTube has a "safety mode" option for filtering out harmful material. In addition, YouTube created the YouTube Kids app to provide parents control over videos that are deemed appropriate for kids in a certain age range [15]. Due to the difficulties in detecting such material, unsettling videos continue to surface [16]–[19] even on YouTube Kids [20], despite YouTube's best attempts to regulate the harmful content phenomenon.

One possible reason for this might be that YouTube becomes susceptible to inappropriate material due to the minute-by-minute posting pace of videos. Furthermore, a lot of the video's metadata—that is, its title, description, view count, rating, tags, comments, and community flags—is used by YouTube's algorithms to make decisions.

Therefore, community flagging and metadata-based video screening are insufficient to ensure children's safety [21]. Safe video names and thumbnails are often utilised for unsettling material on YouTube in an attempt to deceive kids and their parents. One further prevalent tactic used by fraudulent uploaders is the sparse addition of information deemed improper for children in videos. One such instance is seen in Fig. 1, where the video clip and title are suitable for children (Fig. 1(a)), but the movie contains scenes that are improper (Fig. 1(b) and Fig. 1(c)). The troubling aspect about this example, as well as many other situations like it, is that these films have been up for years and have had millions of views with a higher number of loves than dislikes. Numerous more situations (shown in Fig. 1(d)) were also found, whereby YouTube videos or the channel itself are not well-liked, but nonetheless include material that is harmful to children, particularly animated cartoons. Examples make it clear that this issue exists regardless of how popular a channel or video is. Additionally, YouTube blocked the dislike option, which prevented users from receiving statistics-based indirect input on the video content. It is advised to utilise video characteristics rather than metadata features connected to videos for the identification of improper material since YouTube information may be readily changed [22].

2. EXISTINGSYSTEM

For the purpose of detecting unlawful material in movies, Rea et al. [37] presented a periodicity-based audio feature extraction technique that was then integrated with visual characteristics.

Typically, the machine learning algorithms are used as classifiers. Using a Gaussian radial basis function (RBF) kernel and the support vector machine (SVM) technique, Liu et al. [38] categorised the periodicity-based audio and visual segmentation characteristics. Afterwards, they added to the framework [39] by including the visual characteristics and aural representations based on the bag-of-words (BoW) and energy envelope (EE).

Mel-frequency cepstral coefficient (MFCC) audio characteristics with skin colour and visual words were combined with MPEG motion vectors by Ulges et al. [23]. A weighted sum of late fusion is created by combining each feature representation after it has been processed by a separate SVM classifier. In order to identify adult material, Ochoa et al. [40] used sequential minimal optimisation (SMO) and LibSVM SVM algorithms to analyse the spatiotemporal information for binary video genre classification.

The one-dimensional signals of skin colour and spatiotemporal motion trajectory were used by Jung et al. [41]. A pornography

identification system called PornProbe was suggested by Tang et al. [42] and is based on a hierarchical latent Dirichlet allocation (LDA) and SVM algorithm. This method outperformed a single SVM classifier in terms of efficiency by combining supervised learning in SVM with unsupervised clustering in LDA. A multilevel hierarchical framework was provided by Lee et al. [43] by using the various aspects of various temporal domains. The bag-of-visual features (BoVF) was used by Lopes et al. [44] to identify obscenity.

In order to detect child-hazardous material and content uploaders, Kaushal et al. [21] used supervised learning. They did this by providing YouTube information at the video, user, and comment levels to machine learning classifiers (such as random forest, K-nearest neighbour, and decision tree). Reddy et al.'s [45] text categorisation of YouTube comments addressed the explicit content issue with videos. For the final classification, they used bigram collocation and input the features into the naïve Bayes classifier.

Disadvantages

- ANALYSIS OF PRE-TRAINED CNN MODEL VARIANTS is not supported by the current system.
- Analysis of efficient net features with distinct classifier variances is not possible with the current technology.

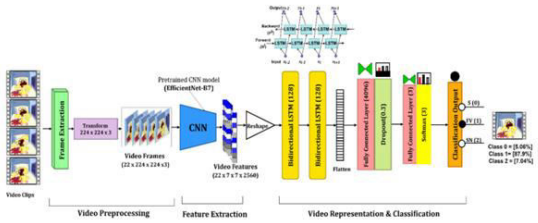
3. PROPOSED SYSTEM

1. The system suggests a unique CNN (EfficientNet-B7) and BiLSTM-based deep learning architecture for the identification and categorisation of improper video material.
2. A manually annotated ground truth video dataset including 1860 minutes (111,561 seconds) of animated cartoons for young children (under 13) is presented by the system. All videos are gathered from YouTube with the use of well-known cartoon characters as search terms. Every video clip has an annotation labelled as safe or hazardous for the lesson. Videos containing graphic sexual material and fantasy violence are categorised as risky. We also want to provide this dataset to the scientific community on a public basis.
3. The CNN-BiLSTM architecture that we have designed is assessed by the system for performance. The validation accuracy of our multiclass video classifier was 95.66%. To identify improper video material, a number of additional cutting-edge machine learning and deep learning architectures are assessed and contrasted.

Advantages

- Convolutional neural networks were used in the majority of image/video classification applications.
- The EfficientNet model is a convolutional neural network model and scaling technique that uses compound coefficient to scale the network depth, width, and resolution equally.

4. SYSTEM ARCHITECTURE



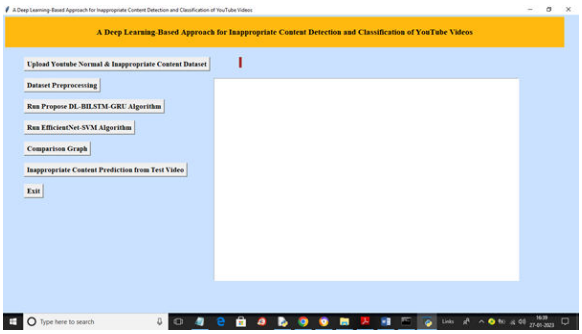
5. MODULES

To implement this project we have designed following modules

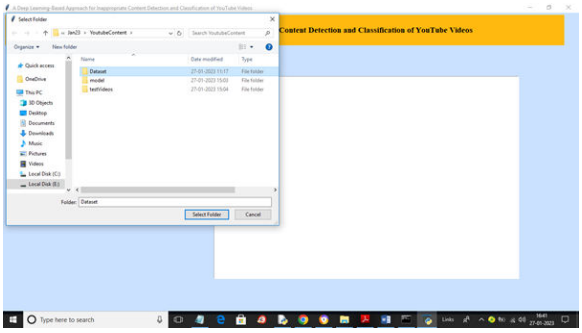
1) **Upload YouTube Normal & Inappropriate Content Dataset:** using this module we will upload YouTube dataset images to application
2) Dataset Preprocessing: using this module we will read all images and then resize all images to equal size and then normalize image pixel values and then shuffle the dataset
3) **Run Propose DL-BILSTM-GRU Algorithm:** using this module we will split dataset into train and test and then input 80% training data to Pre-Trained CNN (EfficientNetB7) algorithm to extract digital content from images and then those features will get retrained with BI-LSTM algorithm to train a model. Trained model will be applied on 20% test data to calculate prediction accuracy
4) **Run EfficientNet-SVM Algorithm:** EfficientNetB7 features will get retrained with existing SVM algorithm and then calculate prediction accuracy
5) **Comparison Graph:** using this module we will plot accuracy comparison graph between propose EfficientNetB7-BILSTM and EfficientNetB7-SVM.
6) **Inappropriate Content Prediction from Test Video:** using this module we will upload any YouTube and if video contains fighting or violence then application will predict as 'Inappropriate Content' otherwise will predict SAFE content.

6. SCREEN SHOTS

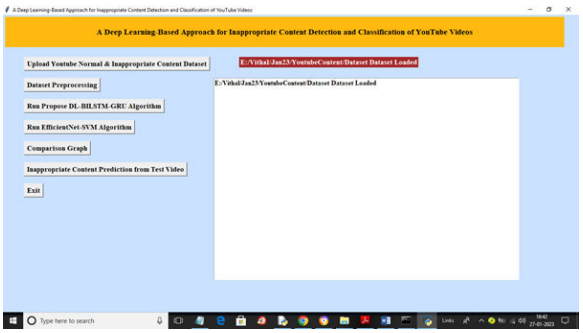
In above screen we have two folders and juts go inside any folder to view training images To run project double click on 'run.bat' file to get below screen



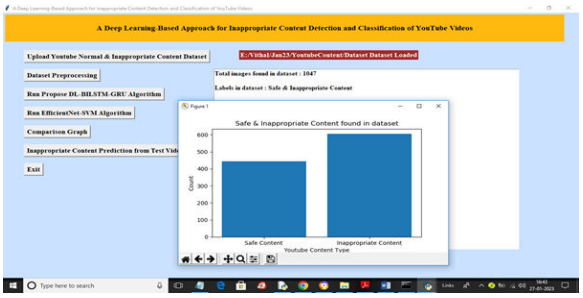
In above screen click on 'Upload YouTube Normal & Inappropriate Content Dataset' button to upload dataset and get below output



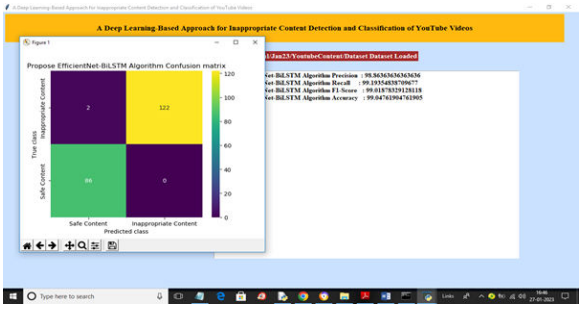
In above screen selecting and uploading entire 'Dataset' folder and then click on 'Select Folder' button to load dataset and get below output



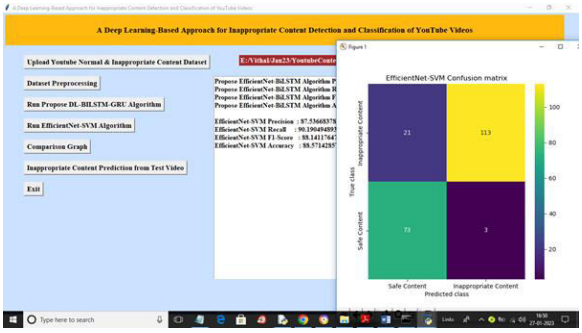
In above screen dataset loaded and now click on 'Dataset Preprocessing' button to read all images and then processes those images for training and get below output



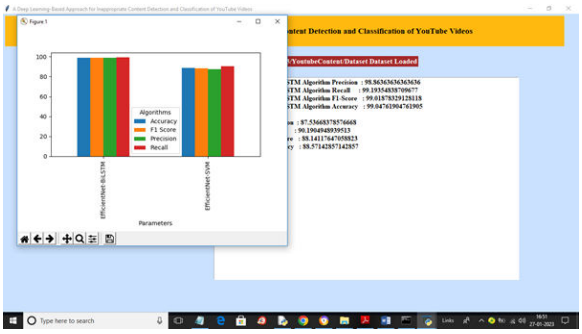
In above screen we can see dataset contains 1047 images and then in graph x-axis represents type of images such as 'Safe and Inappropriate' and y-axis represents count of those images. Now dataset processing completed and now click on 'Run Propose DL-BILSTM-GRU Algorithm' button to train propose algorithm and get below output



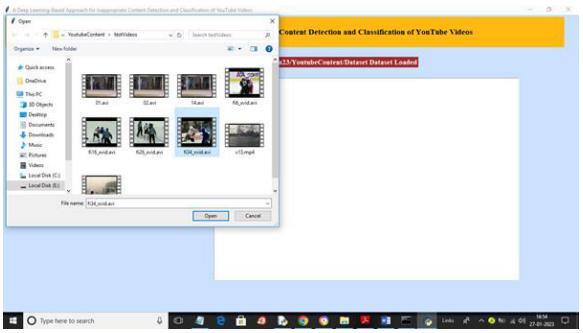
In above screen with propose EfficientNetB7-BI-LSTM we got 99.04% accuracy and in confusion matrix graph x-axis represents Predicted Labels and y-axis represents True Labels and green and yellow boxes contains correct prediction count and blue boxes contains incorrect prediction count which is 2 only. Now close above graph and then click on 'Run EfficientNet-SVM Algorithm' button to get below output



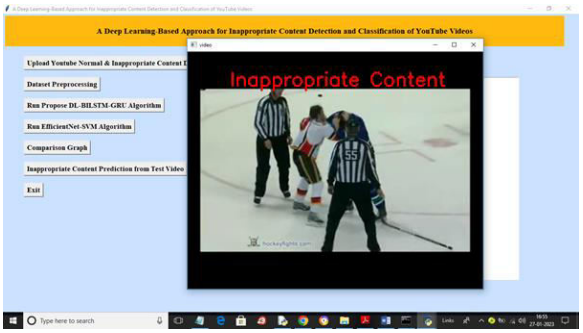
In above screen with EfficientNetB7-SVM we got 88% accuracy and in confusion matrix graph we can see in blue boxes that SVM predicted total 24 incorrect prediction so its accuracy is less. Now close above graph and then click on 'Comparison Graph' button to get below output



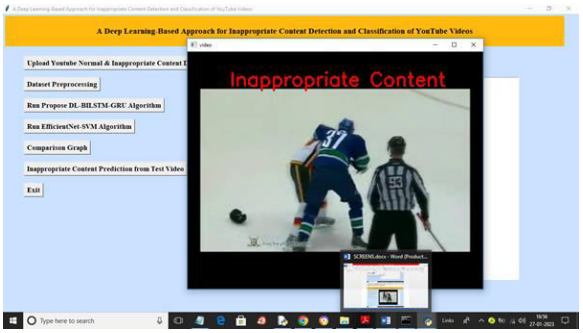
In above graph x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars. In both algorithms propose EfficientNetB7-BI-LSTM got high accuracy. Now close above graph and then click on 'Inappropriate Content Prediction from Test Video' button to upload test video and classify it as Safe or inappropriate.



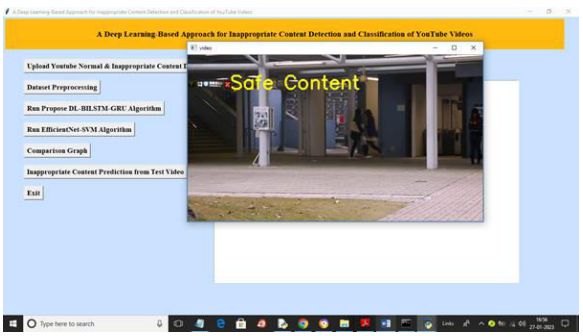
In above screen selecting and uploading video and then click on “Open’ button to play video and perform classification



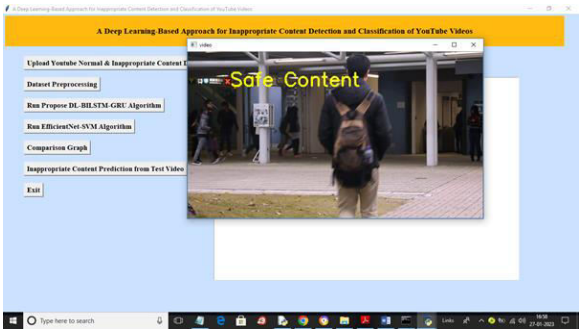
In above screen propose algorithm evaluating playing video and then detecting and classifying it as ‘Inappropriate Content’



In above video also we got classification output



In above we got result as Safe Content



In above screen we got output as Safe Content as peoples are only moving in the video.

7. CONCLUSION

A brand-new deep learning-based framework is suggested for the identification and categorisation of offensive video material for children. The efficientnet-b7 architecture is used in transfer learning to extract video features.

The bilstm network processes the retrieved video features, allowing the model to learn effective video representations and perform multiclass video classification. A carefully annotated cartoon video dataset of 111,156 video clips gathered from YouTube is used for all assessment trials. The evaluation results showed that the efficient net bilstm framework that was proposed (with hidden units = 128) performs better (accuracy = 95.66%) than the attention mechanism-based efficient net-fc, efficient net-svm, efficient net-knn, and efficient net-random forest models that were also experimented with (with hidden units = 64, 128, 256, and 512). Furthermore, by attaining the maximum recall score of

92.22%, the performance comparison with current state-of-the-art models also showed that our bilstm-based framework outperformed other current models and methodologies. The following are the benefits of the suggested deep learning-based children's unsuitable video content identification system:

1) It functions by filtering the live-captured videos by utilising an efficientnet-b7 and bilstm-based deep learning framework, processing the video at a speed of 22 frames per second while taking into account the real-time circumstances.

2) It can help any video sharing site either blur/hide any part with disturbing frames or eliminate the film containing dangerous clips.

3) It may also aid in the creation of online parental control programs that filter harmful information for children automatically using plugins or browser extensions.

Moreover, our approach to identifying objectionable material for children on YouTube is not reliant on the metadata of the videos, which may be readily changed by malevolent uploaders to trick viewers. In order to enhance the model's performance by comprehending the global representations of films, we want to integrate the spatial stream of RGB frames with the temporal stream derived from optical flow frames in the future. Additionally, we want to target the many forms of unsuitable children's material on YouTube videos by increasing the number of categorisation labels.

REFERENCES

[1] L. Ceci. YouTube Usage Penetration in the United States 2020, by Age Group. Accessed: Nov. 1, 2021. [Online]. Available: https://www.statista.com/statistics/296227/us-youtube-reach-age-gender/

[2] P. Covington, J. Adams, and E. Sargin, ‘‘Deep neural networks for YouTube recommendations,’’ in Proc. 10th ACM Conf. Recommender Syst., Sep. 2016, pp. 191–198, doi: 10.1145/2959100.2959190.

[3] M. M. Neumann and C. Herodotou, ‘‘Evaluating YouTube videos for young children,’’ Educ. Inf. Technol., vol. 25, no. 5, pp. 4459–4475, Sep. 2020, doi: 10.1007/s10639-020-10183-7. [4] J. Marsh, L. Law, J. Lahmar, D. Yamada-Rice, B. Parry, and F. Scott, Social Media, Television and Children. Sheffield, U.K.: Univ. Sheffield, 2019. [Online]. Available: https://www.stac-study.org/downloads/ STAC_Full_Report.pdf

[5] L. Ceci. YouTube—Statistics & Facts. Accessed: Sep. 01, 2021. [Online]. Available: https://www.statista.com/topics/2019/youtube/ [6] M. M. Neumann and C. Herodotou, ‘‘Young children and YouTube: A global phenomenon,’’ Childhood Educ., vol. 96, no. 4, pp. 72–77, Jul. 2020, doi: 10.1080/00094056.2020.1796459.

[7] S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson, Risks and Safety on the Internet: The Perspective

of European Children: Full Findings and Policy Implications From the EU Kids Online Survey of 9-16 Year Olds and Their Parents in 25 Countries. London, U.K.: EU Kids Online, 2011. [Online]. Available: http://eprints.lse.ac.U.K./id/eprint/33731

[8] B. J. Bushman and L. R. Huesmann, ''Short-term and long-term effects of violent media on aggression in children and adults,'' Arch. Pediatrics Adolescent Med., vol. 160, no. 4, pp. 348–352, 2006, doi: 10.1001/archpedi.160.4.348.

[9] S. Maheshwari. (2017). On YouTube Kids, Startling Videos Slip Past Filters. The New York Times. [Online]. Available: https://www.nytimes.com/2017/11/04/business/media/youtube-kids-paw-patrol.html

[10] C. Hou, X. Wu, and G. Wang, ''End-to-end bloody video recognition by audio-visual feature fusion,'' in Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV), 2018, pp. 501–510, doi: 10.1007/978-3-030-03398- 9_43.

[11] A. Ali and N. Senan, ''Violence video classification performance using deep neural networks,'' in Proc. Int. Conf. Soft Comput. Data Mining, 2018, pp. 225–233, doi: 10.1007/978-3-319-72550-5_22.

[12] H.-E. Lee, T. Ermakova, V. Ververis, and B. Fabian, ''Detecting child sexual abuse material: A comprehensive survey,'' Forensic Sci. Int., Digit. Invest., vol. 34, Sep. 2020, Art. no. 301022, doi: 10.1016/j.fsidi.2020.301022.

[13] R. Brandom. (2017). Inside Elsagate, The Conspiracy Fueled War on Creepy YouTube Kids Videos. [Online]. Available: https://www.theverge.com/2017/12/8/16751206/elsagate-youtube-kids-creepy-conspiracytheory

[14] Reddit. What is ElsaGate? Accessed: Dec. 14, 2020. [Online]. Available: https://www.reddit.com/r/ElsaGate/comments/6o6baf/

[15] B. Burroughs, ''YouTube kids: The app economy and mobile parenting,'' Soc. media+ Soc., vol. 3, May 2017, Art. no. 2056305117707189, doi: 10.1177/2056305117707189.