applied sciences

MDPI

Review

Children's Safety on YouTube: A Systematic Review

Saeed Ibrahim Alqahtani, Wael M. S. Yafooz *, Abdullah Alsaeedi, Liyakathunisa Syed and Reyadh Alluhaibi

Department of Computer Science, College of Computer Science and Engineering, Taibah University, Medina 42353, Saudi Arabia
* Correspondence: wyafooz@taibahu.edu.sa

Abstract: *Background*: With digital transformation and growing social media usage, kids spend considerable time on the web, especially watching videos on YouTube. YouTube is a source of education and entertainment media that has a significant impact on the skill improvement, knowledge, and attitudes of children. Simultaneously, harmful and inappropriate video content has a negative impact. Recently, researchers have given much attention to these issues, which are considered important for individuals and society. The proposed methods and approaches are to limit or prevent such threats that may negatively influence kids. These can be categorized into five main directions. They are video rating, parental control applications, analysis meta-data of videos, video or audio content, and analysis of user accounts. *Objective*: The purpose of this study is to conduct a systematic review of the existing methods, techniques, tools, and approaches that are used to protect kids and prevent them from accessing inappropriate content on YouTube videos. *Methods*: This study conducts a systematic review of research papers that were published between January 2016 and December 2022 in international journals and international conferences, especially in IEEE Xplore Digital Library, ACM Digital Library, Web of Science, Google Scholar, Springer database, and ScienceDirect database. *Results*: The total number of collected articles was 435. The selection and filtration process reduced this to 72 research articles that were appropriate and related to the objective. In addition, the outcome answers three main identified research questions. *Significance*: This can be beneficial to data mining, cybersecurity researchers, and peoples' concerns about children's cybersecurity and safety.

Keywords: YouTube; kids' digital safety; social networks; inappropriate content; machine learning; deep learning

Citation: Alqahtani, S.I.; Yafooz, W.M.S.; Alsaeedi, A.; Syed, L.; Alluhaibi, R. Children's Safety on YouTube: A Systematic Review. *Appl. Sci.* **2023**, *13*, 4044. https://doi.org/10.3390/app13064044

Academic Editor: Luis Javier García Villalba

Received: 18 February 2023
Revised: 14 March 2023
Accepted: 16 March 2023
Published: 22 March 2023

1. Introduction

Traditionally, kids around the world like to spend hours watching TV series, cartoons, movies, films, DVDs, or recorded films on computers that can be educative or entertaining. In the early stages of childhood, kids gain and develop skills and knowledge that help them build their personalities over time. Furthermore, spending time watching videos aids in the growth and development of such skills. Moreover, traditional ways are easy to control by the government or parents based on kids-friendly program regulations. In this way, the risk and influence on kids' learning and development become limited [1–5].

In the past decade, the development of web and social media, data, video streaming, and the availability of smartphones, laptops, and tablets have led to the expansion of internet usage. This is a result of increased activities using high-speed internet, especially among children and adolescents who spend more time on Social Media Networks (SMN) or surfing the internet. SMNs are among the most desirable and popular sites and platforms for children. They use it for playing games, having one-on-one conversations, sharing information, and watching movies [6–8]. Thus, there are many benefits to kids and teenagers, such as gathering knowledge or enhancing skills, especially during the COVID-19 pandemic, when children spent much time on SMN for distance education platforms or entertainment. However, such platforms without regulation allow children

to access inappropriate content that may be psychologically, intellectually, or emotionally harmful and unsafe [9].

YouTube is one of the most popular social media platforms and was created in 2005. Millions of users who upload, watch, share, and comment on videos use it and display recorded educational videos or stories [10–12]. Kids and teenagers mostly use it as an alternative to traditional TV or entertainment methods. In this regard, digital spaces are consuming large amounts of time for kids. Kids can access a diverse range of YouTube channels through many types of devices, such as tablets, smart TVs, mobile phones, laptops, and desktops [13]. Because many YouTube channels are created to attract kids, the relationship between children and digital media is swiftly growing. According to Statista [14], in 2022, 2.6 billion viewers were on YouTube. Monthly, 70% of this figure access YouTube through mobile phones. Based on recent research in 2022, 65% of kids spend their browsing time on YouTube. Eighty percent of parents in the USA have stated that their kids watch YouTube. It can positively or negatively impact the learning of children at an early age. YouTubers/influencers and companies create hundreds of YouTube channels that target kids. These channels can be beneficial for kids if used for educational or entertainment purposes. At the same time, it can be risky if these channels contain disturbing or inappropriate content for kids that can impact their knowledge, attitude, and foundational and development skills [15]. Therefore, the Children's Online Privacy Protection Act (COPPA) highlighted the requirement to protect children under 13 years for the online community of kids' digital safety. In addition, for that purpose, in 2015, Google introduced YouTube Kids [16]. It can be used to filter YouTube videos based on the appropriateness of kids' age through the analysis of meta-data or suggestions from user reports. On YouTube Kids, children can watch videos from trusted, family-friendly channels, and parents can trust them without any supervision. Additionally, on YouTube Kids, the levels of restrictions are classified as: preschoolers starting from age 2 to 4, young children ages 5 to 8, and older children aged 8 to 12 years. To further explore this matter, based on the statistics for 2022, the top channel watched by kids is ChuChu TV Nursery Rhymes and Kids Songs, which has 57.5 million young subscribers.

However, YouTube Kids is still not a strong platform that can prevent inappropriate or disturbing videos to kids [17–23] due to the increasing number of YouTube channels/videos or user-generated comments. In addition, the decision algorithms rely on the meta-data of videos that are uploaded by YouTubers. Additionally, there are many sites or advertisements not suitable for children that are in the form of visual, audio, and textual content, or through URL links that can be added to YouTube videos through user-generated comments. Such content may appear suddenly while children or adolescents browse educational videos or accidentally click on them. These videos or links contain embedded messages that can negatively impact children. Unfortunately, children will not know that these videos are harmful to them. In this way, regardless of the advantages of the internet, online activities, or YouTube videos, they become a grave danger to a child's mentality and behavior. They are also considered risky to childhood development as health problems such as aggressive thoughts and feelings affect kids' growth. This is due to some of the videos having illegal and inappropriate content, such as child abuse, violence, gambling, horror sounds, scary scenes, and abusive language. Such content is believed to be unsafe for children if no control or monitoring is in place [24].

Therefore, in the last few years, Machine Learning (ML) [25–32], Deep Learning (DL) [15,33–44], and Natural Language Processing (NLP) [45–55] researchers and scientists have contributed to proposing methods, approaches, techniques, and tools to ensure the safety of children on social media, particularly on online videos. Hence, this study aims to conduct a Systematic Review (SR) to provide an overview of the existing methods, tools, techniques, and approaches that were proposed or developed scholarly to protect, prevent, and save kids from inappropriate or illegal video content. The main key stages of conducting a systematic review are followed by the formulation of research questions, identifying inclusion and exclusion criteria, identifying the source materials (electronic

databases), study section, results, and summarization. This study answers three Research Questions (RQs). They are:

RQ1: What are the issues and threats that kids encounter on YouTube videos?
RQ2: What are the existing methods, techniques, and approaches used to protect and prevent inappropriate YouTube video content?
RQ3: What are the challenges in protecting kids from inappropriate content on YouTube videos?

To answer the RQs of this SR, the inclusion and exclusion criteria are identified as follows: (1) Retrieved research papers and articles that are published in IEEE Xplore Digital Library, ACM Digital Library, Springer database, Web of Science, Google Scholar, and ScienceDirect electronic databases. (2) Research papers and articles are written in the English language only. (3) The period of paper publication is between January 2016 and December 2022. (4) Research papers and articles that are related to web videos and YouTube videos from a cybersecurity perspective. (5) Papers that were published in conferences and articles published in indexed databases. (6) Articles and papers that are relevant to machine learning, deep learning, and cybersecurity, in general, which are relevant to computing or information technology research areas. (7) Books, theses, and notes are excluded in this SR.

The rest of this article is organized as follows: Section 2 highlights the most important definitions used in the SR, the impact of the YouTube video on kids, and the research area. Section 3 presents the research methodology that is used in this research. The summarization of the existing approaches is explained in Section 4. The result and discussion of this SR are demonstrated in Section 5. The conclusion of this article is included in Section 6.

2. Background

This section explains the main terminologies and definitions that are used in this SR and relevant to the research area. Further, it illustrates the impact of YouTube videos on kids.

2.1. Terminologies and Definitions

These are some of the terminologies used in this SR. These terms are defined below:

- **Social media networks:**

Social media networks are a platform in which users can communicate together or seek knowledge through the desired network on a daily basis.

- **Inappropriate content for kids:**

Content that is not suitable for younger audiences (kids) and inappropriate content which is harmful to children are classified as inappropriate content, such as harassment, nudity, violence, disturbing, terror, racist, and abusive language.

- **Kids' digital safety:**

Digitally protecting children from inappropriate/harmful content online.

- **YouTube platform:**

A platform that provides video content to watch online. It consists of a broad spectrum of content and is available for all ages and audiences throughout the internet.

- **YouTube Kids:**

A platform that produces content specifically for children to protect them from inappropriate online content.

- **Kids:**

"Kids" is the name of the age group that ranges from 2 to 12 years.

- **Teenagers:**

"Teenagers" is the name of the age group that ranges from 13 to 19 years.

2.2. Impact of YouTube Videos on Kids

YouTube has become a go-to source of content for children and their parents because it provides an infinite stream of videos that cater to user tastes regardless of age. With new competitors appearing daily and using repetition, familiarity, and long strings of keywords to place their videos in front of kids who are willing to spend hours clicking on whichever thumbnails spark their attention, YouTube is leading the market for kid-oriented videos [56]. All children adore YouTube, whether it is for amusement, free academic assistance, or tutorial videos. YouTube has established a foothold in children's lives. However, YouTube safety for children is a complex subject. Kids' ability to browse videos on YouTube has great advantages, but it also carries some danger of harm. The positive and negative impacts of YouTube videos are as follows.

Pros:

1. **Helps in upskilling**: With the help of videos, you may learn to cook, do crafts, dance, and even fix broken things. It is an excellent method to develop your skills and to make use of your time effectively [57].
2. **Fostering Creativity**: Kids who are exploring the YouTube platform come across a wide variety of content. YouTube provides users with a variety of solutions to the same problems whenever they encounter them, fostering their creativity and broadening their viewpoints [58].
3. **An Academic Resource**: YouTube is a fantastic resource for learning life skills, but it is also useful for obtaining academic assistance. When your kid is struggling with their homework, you can aid them by pointing them toward YouTube. However, be sure to also familiarize them with the procedure for determining the credibility of the sources [57].

Cons:

1. **Exposure to Inappropriate Content**: Children and teenagers have infantile minds because they have not reached adulthood. Being exposed to things that are inappropriate for their age can have a negative impact on their way of thinking from a young age. There is no monitoring of content on YouTube channels [56,57].
2. **Consumption of Wrong Information**: Young children are subjected to a plethora of videos for entertainment and as a tool for learning new things. However, there is no guarantee that the information supplied on these channels is correct. Any misinformation can lead to problems in the future [56,57].
3. **Creating False Projection**: Even though we know the internet is a scam, we continue to fall for it. The same thing happens with children who are drawn to YouTubers' pompous lifestyles. This results in discontent and the formation of fake promises in children [57].

3. Methods

This section presents the research methodology used to carry out a systematic review of methods and approaches that are used to protect kids from inappropriate YouTube content. It contains four main phases. They are the strategy of search, study selection, data extraction, paper selection, and results, as shown in Figure 1.

3.1. Strategy of the Search Phase

This was the initial phase in the systematic review indicating the research papers that are related to the computing research area in protecting kids from illegal or inappropriate content on social media, especially through cybersecurity for kids and safer YouTube videos. Some terms/keywords were identified: social media, YouTube, inappropriate content, online safety, video content, YouTube, digital safety, online behavior, video analysis, audio analysis, content analysis, child safety, social network security, cybersecurity, safe contains, kids, child, children, parents, and early childhood. Such terms were used to review the research articles and narrow the scope of research to make it easier to address research

questions 1 and 2 and then select the relevant articles. In this stage, the pre-print research has been excused, and all the SR inclusion and exclusion criteria were applied.

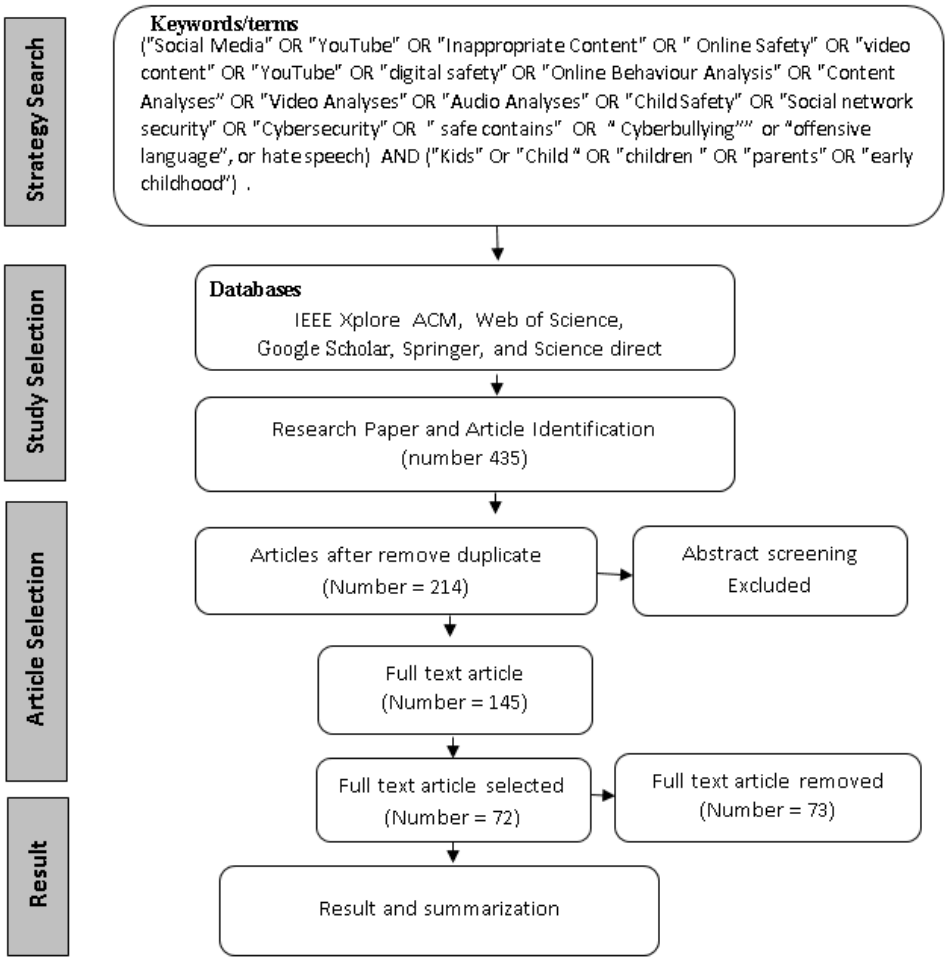

Figure 1. SR methodology framework.

3.2. Study Selection Phase

The source of databases was selected based on the relatedness of computer science and information technology areas, namely, IEEE Xplore Digital Library (https://ieeexplore.ieee.org/ (accessed on 15 March 2023)), ACM Digital Library (https://dl.acm.org/ (accessed on 15 March 2023)), Springer database, Web of Science (https://www.webofknowledge.com/ (accessed on 15 March 2023)), Google Scholar (https://scholar.google.es/ (accessed on 15 March 2023)), and ScienceDirect (https://www.sciencedirect.com/ (accessed on 15 March 2023)) database. The systematic search focuses on the research papers and articles published in indexed conferences and journals from January 2016 to December 2022. Pre-print papers were excluded, and the inclusion and exclusion criteria were applied. Table 1 and Figure 2 illustrate the number of research articles and the database sources, while Figure 3 shows research articles per country.

Table 1. Research articles and databases.

Databases	Found	Candidate	Selected
Google Scholar	144	42	24
IEEE Xplore	63	22	10
ACM	66	32	15
Springer	76	21	11
ScienceDirect	86	28	12

Figure 2. Number of research articles from database sources.

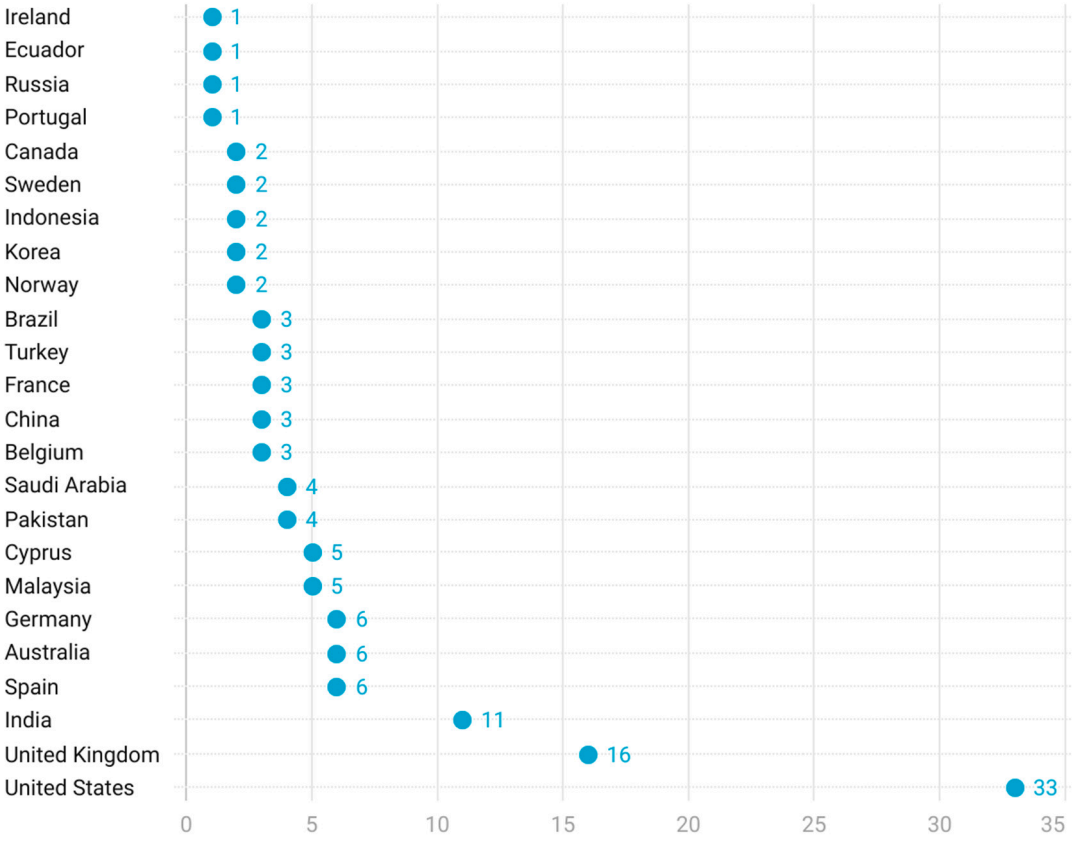

Figure 3. Research articles per country.

3.3. Data Extraction and Paper Selection Phase

This stage included three main stages: retrieved articles, initial selection, and selected articles. The articles retrieved and downloaded were the research papers from the mentioned databases without any pre-processing methods. The initial section in the data extraction process began to filter the papers and select the most relevant ones to the scope of this systematic review based on the inclusion and exclusion criteria (i.e., the title, abstract, and keywords were read). The abstract is an important part of a paper that provides a summary of the research, which helped us select the related and relevant research articles. Additionally, in this stage, we skimmed the research papers in order to make accurate decisions to include or exclude the research articles. Thus, some of the research articles were excluded if they did not meet the inclusion and exclusion criteria addressed in the SR.

In addition, duplicate and non-relevant articles were removed. In the third stage, the full text of the paper was retrieved to identify the methods, techniques, approaches, and tools. In this stage, all the discussion between the team members was conducted to remove any confusion or ambiguity in identifying the research articles. In addition, the initial idea of how to categorize the research articles based on the existing approaches used scholarly was identified. This categorization was used in the next phase.

3.4. Results and Summarization Phase

This was the final phase in the SR. The process of summarization and categorization of the existing methods was executed in this stage. The total number of relevant research articles was 72. These articles were categorized into four main approaches: textual data analysis, meta-data analysis, video frame analysis, and audio-based analysis.

Then, the research articles were summarized based on seven main aspects. First, was the paper's main objective. Second, the focus given to the approaches utilized: machine or deep learning or content analysis focusing on statistics. Third, whether the researchers retrieved, collected, and constructed their dataset from YouTube videos using YouTube API or other used publicly available datasets. Fourth, this SR focused on the language utilized in applying the experiment using it. Fifth was the type of methods or models that were used in the textual data representation or meta-data collection/extraction frame analysis or audio conversion. Sixth was the accuracy of the model performance. At the end of each approach, a critical analysis was also highlighted, and then the future perspective was outlined.

4. Approaches and Existing Methods

This section demonstrates the findings of the SR and existing methods that are proposed to detect or prevent inappropriate, illegal content on YouTube Kids. These methods can be categorized into four methods: textual data analysis, meta-data analysis, video frame analysis, and audio-based analysis, as shown in Figure 4.

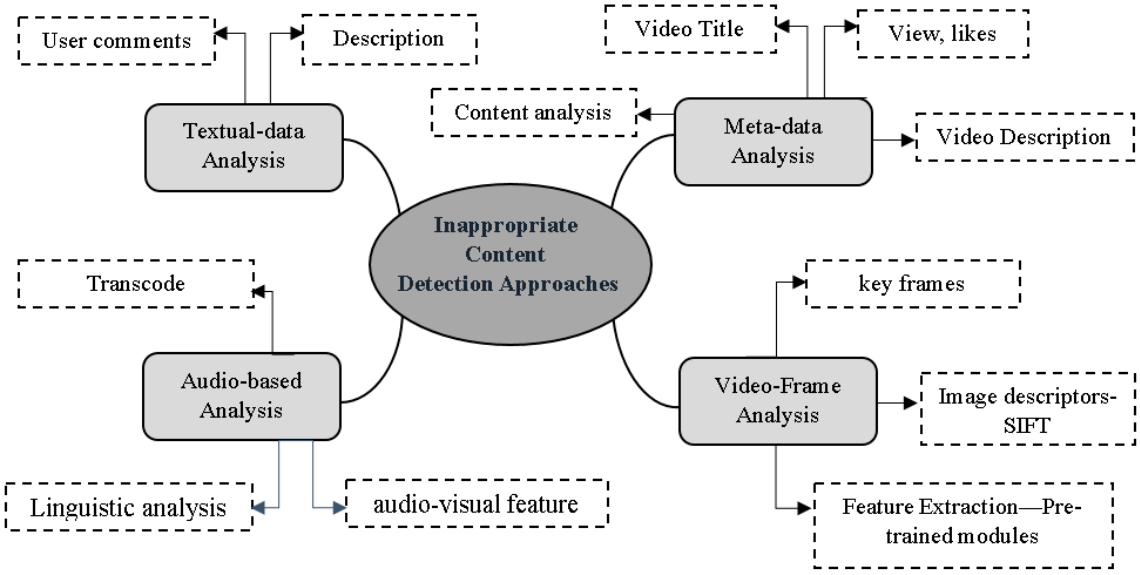

Figure 4. An overview of existing methods for preventing inappropriate content on YouTube video content.

All the aforementioned methods are carried out through two main approaches: AI applied techniques—such as machine learning, deep learning, transfer learning, and computer vision—and content analysis with a statistics approach. In the first approach, the common phases that are used can be generalized as data preparation, data cleaning, annotation process, data pre-processing, feature extraction and representations, building a machine deep learning model, model evaluation, and model performance analysis, as shown in Figure 5. The second approach is based on content analysis and statistics that can be collected from the meta-data of the YouTube videos.

In the data preparation phase, the data can be collected from targeted YouTube channels using the YouTube API or datasets that are available to the public. The form of the data can be textual, meta-data of videos, audio, and image frames from YouTube videos. The output of this phase is an initial dataset that is used in the next phase. In the second phase, which is also known as data cleaning, if the data are extracted from YouTube channels, it is

required to apply cleaning techniques such as noise or not-required data. After this phase, the annotation process is required to be performed in order to label the data into binary classes or multi-classes. The commonly used classes are "appropriate", "inappropriate", "suitable", and "distributing", or 0 or 1. The output of this phase is a dataset that is used as input for the machine/deep learning models in the next phase. In machine learning, there are comment classifiers such as Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), k-Nearest Neighbors (KNN), and Support Vector Machine (SVM). While in deep learning, there are many proposed architectures, such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Bidirectional Long Short Term Memory (BiLSTM), or hybrid models of two or more as well.

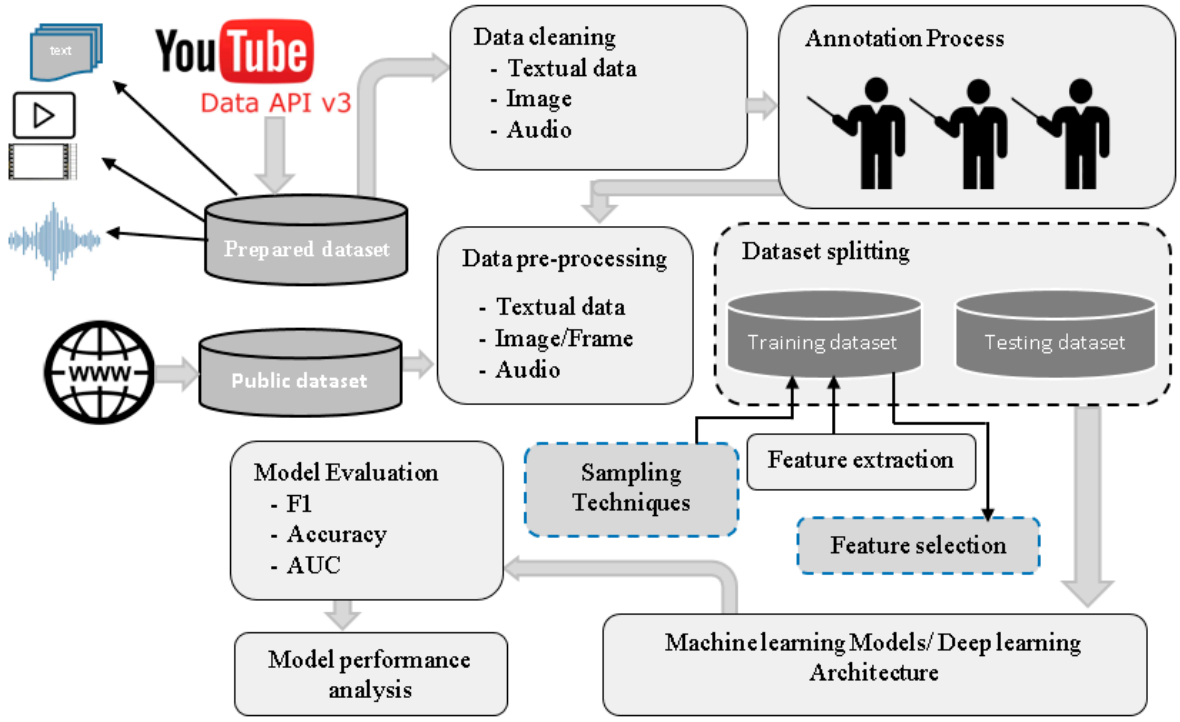

Figure 5. A general view of methods used for detecting inappropriate YouTube Kids videos using machine/deep learning approaches.

Before using the model, there are features such as extraction and representation phases. In this phase, there are several representation methods such as Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), word embedding, Word2Vec, GloVec, FastText, and Bert. The output of this phase is used as input for the model. In addition to the pre-processing steps, methods such as normalization, stemming, lemmatization, and case folding can be used to clean the text. In building a machine learning model, the dataset is divided into two parts—one for data training to train the model and one for data testing to test the model's performance in detecting inappropriate content. Before the final phase, called model evaluation, several metrics can be used to evaluate the model, such as accuracy, area under the curve (AUC), and F1-score. The final phase is the performance analysis, which is used to compare the results and the discussion part.

4.1. Textual Analysis Approach (TA)

In this approach, the researcher attempts to detect videos that contain inappropriate and abusive content. This can be performed by collecting the user comments or converting the video manuscript to textual data through the transcribe method. Then, a dataset is built with annotators to classify the user comments, or a publicly available dataset is used. Then, several NLP techniques are utilized to analyze the contents and extract features of

video from user-generated comments. Generally, uploaders or commenters publish this content. Therefore, the existing methods focused on detecting, identifying, and tracking unsafe videos or user comments that may impact users through analysis of the textual information can be found in the user comments, as shown in Figure 6, video description, or through analysis of the transcribed manuscript. In addition, researchers employ NLP approaches such as sentiment analysis to detect inappropriate content in videos based on user comments by extracting features. All the methods used in this approach rely on machine learning, deep learning approaches, and transfer learning, and they utilize natural language processing techniques.

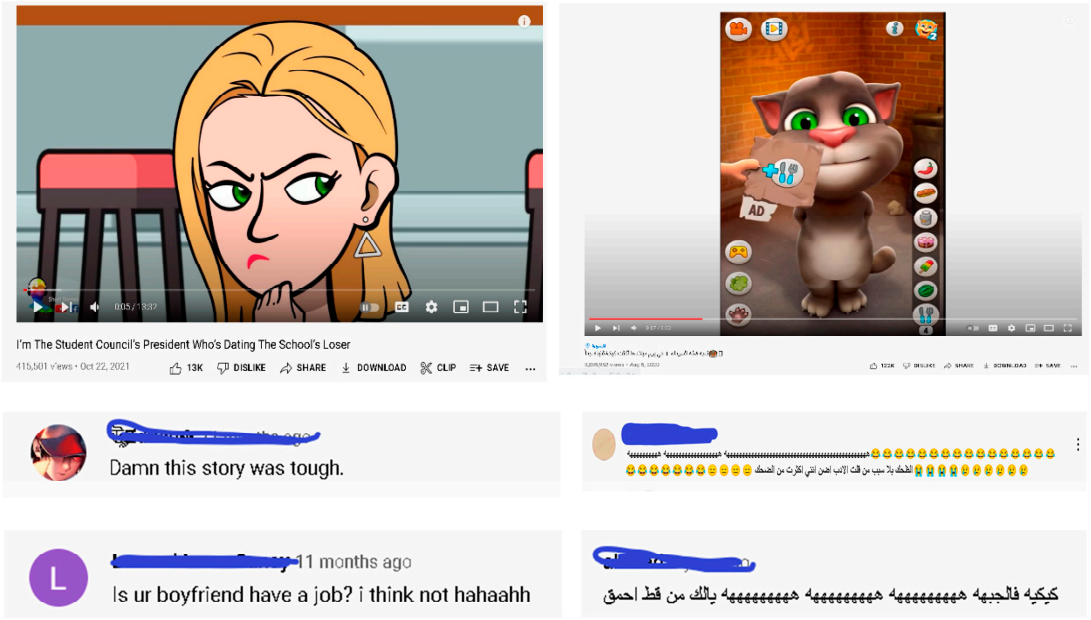

Figure 6. Kid's video with user comments.

The machine learning approach is considered the traditional method, and early research works were developed using ML classifiers. It uses inappropriate content on YouTube videos as a supervised machine-learning classification problem. Therefore, many classifiers such as NB, LR, DT, RF, KNN, and SVM were used. In addition, in all the existing methods described in the literature, the NLP techniques are also used as pre-processing steps, which helps in achieving high accuracy in identifying and detecting inappropriate text. Ashraf et al. [25] introduced a new dataset called CAALDYC for detecting abusive language on YouTube from user-generated comments based on the English language. This dataset focuses on labels of abusive language on politics, religion, and others that were annotated into abusive or non-abusive. These data were collected from 19 YouTube videos from BBC, CNN, or TV, with approximately 2304 user comments and 6139 replies. Then, they applied eight machine learning and deep learning classifiers using *n*-grams and GloVec as feature representation. After carrying out several experiments, the best accuracy, 91.96%, was achieved in terms of F1-score when using the Adaboost classifier with *n*-gram features on the dataset with replies. For the comment dataset, the model achieved 87.87% and 89.86 in terms of accuracy using DT with *n*-gram and 1D-CNN + MP, respectively. However, such a dataset is considered small.

Similarly, to detect and prevent inappropriate content on YouTube, Hani et al. [26] used machine learning classifiers SVM and a neural network that employs the NLP method, including n-grams as pre-processing steps. They use TFIDF as feature representation. An experiment based on a publicly available dataset from the Kaggle website, which contains 12,773 comments collected from conversations and messages, while the experiments were carried out on only 1608 as a balanced dataset for both classes was applied. Both models'

accuracy reached 89.87% and 91.76% using SVM and neural networks, respectively. In the same way, Alcântara et al. [27] used machine and deep learning in Portuguese and transcoding methods to convert the videos into textual data. Then, these data were annotated into two classes, namely offensive and non- offensive, using 400 videos. Machine learning, deep learning, and transfer learning classifiers were applied. In addition to improving model performance, they used different representation techniques such as *n*-grams, GloVec, Word2Vec, and FastText. n-grams achieved the best results, and the F1-score reached 74%. However, the size of the dataset is smaller and is not used for deep learning approaches. In this approach, the model required more training to achieve higher accuracy.

In the same way as other researchers, they utilized machine learning classifiers using a publicly available Arabic dataset collected from a YouTube video. Alsubait and Alfageh [28] used three classifiers: multinomial NB, complement NB, and LR classifiers, to compare their performance in automatically detecting cyberbullying in two different feature representations using TFIDF and count vectorizer. The results show that LR outperforms the other two classifiers in vectorizer, which achieved 78.6% in F1-score, while complement NB recorded 78.5%. Mouheb et al. [29] used 25,000 user comments from YouTube and Twitter and applied an NB classifier after several NLP pre-processing steps. The NB recorded 95% in terms of accuracy in detecting inappropriate information. The Bengali language is used in [30,31], and Ahmed et al. [30] utilized classifiers SVM, LR, NB, and ensemble methods (XGBoost). The dataset is based on the Bengali language collected from YouTube user comments and other sources. They built three datasets—the first dataset consisted of 5000 Bengali texts, the second dataset consisted of 7000 randomized texts, and a combination of the two sets consisted of a total of 12,000 user comments. These were divided into two classes: bullying and non-bullying based on the three datasets. Before, experiments were carried out with several NLP tasks, which were performed to prepare the dataset for ML and ensemble classifiers on three datasets. The model achieved an accuracy of 76%, 84%, and 80% using the first dataset with SVM, the second dataset with NB, and the third dataset with NB. Awal et al. [31] applied NB in the Bengali language that was collected from YouTube and translated from English language. The dataset had 2665 user comments in the form of two classes: abusive and non-abusive. Ensemble methods [32] are used to detect cyberbullying based on personality traits through collected user comments from YouTube and applied the classifiers, which achieved 95% model accuracy.

All the aforementioned research is used to detect inappropriate content based on two classes (binary classes), and there are research works made attentively by studying the same issue from different perspectives. Therefore, the authors in [4,33–35] proposed a variety of methods to detect inappropriate content in user comments by the affected range of kids and examine URLs in user comments. The authors collected 3.7 million comments posted on children's videos. Then, in [33], the user comments were divided into five groups based on age. They found that the group aged 13–17 was the most exposed to inappropriate comments, followed by the 6–8 age group. In [4], Alshamrani et al. studied the presence of inappropriate and malicious content in comments of YouTube videos targeting kids and teenagers by investigating the existence of malicious and inappropriate URLs found inside the comments posted. Among these comments, 8677 URLs were extracted using a regular expression. Moreover, 107 various topic categories associated with these URLs were extracted. The top ten categories included "Entertainment", "Television", "Streaming Media", "Society", "Social Networking", "News/Weather", "Technology", "Music", "Hobbies", and "File Sharing". The authors stated that the URLs in comments in videos targeting kids are very likely to be blindly clicked. Finally, the authors checked the validity of a given URL by accessing them and investigating the response of HTML requests.

Alshamrani et al. [35] studied the detection of toxic behaviors presented in YouTube video comments and new video captions. They collected more than 7.3 million comments posted in news videos and 10,000 captions on those videos. Two datasets were utilized for training the ensemble classifier: Wikipedia Ground Truth and YouTube Ground Truth datasets. The former dataset consists of 160,000 comments manually annotated, with

15,294 labeled as toxic, 143,000 safe, 1405 hate, and 8449 obscene. The latter dataset was manually annotated, which consisted of 5958 YouTube comments, with 1832 labeled as safe, 788 identified as hate, 2367 obscene, and 4126 toxic. The pre-processing of comments includes transforming words to numerical vectors using the pre-trained Word2Vec. The main objective of this study was to detect and classify comments in terms of toxic behaviors. In this way, a deep neural network ensemble model was used to classify comments into three categories: obscene, toxic, and identity hate. Additionally, the association between toxic behavior and extracted topics from their captions were identified with the topic modeling using LDA. The authors examined the relationship between various toxic behaviors and the content of news videos across 15 topics, demonstrating that the comments related to religion, violence, and crime topics were the highest rated of toxic behavior, while comments related to the economy had the lowest rate of toxic behavior. Differently, AlHarbi et al. [36] employed entropy, chi-square, and PMI to detect cyberbullying based on collected user comments from YouTube and Twitter by using approximately 100,327 user comments. They were classified into two classes: bullying and non-bullying, using three annotators. The results show that PMI outperforms entropy and chi-square.

For the last few decades, this approach has been used by many researchers due to its accuracy in detecting and identifying inappropriate content when using the deep learning approach such as CNN, RNN, LSTM, BiLSTM, or hybrid models of two or more. Sometimes, kids search for YouTube videos, and due to spelling mistakes, inappropriate videos appear to them. On the other hand, Yenala et al. [37] examine complete sentences in a search on YouTube, or any other social media platform, as sometimes obscene words are automatically suggested to complete sentences. Additionally, the same concept is used in user conversion when the application autocompletes sentences. They propose a novel method called Convolutional Bi-Directional LSTM (C-BiLSTM) using a deep learning approach that can be used in social media networks such as YouTube to filter inappropriate content during the search process for videos. They introduced a novel dataset built from 79,041 user comments. This is divided into two classes, namely clean and inappropriate. The hybrid deep learning approaches are applied using CNN and LSTM with NLP methods such as *n*-grams based on characters as input. They use BLSTM due to the length of comments, approximately 200 words, considering the misspelling and performance of a check based on the characters of the words. The model achieved 87% in terms of accuracy using C-BiLSTM. Cunha et al. [38] utilize different deep learning architectures using CNN to detect cyberbullying. This applies the sentiment assist concept based on YouTube user comments to show the quality of the YouTube videos. Two videos with around 1000 user comments are considered, and the user comments are annotated into three classes. The best model performance that has been achieved is 84% in terms of accuracy.

The authors in [39,40] highlighted the importance of analyzing the content of videos or movie scripts. Martinez et al. [39] propose a way of predicting violence in movies. In this method, they use a dataset collected from 945 Hollywood movie scripts, which are categorized into three groups based on the level of violence: low, medium, and high. Pre-processing methods of NLP are then applied to extract the abusive features. They experimented by using the SVM classifier and RNN, and several NLP methods to prepare the dataset. The model performance achieved 60% in terms of accuracy. Chen et al. [40] concentrate on online streaming video services and detecting bad words. They developed a self-attention model to predicate the level of profanity based on sentence level using the collected dataset of 150 K titles from different genres, and the model achieved 90.6% using the self-attention model in terms of accuracy. In an effort to emphasize the existing unsafe nature of the YouTube Kids channel, Gkolemi et al. [15] studied the new "made for kids" feature. They were also the first to study the characteristics of YouTube accounts publishing videos made for kids. They discovered that 25% of channels with suitable content are set to "made for kids", while only 3% of channels with inappropriate content are set as such. They analyzed 27 different characteristics of channels and how these features are associated with the type of channel and the content it publishes. Among these features are country and

channel creation date, statistics such as subscriptions and video views, keywords and topics, social media links, polarity, and sentiment of description. They applied several experiments using MLC, such as NB, RF, LR, and Neural Net. The RF classifier outperformed them all with an accuracy rate of 79%. It outperformed all the other classifiers in TPRate, Precision, Recall, F1, and AUC. However, it struggled heavily in FPRate, scoring the lowest result out of all the other classifiers. These experiments carried out on the dataset consisted of a set of 2041 random videos from 1338 YouTube channels, which were then placed into two categories: disturbing and suitable. In [41], Vishal Anand proposed a system in order to identify inappropriate content on YouTube by using machine learning classifiers and deep learning architectures. They started by building an offensive dictionary that contained harsh terms that were used to retrieve YouTube videos through the YouTube API. It was used to detect offensive YouTube comments and then rank them on a scale of 0–100% for how inappropriate they were, based on the sentiment analysis concept. Their dataset consisted of 20 k+ videos based on keywords and extracted all the critical information available and categorized them into offensive and non-offensive. They applied several experiments with various combinations of features sentiments from comments, thumbnail images, text-based features, and video and frame-level features using SVM, CNN, LR, and LSTM. The model achieved the highest accuracy, reaching 86.4% by using GRU and Hierarchical Attention Network with GloVec. Dadvar and Eckert in [42] examined DL models on different cyberbullying datasets collected from Twitter, Wikipedia, Formspring, and YouTube. Additionally, they applied different feature representation methods through embedding, such as SSWE and GloVec. The experiment shows that BLSTM with attention models outperforms all other classifiers using the same datasets. The model accuracy achieved is 96% using the YouTube dataset.

Mollas et al. [43] propose two datasets, one for binary classification, which consists of 998 user comments collected from YouTube videos, while the other contains 433 user comments of multi-classes. These datasets can be used for detecting inappropriate content. To assess the proposed datasets, several ML classifiers LR, SVM, RF, and NB were used. Additionally, Gradient Boosting, several deep learning architectures such as CNN and BiLSTM, and transfer learning concepts such as BERT and DistilBERT were also used. The best accuracy rate was achieved through DistilBERT, which reached 80%. In the same way, Vasantharajan and Thayasivam [44] examined the detection of offensive content/speech on YouTube based on textual data from multiple languages while concentrating on Tamil–English. They used a public dataset extracted from 22,299 user comments on YouTube. Then, they divided it into six classes: non-offensive, offensive—targeted–insult–individual, offensive—targeted–insult–group, offensive—targeted–insult–other, offensive—untargeted, and not-Tamil. The dataset is considered an imbalanced dataset. They applied it as a server extensive experiment using multiple deep learning and transfer learning models from HuggingFace, such as CNN-BiLSTM, mBERT-BiLSTM, DistilBERT, XLM-RoBERTa, and ULMFiT. The best-performing model was ULMFiT, which had a recorded accuracy of 74%. In addition, it was suggested that a fresh and adaptable method of selected transliteration and translation procedures were to be used to obtain the most from assembling and fine-tuning.

In the approach of NLP, some researchers [58,59] used topic modeling approaches to identify inappropriate content in YouTube videos. Obadimu et al. [58] applied Latent Dirichlet Allocation to detect inappropriate content in YouTube user comments; 8276 pro-NATO and 46,464 anti-NATO comments were collected from YouTube channels. The user comments were given a score between 0 and 1 under five proposed classes: threats, hate, sexually explicit, insults, and identity-based attack. They performed an analysis and visualization for the frequency of the words. However, this only focused on the existing issues and showed the data analysis. Additionally, Bhuiyan et al. [55] utilized sentiment analysis to retrieve effective YouTube videos using SentiStrength. They collected and analyzed around 1 million user comments from about 1000 YouTube videos in 10 categories. Their methods achieved 75% in terms of accuracy. Differently, Reddy et al. [24] proposed a restriction model when accessing a video to determine whether the user should watch the

video or not based on age. The proposed restriction model is based on sentiment analysis and age detection. The model captures a video stream from a webcam. Then, the photo is captured from the stream. The captured images are pre-processed using grayscale in the cv2 module. The cascade classifier is utilized to detect faces in the images. Once the age detection phase is complete, the sentiment analysis model is applied to the comments extracted from the given video.

Table 2 provides a summary of the existing methods in the literature that are used for detecting and filtering inappropriate content in YouTube videos.

4.2. Meta-Data Analysis Approach

In this approach, two types of methods have been used, namely, meta-data content analysis and content analysis through statistics. In the meta-data content analysis method, researchers attempt to detect and filter inappropriate YouTube videos by analyzing the meta-data of the YouTube videos, such as video title or description, subtitle, rating, number of views, likes, and dislikes as shown in Figure 7. In content analysis through statistics, researchers use a concept of analysis of the video content through surveys or watching videos to determine whether the video is safe or not.

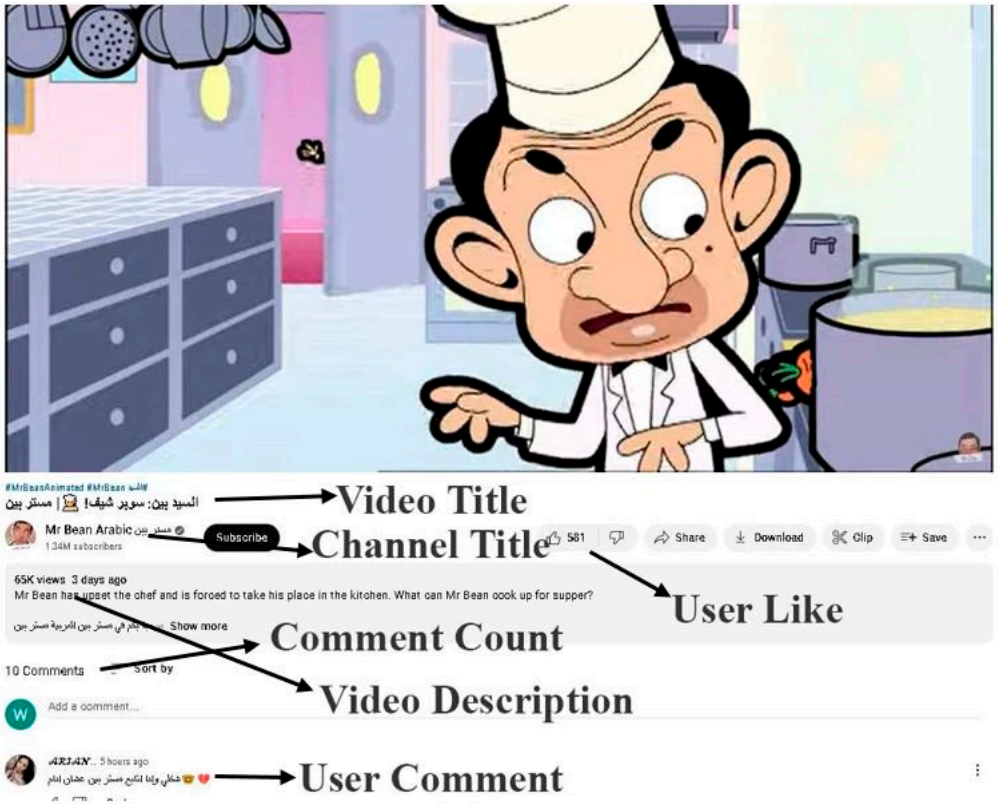

Figure 7. Example of YouTube video meta-data.

Ribeiro et al. [60] analyzed 330,925 videos to carry out an inspection regarding user radicalization on YouTube. These videos were posted on 349 channels that fall under one of these four categories: Alt-lite, Alt-right, Media, and Intellectual Dark Web (IDW). The meta-data, posted comments, and channel recommendations were collected for each video. The analysis presented in [60] includes an inspection of the intersection of commenting users and showed that there is significant user migration in the studied communities. The authors proposed a recommendation algorithm to investigate more than two million videos and claimed that the Alt-right videos are only accessible through channel recommendation, while Alt-lite is simply accessible from the IDW channel.

Table 2. Comparative study among existing methods.

Author(s)	Language	Classes	Size	Representation	Classifiers	Accuracy
Ashraf et al. [25]	English	abusive or non-abusive	2304 comments 29 YouTube	GloVec TF-IDF *n*-grams	LR, MLP, RF, NB, DT, Adaboost, VotingClassifier, CNN, LSTM	87.87% using DT (*n*-gram) 89.86% using CNN + MP with GloVec 91.96% using Adaboost
Hani et al. [26]	English	cyberbullying non-cyberbullying	1608 comments	TFIDF	SVM Neural network	92.8% using neural network 90.3% using SVM
Alcântara et al. [27]	Portuguese	non offensive offensive	400 videos	*n*-grams Word2Vec, FastText, Wang2Vec, and GloVec Transcode Video description	NB LR RN CNN LSTM Transfer learning (BERT, ALBERT)	74% using BERT
Alsubait and Alfageh [28]	Arabic	positive negative	15,000 user comments	TFIDF BoW	NB LR	78.6% LR (BoW) 78.5% NB (TFIDF)
Mouheb et al. [29]	Arabic	bullying and non-bullying	25,000 user comments	N/A	NB	95%
Ahmed et al. [30]	Bengali	bullying and non-bullying	Three datasets (5000, 7000, and 12,000) user comments	TFIDF	NB SVM LR XGBoost	Dataset 1, 76% using SVM Dataset 2, 84% using NB Dataset 3, 80% NB
Awal et al. [31]	English	abusive non abusive	2665 user comments	BOW	NB	80.57%
Balakrishnan t al. [32]	English	bullying and non-bullying	5152 user comments	N/A	RF AdaBoost	97% using Adaboost
Alshamrani et al. [33,34]	English	toxic, obscene, insult, threat and identity hate	3.7 million comments	Word2Vec GloVec	DNN	In [29], most affected age 13–17 years In [30], examine inappropriate URLs
Alshamrani et al. [35]	English	toxic, obscene, identity hate	7.3 million comments (14,506)	Word2Vec	LDA DNN	Found 69% videos contains inappropriate content
AlHarbi et al. [36]	Arabic	bullying and non-bullying	100,327 user comments	BoW	PMI Entropy Chi-Square	Average F1 81%
Yenala et al. [37]	English	inappropriate clean	79,041 user query	word embedding	SVM CNN LSTM BiLSTM C-BiLSTM	89% F1 using C-BiLSTM

Table 2. *Cont.*

Author(s)	Language	Classes	Size	Representation	Classifiers	Accuracy
Cunha et al. [38]	Portuguese	positive negative neutral	2000 user comment	BoW	CNN	84%
Martinez et al. [39]	English	low, medium, high	945 movie scripts	TFIDF word2vec	SVM RNN	60% using RNN (GRU)
Chen et al. [40]	N/A	profane non profane	150 K titles of video streams	TFIDF	DistilBERT Self-attention model	90.6%
Gkolemi et al. [15]	N/A	disturbing and suitable	2041 videos	Features extraction Statistics	NB, RF, LR and Neural Net	79% using RF
Anand et al. [41]	English	offensive and non-offensive	20,000 videos	TFIDF GloVec Fast Text	LR SVM CNN LSTM GUR + Hierarchical Attention Network	86.4% using GUR + Hierarchical Attention Network 67.70% using SVM
Dadvar and Eckert [42]	English	bullying and non-bullying	Formspring 12,000 Wikipedia 10,000 Twitter 16,000 YouTube 54,000	GloVec SSWE	CNN, LSTM, BiLSTM	96% using BLSTM with attention models (YouTube dataset)
Mollas et al. [43]	English	ishate vs. nohate (binary) (multi-label) violence, directed vs. generalized, gender, race, national origin, disability, sexual orientation, religion)	998 user comments 433 (multi-label)	TFIDF	SVM NB LR RF Distil BERT LSTM BiLSTM	80% using Distil BERT 66.94% using RF 66.43% using SVM
Vasantharajan and Thayasivam [44]	Tamil–English	non-offensive, offensive—targeted–insult–individual, offensive—targeted–insult–group, offensive—targeted–insult–other, and offensive—untargeted, non-Tamil	22,299 user comments	GloVec	CNN-BiLSTM, mBERT-BiLSTM, DistilBERT, XLM-RoBERTa, and ULMFiT	74% using ULMFiT
Obadimu et al. [58]	English	threats, insults, hate, sexually explicit, and identity-based attack	8276 (pro-NATO) 46,464 (anti-NATO)	Based on score	Analysis and visualization using word cloud	

Nicoll and Nansen [61] studied toy-unboxing videos to investigate children's participation on YouTube. In toy-unboxing videos, children record themselves reviewing and unboxing toys. The authors analyzed the content of 100 toy-unboxing videos to investigate the children's place in the genres of YouTube. Five categories of these videos were considered: genre, narration, production, product, and branding. The findings in this research indicate that children's production patterns as amateur producers of content are formed by the standardized and common conventions of video genres. In addition, the main findings in the research indicate that there is a relationship between amateur and professional unboxing videos with respect to the methods of filming, editing, narrating, and producing. The authors argued that the mimesis concept is beneficial for attaining insight into a video's popularity among young YouTubers and the collective waves of imitation.

Through different approaches to protect and monitor kids on YouTube videos, some researchers focus on the recommendation applications that appear during the videos, along with the inclusion of inappropriate content through advertisements. Liu et al. [23] conducted a search for inappropriate YouTube advertisements or ads for short. They analyzed the advertising patterns of 24.6 K diverse YouTube videos appropriate for young children and found that 9.9% of the 4.6 K unique advertisements shown on these videos contain inappropriate content for young children. Moreover, Liu et al. perceived that 26.9% of all the 24.6 K appropriate videos included at least one inappropriate ad for young children. Ferreira and Agante [62] emphasized the issue of advertisement and video recommendations by using YouTube Kids video data in social media marketing. In addition, they studied how such strategies affect kids' health. Such skippable and non-skippable advertisements appear frequently during the run time of videos.

Similarly, Feller and Burroughs [63] investigated the evolving transformation in advertising and digital media industries targeting kids through YouTube. The analysis showed companies' strategic use of the media industry and their expertise to transform YouTube stars into global brands. PocketWatch was utilized as a case study and discussed the advertising department formed by PocketWatch to exploit kids' partners by creating advertising based on setting practices and those that cause problems. The study shows how governmental regulation and policy changes affect the development and growth of digital media that targeted kids and act as commercial forces that caused an increase in the number of kids that transitioned from YouTube stars to brands.

In the same manner, Tan et al. [64] examined food ads targeting children on YouTube in Malaysia. SocialBlade.com was utilized to identify the most common 250 YouTube videos targeting children. The ads that appeared while watching these videos were analyzed. The number of ads detected in these videos was 187. Moreover, 38% of these ads were food and beverage. Of those, 56.3% promoted non-core foods. The finding from the analysis showed that the most popular marketing techniques utilized were taste appeal, novelty, the use of animation, fun appeal, use of promotional characters, price, and health and nutrition benefits. The outcome of the conducted analysis was that unhealthy food ads are the dominant content targeting children on YouTube. Similarly, Araceli Castelló-Martínez and Tur-Viñes [65] highlighted the connections among the advertising done by food brands, practices on YouTube, and child obesity in the Latino demographic. They analyzed and compared advertising by food brands on television with videos by child YouTubers (influencing) in Spanish. A content analysis of 304 videos was used to undertake an exploratory study, with 12 factors divided into two categories: the prevalence of ultra-processed versus healthy products in advertising and the marketing approach. The results draw attention to the growth of hybrid media forms, as well as significant differences in the strategies used by brands and young YouTube stars. The results highlight the lack of content advertising warnings directed at young people. Because these media outlets have such a great influence over a vulnerable audience such as children, it is urged that they assume more accountability. It showed that food brands were targeting children with advertisements for foods with low nutritional value that can be harmful, take a toll on a

child's health, and cause obesity. They disguise themselves by evoking joyful and cheerful emotions, when really it is just a scheme from which the food brand can benefit.

Pattier [66] provided insights regarding the use of YouTube videos as educational resources for children and investigated the factors that affect how they may be successfully used by educators. In this regard, 41 educational YouTube channels in Spain were utilized to examine various aspects of the creation of content. These aspects included: video structure, editing, recording, "edutuber" personality, social media usage, and statistics. The analysis of these YouTube channels showed that the science channels acquire more subscribers than other types. In addition, the study claimed that the science channels were very prolific in production, as the number of videos uploaded in one week ranged from one to two. Moreover, few science videos were created by content creators compared to their counterparts in other educational video areas.

Yeo et al. [67] stated that the content of videos is not restricted by federal policy, especially on free video platforms. The authors identified the ten most popular kids' channels on YouTube for content analysis. Five videos with the highest view counts were watched to ensure that these videos were child-directed. The duration of ads and the proportion of inappropriate ads in these videos were noted. They concluded that the total number of ads in the 50 videos was 286. The sidebar ads were the most popular ones at 41.6%, followed by pre-roll ads at 16.8%, banner ads at 14.7%, interstitial ads at 14.3%, and post-video ads at 12.6%. This study was limited to English-language videos.

Araújo et al. in [68] studied the effect of advertisements posted on services targeting kids on YouTube and investigated the demographics of users. To achieve this, data from 12,848 videos were collected from 24 channels in Brazil and 17 channels in the UK and the USA. The collected data included video statistics, comments, and replies. The number of comments collected from these channels was more than 14 million. The authors utilized a free tool for face recognition and text analysis to identify the race, gender, and age of users and to characterize their behavior. Moreover, Named Entity Recognition (NER), with the assistance of video meta-data, was used to extract entities such as brands, products, names, etc. The sentiment analysis concept was used to retrieve the public perception of videos published by analyzing the comments. The children's activities were identified, and this showed that the minimum age of kids utilizing the YouTube platform is 13 years. The study illustrates that the proportion of Black users was very small when compared to Asian and White users.

Baldwin et al. [69] examined the impact of online behavior and social media on unhealthy food consumption by children. The authors carried out an online survey targeting ten- to sixteen-year-old children. The number of children invited to participate in the questionnaire was 582, and 417 responses were received. The analysis of these responses showed that 304 children utilized social networking platforms. Fifty-two children stated watching their favorite food brands on social media ads. Additionally, 25 children reported that food brands are seen in social media hashtags. The number of children that watched YouTube was 329. The main finding of this study was that consuming unhealthy foods and beverages is closely associated with social media behavior. Owing to watching YouTube ads, children were encouraged to purchase unhealthy food.

Tur-Viñes et al. [70] identified young influencer practices on YouTube and the existence of brands in the posted content. To achieve this, the authors selected the top five most-subscribed channels targeting children. For each selected channel, the five most-viewed videos were selected to conduct content analysis, and each video lasted up to 15 min. Several aspects were studied, and each of them consisted of indicators. These aspects were: promotion related to viewers' participation, interaction type, speaking style, editing content complexity, brand content analysis, and the usage of an introduction and outroduction.

Papadamou et al. in [18] proposed a classifier with the ability to recognize inappropriate and unsuitable content intended for kids on YouTube. The authors collected and reviewed child-oriented videos manually and labeled them into one of four categories: suitable, irrelevant, restricted, and disturbing. The collected dataset included 1513 suitable,

419 restricted, 929 disturbing, and 1936 irrelevant videos. The authors collapsed the initial labels of the videos into two categories to distinguish inappropriate from appropriate videos and obtain an accurate classifier. To detect disturbing videos, a deep learning classifier was developed aiming to enhance the performance compared to the baseline models. Several features were utilized in the developed classifier, including title, tags, thumbnail images, statistical meta-data, and style features. The proposed classifier attained an accuracy of 84.3%. The performance obtained in the developed classifier is still lower than the desired performance, but this reflects the similarity between disturbing, restricted, and suitable videos. It was found that content creators attempted to upload disturbing videos with characteristics similar to the kind of child-oriented videos.

Researchers used techniques that utilized a meta-data or content analysis of the YouTube video, as demonstrated in Table 3. In the meta-data approach, they used meta-information, description, and textual attributes of the video. Some of these features include information published by the video uploaders/publishers. Therefore, such features are not sufficient and not accurate enough to filter an inappropriate YouTube video. This is due to such information being provided by the video uploader. Thus, the video uploader will not include terms that refer to unsafe or inappropriate content of the video. In addition, the recommended algorithm and content analysis approach of the YouTube platform suggest a similar video for kids and is usually used in brand and food advertisements.

Table 3. Comparative study on meta-data analysis.

Author(s)	Dataset	Language	Parameters	Class(s)	Finding
Papadamou et al. [18]	4797	N/A	title, tags, thumbnail images, statistical meta-data, and style features	Suitable, irrelevant, restricted, and disturbing	Proposed DL classifier using LSTM, accuracy 84%
Ribeiro [60]	330,925 videos	English	number of subscribers and views, comments captions	Alt-lite Alt-right Media IDW	Proposed recommendation algorithm
Nicoll and Nansen [61]	100 videos	N/A	statistics analysis	Genre, narration, production, product, branding	Analyzing children's toy-unboxing videos
Ferreira and Agante [62]	N/A	N/A	analysis of the existing regulation	N/A	Strategies affect kids' health
Tan et al. [64]	250 videos	N/A	watching videos	Calculating the advertising in videos	Advertising about unhealthy foods
Castelló-Martínez and Tur-Viñes [65]	304 videos	Spanish	watching videos	Content analysis	Food brands marketing through YouTube
Pattier [66]	41 channels	Spain	statistics using SPSS, parameters country, number of subscribers, number of views, knowledge area	Structuring, editing, recording, edutuber personality, social media usage, and statistics	Analysis YouTube channel (educational channels of sciences)
Yeo et al. [67]	5 videos	English	watching videos	Calculating the duration of the advertising	Inappropriate advertising during watching video
Araújo et al. [68]	12,848 videos	English	list of channels, video statistics, comments, and replies	Content analysis	Impact of advertising
Baldwin et al. [69]	417 responses	N/A	purchasing unhealthy food online	Survey response	Survey about advs. on YouTube
Ishikawa et al. [71]	5 channels	Spanish	watching videos	Content analysis	Presence of brands on YouTube videos

4.3. Video Frame Analysis Approach

Video content analysis is concerned with the extraction of meta-data from raw video to be utilized as features in applications such as event detection, tracking, search, clas-

sification, and summarization. It has gained much attention in recent years because of rapid advances in artificial intelligence and deep learning techniques. Furthermore, several studies investigated specific video content classification that could be used for classifying content appropriate for children. For example, violent scenes could be automatically detected by the presence of gunshots, blood, explosions, hate speech, alcohol, drug, and aggressive human actions such as hitting, kicking, punching, slapping, and screaming in audio or video signals. In this approach, researchers gave more attention to detecting and filtering inappropriate YouTube videos that can impact kids by extracting the image frame from the video. Then, features that contain inappropriate content for kids are extracted. Following is a detailed review of recent related work on video content analysis for detecting inappropriate content in videos.

Ishikawa et al. [71] developed deep learning-based approaches for automatically extracting discriminative space–temporal information for filtering "Elsagate" content from disturbing cartoons. They extracted the static information from the videos by sampling one frame per second, and the temporal information was extracted from MPEG motion vectors. Four different deep learning architectures—SqueezeNet, MobileNetV2, GoogLeNet, and NASNet—were evaluated on a public dataset for "Elsagate" classification problems. After processing static and motion data using an SVM classifier, late fusion was finally applied, yielding independent classification scores that were combined to create a single score for the final classification. They claim that when applying transfer learning from ImageNet, NASNet performed better and was a more effective feature extractor.

A video-based detection system was developed by Borg et al. [72]. It utilized a YOLOv3 CNN with automatic feature extraction and an RNN to extract temporal information included in videos. They explained how their method may be utilized for both video-level labeling and localizing pornographic content inside videos. They outline an effective technique for locating sexual items inside pornographic video segments and explain how the categories of sexual objects found could be used to determine the intensity (or "harmfulness") of the pornographic content.

Yousaf and Nawaz [17] proposed a deep-learning model to detect and classify inappropriate video content. The efficientNet-B7 model based on CNN is utilized to extract video textual descriptions to feed them to a BiLSTM network, which then classifies videos. The authors manually collected a dataset by searching popular video cartoon names through the YouTube API; irrelevant videos were then filtered out, which resulted in 1126 videos. These videos were split into one-second duration clips, and each of them was manually annotated into safe, sexual–nudity, or fantasy violence classes. The total number of annotated video clips was 111,561, including 27,003 clips belonging to the sexual–nudity class and 26,650 clips belonging to the fantasy violence class. The experimental results showed that the proposed Efficient Net-BiLSTM with 128 hidden units performed well, outperforming other variants of models, including Efficient Net-FC, Efficient Net-SVM, Efficient Net-KNN, Efficient Net-Random Forest. The authors performed experiments to investigate the impact of the attention mechanism on the performance of the proposed Efficient Net-BiLSTM. The outcomes of the experiments with attention mechanism showed that the proposed Efficient Net-BiLSTM with 128 hidden units without attention mechanism outperformed Efficient Net-BiLSTM with attention mechanism-based models with hidden units 64, 128, 256, and 512.

Garcia et al. [73] developed an application based on skin tone detection filters and a pixel-based approach to identifying pornographic videos and images. With the nudity detection algorithm serving as the system's foundation, the video frames were pre-processed, segmented, and categorized to analyze skin-colored pixels in YCbCr space and then classified as non-skin or skin pixels. The percentage of skin pixels relative to frame size is calculated to be part of the mean baseline that determines whether the files are nude and filters them. A total of 1239 multi-media files were collected from the Web, with 253 videos and 986 images. Using the supplied dataset, the application achieved an accuracy of 80.23% and a precision of 90.33%.

Singh et al. [74] proposed KidsGUARD, a fine-grained technique for detecting sparsely distributed child-unsafe content. They used VGG16 CNN, which was pre-trained on approximately 1.3 million ImageNet data as the model learned to classify each image into one of 1000 groups. They used VGG16 to extract a feature vector from the frame; further, the encoded video representations are fed into the LSTM classifier to detect sparse child-unsafe video content. To test their method, they created a dataset of 109,835 video clips that were specifically filtered for child-unsafe content. They achieved a high precision of 80% and a recall of 81% and found that deep learning approaches outperformed baseline encodings such as Vector of Locally Aggregated Descriptors (VLAD) and Fisher Vector (FV) methods in detecting child-unsafe video content.

Kushal et al. [75] suggested KidsTube, another approach for the detection of child-unsafe content and its promoters. They employed two techniques: a deep learning-based Convolutional Neural Network (CNN) algorithm that uses video frames and supervised classification, which makes use of several video-level, comment-level, and user-level features. For video frame analysis, to detect the different images in a video frame, the CNN was implemented. The CNN-based technique selected crucial frames that exhibited abrupt or gradual changes based on the visual similarity of the adjacent frames in the video. Frame similarity was determined by utilizing the descriptive effectiveness of the Hue, Saturation, and Value (HSV) and Speeded-Up Robust Features (SURF) histogram descriptors. An accuracy of 74% was achieved with CNN based approach, and with supervised classification, they obtained an accuracy of 84%. Further, to identify the community of safe and unsafe content promoters, a network-based approach was applied.

Bhatti et al. [76] designed and implemented an Explicit Content Detection (ECD) system Safe for Work (NSFW) or not suitable media image or video content. The strategy was built on methods for image processing, skin tone detection, and pattern recognition. The image was first transformed using the YCbCr color scheme to identify various objects that were not of relevance. Second, the image's skin tone detection threshold was calculated to filter out different image segments. To identify whether or not an image contains pornographic content, the image explicitness probability was computed. The images were sent to the ECD-CNN model, which was implemented using the CafeNet deep learning framework [77], for classification into explicit and non-explicit images.

A multi-modal deep learning classifier was developed by Chuttur et al. [78] to detect inappropriate cartoon videos for kids. For image analysis, they used the output from VGGNet, and for text analysis, they used the output from LSTM. The VGGNet network is used to identify cartoon image characters, and the LSTM network is utilized to analyze user comments and closed captions related to a video. The VGGNet model was trained and tested on a manually annotated image dataset of 6000 cartoon characters, while the LSTM model was trained and evaluated on a dataset consisting of around 290,000 labeled text records. The LSTM network had a testing accuracy of 94%, whereas the VGGNet network had a testing accuracy of 99%.

Tahir et al. [21] created a deep-learning architecture to detect inappropriate video platform content. Their dataset includes over 1000 videos with violent, fake, or explicit content. They used frame histograms to represent visual features of video frames and cosine-distance between successive frames to detect changes in scenes when processing data. They used a VGG19 pre-trained CNN that had been trained on ImageNet to extract features. The classification model was fed with three different types of features: frame features, movement features, and audio features. Frame features were extracted by running them through a pre-trained CNN, and they noticed that fake videos looked different than real ones. For movement features, the difference in movement characteristics between fake and real cartoon characters was quite noticeable. The characters in a fake cartoon barely move their limbs. Their motion was jerky and hard, so features related to character movement are critical for the CNN model. To extract audio features, each frame's audio spectrogram was passed through the CNN, and its output was recorded. Almost all of the fake and explicit videos had little or no audio. All three—audio, movement, and frame

features—were fed to the VGG19 CNN and LSTM models for classification into violent, fake, or inappropriate video content. They achieved an accuracy of more than 90%.

De Freitas et al. [79] presented a multi-modal (audio and image features) architecture based on Convolutional Neural Networks (CNNs) for detecting inappropriate scenes in video files. Image features were extracted using InceptionV3 as the visual backbone, and audio features were extracted using VGG as the audio backbone. Image features were scaled to a size of 1024 by PCA, and audio was scaled to 128 sizes by PCA. Finally, the image and audio embeddings were concatenated to form the final video embeddings with 1152 dimensions. The video embeddings were then fed into a Support Vector Machine for classification. Appropriate classes received an F1-score of 98.95%, while inappropriate classes received an F1-score of 98.94%.

Khan et al. [80] proposed multi-media content detection techniques for detecting violent content in videos. Initially, key frames were extracted by comparing histogram differences between successive frames. The SIFT feature extraction technique was used to identify unique interest points. The SIFT descriptors were then vectorized using the Gaussian mixture model and Fisher vector encoding. For classification, Spectral Regression–Kernel Discriminant Analysis (SR-KDA) was used. SIFT features encoded with Fisher vector and SR-KDA correctly classified images containing specific violent content, and an accuracy of 97% was obtained.

Alghowinem [2] proposed a multi-modal fusion of audio acoustic, audio transcripts, and video frame methods as an additional layer of content filtering for detecting violent scenes in YouTube videos. The thin-slicing theory was used to extract frames from videos, where several one-second slices were selected randomly from the clip and extracted. The use of a one-second slice ensured that content was analyzed both temporally and in real time. As a pre-processing step for feature extraction, the extracted video slices were segmented into image frames and treated as key frames. Furthermore, CNN, BoVW, and TRoF were used to detect violent scenes. Acoustic features such as spectrograms, MFCCs, and energies were fed into deep neural networks, Gaussian mixture models, and ensemble classifiers, with the best classification results used. The audio was automatically transcribed and analyzed for each slice using automatic speech recognition techniques. Table 4 shows investigated research that falls under this approach.

Researchers proposed two types of approaches for detecting inappropriate content in videos. For instance, the authors in [2,17,21,71,72,74,75,78,79] attempted to extract key frames from videos for extracting significant features and use them to build machine learning classifiers using deep learning and transfer learning techniques. Transfer learning techniques like NASNet, Efficient Net-BiLSTM, ResNet 50, InceptionV3, and VGGNet were applied and outperformed YOLOv3 CNN, RNN, and LSTM deep learning techniques. The performance of transfer learning techniques was better when compared to traditional machine learning and deep learning techniques in terms of speed and performance because the models used knowledge (features, weights, and so on) from pre-trained models that have already been trained to recognize the features. Pre-trained networks were more efficient than training neural networks from scratch. The second category of video frame analysis is based on computer vision techniques [73,76,80], in which handcrafted features were extracted by combining SIFT feature descriptors and segmentation techniques with traditional machine learning classification algorithms such as Support Vector Machines to detect skin and non-skin regions. It was also discovered in [80] that computer vision techniques combined with CNN and ResNet50 deep learning techniques performed efficiently in detecting inappropriate content in video frames.

Table 4. Comparative study on video frame analysis.

Author(s)	Dataset	Method(s)	Class	Accuracy
Alghowinem [2]	YouTube videos	CNN, BoVW, TRoF	hate speech, violence, complex language, sexual references, drug, and alcohol	N/A
Ishikawa et al. [71]	Elsagate videos	NASNet	sensitive non-sensitive	92%
Borg et al. [72]	Videos	YOLOv3 CNN, RNN	19 categories	87%
Yousaf and Nawaz [17]	Cartoon videos	Efficient Net-BiLSTM	safe unsafe	95%
Garcia et al. [73]	Videos, Images	YCbCr, Threshold	skin non-skin	80%
Singh et al. [74]	ImageNet	VGG 16, LSTM	safe unsafe	81%
Kaushal et al. [75]	Cartoon videos	CNN	safe unsafe	74%
Bhatti et al. [76]	Videos	YCbCr, ECD-CNN, ResNet 50	explicit non-explicit	95%
Chuttur et al. [78]	Cartoon images and labeled text	VGGNet LSTM	appropriate inappropriate	99% 94%
Tahir et al. [21]	Cartoon videos	VGG19 LSTM	violent, fake, or inappropriate	90%
De Freitas et al. [79]	Videos	InceptionV3, VGG, SVM	appropriate inappropriate	F1-score 98.95% 98.94%.
Khan et al. [80]	Cartoons	SIFT, SR-KDA	violent non-violent	97%

4.4. Audio Content Analysis Approach

Some researchers have attempted to use an audio content analysis that might be an important addition to the video material that helped determine whether a video is suitable for children through transcription analysis/transcribed text. The sound language and auditory cues were both present in a video's signal. Automatic speech recognition algorithms were utilized to extract the linguistic component of an audio source using NLP methods. Alghowinem in [2] claimed that the content filtering approaches that were based on meta-data attributes were not appropriate. Therefore, real-time content filtering is proposed to keep undesired content from kids. Their method relies on analyzing video and audio content to add more restricted filtrating. The collected videos were annotated into two classes: appropriate and inappropriate for kids. The thin-slicing theory was utilized to extract audio and image frames in order to reduce computational time and resources. The audio was passed to automatic speech recognition techniques for linguistic content analysis. Additionally, the audio signal was analyzed to extract scenes and events. The image frames were analyzed to detect and avoid appropriate scenes. Several features were extracted from each image frame, including color, shapes, objects, and motion features. The proposed method utilized a fusion of three modalities (video, audio, and text transcription from audio) in order to increase the performance of the classification process.

Ali and Senan [81] proposed deep neural network models to detect and classify the violent content appearing in videos. The purpose of their study was to examine the impact of several deep network architectures with different numbers of hidden layers and hidden nodes. These architectures were implemented based on the try-error method to assess the effect of the number of hidden layers and hidden nodes on the classification performance. The VSD2014 dataset was used in the experiments, where 86 videos with 434 attributes (features) were selected. Among these features, eight of them were chosen to represent audio features: Amplitude Envelope (AE), Zero-Crossing Rate (ZCR), Root-Mean-Square Energy (RMS), Band Energy Ratio (BER), Frequency Bandwidth (BW), Spectral Centroid

(SC), Spectral Flux (SF) and Mel-Frequency Cepstral Coefficients (MFCC). On the other hand, four visual features were selected, namely Local Binary Patterns (LBP) with 144-dimensional, Color-Naming Histogram (CNH) with 99-dimensional, Color Moments (CM) with 81-dimensions and Histogram of Oriented Gradients (HOG) with 81-dimensions. The experimental results show that deep neural networks reached an accuracy score of less than 60%, indicating that many improvements were needed.

Hou et al. [82] proposed a bloody video detection system based on audio–visual feature fusion. The authors utilized ResNet and bidirectional LSTM (bi-LSTM) in order to extract static images and temporal and spatial features. The ResNet-50 model was utilized to detect static image characteristics from the bloody videos. The resultant feature maps were passed into the bi-LSTM model. The process of extracting audio features consisted of passing the original waveform of audio into the CNN model. The output of the trained CNN model grouped the audio data into either bloody or non-bloody classes. The authors then proposed the idea of feature fusion by introducing an audio–visual feature fusion layer and passing them to the sigmoid activation function for classification purposes. Owing to the lack of bloody video dataset availability, the authors collected 50 bloody videos on YouTube and collected non-bloody video samples from the MediaEval 2015 dataset. The conducted experiments showed that the fusion-based model outperformed the audio-based and visual-based classifiers. The fusion-based model attained an accuracy of 95% taking advantage of combining the visual and audio features into the same feature maps.

Chaudhari et al. [83] proposed an automated approach to tackle disrespectful content in videos. The input to the proposed system is a video, and the audio was extracted from the given video. The extracted audio was converted into text using speech-to-text techniques in order to find violent and disrespectful words. The video and audio segments were collected from the beginning until the end of the time segment to consider all words in the audio during analysis. These segments were passed to the detection algorithm to detect profane content and then fed to the face detection algorithm. The lip detection algorithm was then applied to detect lips between profane segments in order to apply the lip-blur algorithm. In other words, the segment of the video that contained disrespectful content was replaced with muted audio, and the speaker's lips were pixelated using the Gaussian blur to prevent lip reading.

Aren et al. [84] implemented an unprecedented exploration into the large-scale discovery of recurring audio events in a diverse corpus, where results were promising. Their idea is to apply a streaming, non-parametric clustering algorithm to both spectral features and out-of-domain audio embeddings. They use a small collection of manually annotated audio events to estimate the intrinsic clustering performance. Additionally, in order to provide a valuable technique for unsupervised active learning, they show the benefit of the discovered audio event clusters in weakly supervised learning and informative activity detection. They use weakly supervised learning to exploit the association of video-level meta-data and cluster occurrences to temporally localize audio events. In addition, they use informative activity detection to estimate the semantic saliency of a cluster based on the corpus statistics of the discovered event clusters.

Krithika et al. [85] studied the interaction between YouTube Kids video content and automatic speech recognition systems. They found that existing state-of-the-art automatic speech recognition systems may produce text content highly inappropriate for kids while transcribing YouTube Kids' videos. They used a set of audios for which the well-known automatic speech recognition systems generate inappropriate content for kids. In addition, they showed that some of these errors can be resolved using language models.

5. Results and Discussion

This section discusses the analysis and answers the addressed research questions of this SR. It explores and analyzes the recent techniques that are used for children's safety by detecting or filtering inappropriate content in kids' YouTube videos. This is performed through the collection, exploration, and analysis processes of the existing methods, tech-

niques, and approaches in the literature review from January 2016 to December 2022. In this SR, 72 research articles were analyzed, which are related to its objectives. The findings of this SR answer the research questions that were formulated. These RQs are as follows:

RQ1: What are the issues and threats that kids encounter on YouTube?

Through the analysis process of the relevant research articles concerned with kids' digital safety in this SR, we found that researchers from different disciplines have captured attention concerning this issue and how social media platforms influence children in various ways, such as mental health [86–96], food and beverage marketing [3,64,69,97–103], cyberbullying [104–111], and childhood obesity [112–115]. Additionally, internationally recognized organizations raised awareness and were concerned about children's safety online in the digital world. Some of these organizations are as follows: European Data Protection Law [116], American Academy of Pediatrics [115], Children's Online Privacy Protection Act [117], and Children's Media Regulations [118]. This is due to the issues and threats that are increasing rapidly. Therefore, all research findings reflect the threats and issues that exist in the online world for children.

These threats and issues are foundational and develop skills, kids' attitudes and behaviors, addiction to social media, and health issues. In foundational and development skills, during the early childhood stage when the child spends long periods of time watching content found on YouTube. The type of content viewed on YouTube at such an early stage in a child's development can have a significant impact on their character and personality [1,119]. Therefore, if a child is exposed to harmful content at this important stage, it can have a negative impact on the child because children learn primarily through observation during this time [120–122]. In such a way, children act, behave, and cope with other people in society.

Additionally, kids' attitudes and behaviors are considered one of the threats that can be affected by social media platforms [3,15,35,123]. Kids' behavior can be altered through the content viewed online, in a positive or negative way, whether it is by reading the comments of other users online, watching videos that are classified as harmful, or by looking at advertisements that appear randomly in the video. Therefore, YouTube is considered an unsafe platform for kids [19].

On the matter of health issues, social media has been linked to depression, anxiety, and loneliness. It can make people feel isolated and alone. Social media has repeatedly been proven to be a direct link to a plethora of psychiatric illnesses, including the above-mentioned psychological problems and various forms of eating disorders. The prolonged use of social media platforms by today's youth has negative effects on mental health. In fact, it has been discovered that youths who use social media more frequently are more likely to suffer from depression, anxiety, eating disorders and body dysmorphic disorders, and self-harm and suicidal tendencies. Social media undoubtedly contributes significantly to the rising rate of instances of melancholy, anxiety, and eating disorders, which in turn raise the risk of suicide. According to studies, suicidal people to broadcast or celebrate self-harm and other types of upsetting material frequently use social media as a platform. Predominantly, the YouTube platform can have a strong impact on the mental health of kids. One of the most common mental issues a child can acquire through exposure to social media is depression [25,123].

Another threat that parents encounter is their child's addiction to social media. Kids may develop an addiction as a result of the constant stream of videos, especially on YouTube [18,21,124,125]. The fast-paced, attention-grabbing attitudes of content creators attract the attention of children. Kids often return to these channels/content creators because the uploader usually has an advertised schedule that encourages children to come back and check to see if there have been any new uploads. To put it into perspective, the active community on social media, particularly YouTube, makes children always come back for more.

RQ2: What existing methods, techniques, and approaches are used to protect kids and prevent potential issues?

The aforementioned issues and threats arose due to the increasing popularity of YouTube video watching among children. Several researchers have made many attempts to solve such issues using machine learning approaches early on that required small datasets. Recently, deep and transfer learning approaches contributed to achieving better accuracy in detecting unsafe videos and inappropriate video content. These methods are applied to data collected from the YouTube platform. These data can come in the form of textual data, such as user-generated comments, video descriptions, meta-data, videos in both image and audio, or content analysis methods based on statistical approaches. These attempts concentrate on several methods to provide proper solutions. Some authors analyzed the textual data through different methods. Others suggested slicing the frames of the video and extracting information or identifying unusual patterns for kids in order to detect inappropriate content through using NLP methods for textual data, computer vision methods for video, or transcode that converts the speech to text and then perform further processing by training. Additionally, some went to record to achieve high accuracy in the detection process using the machine, deep, transfer learning models.

Therefore, some studies focus on particular languages such as English vocabulary [8,23–26,31–35,37,42–44,58], Arabic [28,29,36], Portuguese [27,38], Bengali [30], and Tamil [44]. We noticed that great consideration is given to the English language, as it is a universal language. However, little attention has been given to the Arabic language, despite the fact that it is the fourth most spoken language on the web. Other studies proposed methods to improve the model performance in terms of accuracy [25–34]. Studies [15,25–32,37,39,41] used a machine learning classifier, and the best accuracy has been recorded using NB [28–31], in which accuracy reached 78.5%, 95%, 84%, and 80.57%, respectively, followed by SVM, which achieved 90.3%, 76%, and 66.43%, [26,30,43] respectively. As a result of the majority of studies' findings, the NB classifier outperformed other ML classifiers. In addition, ensemble classifiers have been utilized and recorded higher accuracy compared to ML classifiers such as Adaboost, which in studies [25,32] recorded 91.96% and 97%, and RF in [15,43] achieved 79% and 66.94%, respectively. Moreover, the best accuracy has been recorded using the deep learning approaches, particularly using RNN and its variants; these studies [26,37,39–44] have obtained a higher accuracy compared to the previously mentioned methods, the accuracy reached 92.8%, 89%, 60%, 90.6%, 86.4%, 96%, 80%, and 74%, respectively. Furthermore, NLP methods and pre-trained models that used word embedding concepts such as Word2Vec and GloVec were utilized in [27,33–35,37,39,41,42,44] and outperformed word representation such as BoW and TFIDF that were used in [25,26,28,30,31,36,39–41,43]. On the other hand, transfer learning has been used in some of the recent studies [27,40,43,44] by using BERT or models based on BERT.

In frame-video analysis, we found that researchers utilized DL more than ML due to its accuracy in detecting inappropriate content. Researchers proposed two types of approaches for detecting inappropriate content in videos. For instance, authors in [1–3,5,6,8,10,11,13] attempted to create video-related features and used them to build machine learning classifiers using deep learning and transfer learning techniques. They employed features based on video key frames. Transfer learning techniques such as NASNet, EfficientNet-BiLSTM, ResNet 50, InceptionV3, and VGGNet were applied and outperformed deep learning techniques such as YOLOv3 CNN, RNN, and LSTM. Transfer learning techniques outperformed traditional machine learning and deep learning techniques in terms of speed and performance. Because the models used knowledge (features, weights, and so on) from pre-trained models that have already been trained to recognize the features, it is more efficient than training neural networks from scratch. The second category of video frame analysis is based on computer vision techniques [4,12] in which handcrafted features were extracted by combining SIFT feature descriptors and segmentation techniques with traditional machine learning classification algorithms such as Support Vector Machines to detect skin and non-skin regions. It was also discovered [7] that computer vision techniques combined with CNN and ResNet50 deep learning techniques performed better in detecting inappropriate content in video frames.

On the other hand, some research studies highlighted the issue of advertisements that appear suddenly to children while they are watching YouTube videos. These methods are based on content analysis and measure their effectiveness on kids. The research used statistical methods or surveys as tools to identify this phenomenon that is increasing among kids on the YouTube platform. Therefore, the research studies focus on the study of the effects of advertisements on kids [62,63,67,68]. Recently many studies focused on food and brands [64,65,70] in order to attract kids. It is one of the many food-marketing strategies. In [69], we highlighted the relation between unhealthy food consumption and YouTube. In addition, some research examines unboxing videos [61].

RQ3: What are the challenges in protecting kids from inappropriate content on YouTube?

YouTube is considered a tool for education and entertainment. The misuse of this tool negatively affects children. In this digital era, the majority of kids access the internet through different mobile devices. Because to this, families and kids' international organizations that work to protect kids are concerned and encounter great challenges when faced with the question of how to protect kids in the digital world [1,4,7,24,34,85,126–130]. To achieve this objective and to answer the research question, there are three main challenges to adding a layer to the social media platform. The first is finding a multi-lingual dataset in order to obtain high accuracy in detecting inappropriate content [23,131–135]. Such a dataset requires efforts to construct it using a large repository of multi-lingual text that contains inappropriate content by utilizing word synonyms such as WordNet and multi-lingual dictionaries. The second is refining and improving model accuracy by using three or four approaches utilizing deep learning architectures, in addition to using an approach of transfer learning by training deep learning models using the large multi-lingual dataset. Such will improve model performance in terms of accuracy and F1-score to detect and classify inappropriate content. Lastly, for notification purposes, notification should be sent to parents in the event that children access inappropriate content, in accordance with a set of guidelines that includes user-generated comments. A safety layer can be added to YouTube's architecture or a new layer can be added that filters YouTube videos [2,136,137] and user content with the option of parental control using internet-of-things methods for notifications.

6. Conclusions

The number of YouTube videos spread quickly in tremendous ways. Furthermore, many digital devices in kids' hands helped in that spread, and because of this, much inappropriate content has an impact on kids. This is considered the most important issue on the YouTube platform. After the SR search, 72 papers were retrieved, concentrating on detecting, preventing, and monitoring inappropriate content on the kids' platform. Therefore, the findings and conclusions of the included research articles were categorized into four approaches based on textual analysis, video analysis, audio analysis, and meta-data content analysis.

However, the challenge still exists in trying to find the best way to detect and filter inappropriate videos for kids using an automatic strategy by machine and deep learning approaches in different ways. In this research gap, researchers are required to find methods to improve the detection, monitoring, and prevention of unsafe YouTube videos on kids' platforms. This is still an interesting research area for researchers and Ph.D. students, as well as for the industry to propose a security layer in order to achieve security for kids on social media and not only on YouTube. However, our focus was only on the YouTube platform, while there are other social media platforms that also impact kids and society.

The main contribution of this SR is to give an overview of the existing methods, approaches, and techniques that are adapted to protect kids on the YouTube platform. Kids' digital safety is an active research area. Therefore, this SR can be beneficial and can be used as a reference for many researchers as a starting point in the area of social media security for kids.

Author Contributions: Conceptualization, W.M.S.Y., S.I.A. and A.A.; methodology, W.M.S.Y.; formal analysis, S.I.A., W.M.S.Y., A.A., L.S. and R.A.; resources, S.I.A., W.M.S.Y., A.A., L.S. and R.A.; data collection, S.I.A., W.M.S.Y., A.A., L.S. and R.A.; writing, S.I.A., W.M.S.Y., A.A., L.S. and R.A.; visualization, W.M.S.Y.; review, W.M.S.Y. and S.I.A.; supervision, S.I.A. and W.M.S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Taibah University, Department of Scientific Research, the Research Capabilities Initiative, project ID: RC-442-72.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Izci, B.; Jones, I.; Özdemir, T.B.; Alktebi, L.; Bakir, E. YouTube & Young Children: Research, Concerns and New Directions. In *Crianças, Famílias e Tecnologias. Que Desafios? Que Caminhos?* Lisbon School of Education: Lisbon, Portugal, 2019; pp. 81–92.
2. Alghowinem, S. A Safer YouTube Kids: An Extra Layer of Content Filtering Using Automated Multimodal Analysis. In *Advances in Intelligent Systems and Computing*; Springer: Berlin, Germany, 2018; Volume 868.
3. Vanwesenbeeck, I.; Hudders, L.; Ponnet, K. Understanding the YouTube Generation: How Preschoolers Process Television and YouTube Advertising. *Cyberpsychol. Behav. Soc. Netw.* **2020**, *23*, 426–432. [CrossRef]
4. Alshamrani, S.; Abusnaina, A.; Mohaisen, D. Hiding in Plain Sight: A Measurement and Analysis of Kids' Exposure to Malicious URLs on YouTube. In Proceedings of the 2020 IEEE/ACM Symposium on Edge Computing, SEC 2020, San Jose, CA, USA, 13 November 2020.
5. Hanson, G.; Haridakis, P. YouTube Users Watching and Sharing the News: A Uses and Gratifications Approach. *J. Electron. Publ.* **2008**, *11*. [CrossRef]
6. O'Keeffe, G.S.; Clarke-Pearson, K.; Mulligan, D.A.; Altmann, T.R.; Brown, A.; Christakis, D.A.; Falik, H.L.; Hill, D.L.; Hogan, M.J.; Levine, A.E.; et al. Clinical Report–The Impact of Social Media on Children, Adolescents, and Families. *Pediatrics* **2011**, *127*, 800–804. [CrossRef]
7. Fabian-Weber, N. 8 Dangers of Social Media to Discuss with Kids and Teens. Available online: https://www.joyoffaith.com/assets/8-dangers-of-social-media-to-discuss-with-kids-and-teens.pdf (accessed on 13 March 2023).
8. *Miller Caroline Does Social Media Use Cause Depression?* Child Mind Institute: New York, NY, USA, 2020.
9. Temban, M.M.; Hua, T.K.; Said, N.E.M. Exploring Informal Learning Opportunities via Youtube Kids among Children during COVID-19. *Acad. J. Interdiscip. Stud.* **2021**, *10*, 272. [CrossRef]
10. Pires, F.; Masanet, M.J.; Scolari, C.A. What Are Teens Doing with YouTube? Practices, Uses and Metaphors of the Most Popular Audio-Visual Platform. *Inf. Commun. Soc.* **2021**, *24*, 1175–1191. [CrossRef]
11. Szmuda, T.; Syed, M.T.; Singh, A.; Ali, S.; Özdemir, C.; Słoniewski, P. YouTube as a Source of Patient Information for Coronavirus Disease (COVID-19): A Content-Quality and Audience Engagement Analysis. *Rev. Med. Virol.* **2020**, *30*, e2132. [CrossRef] [PubMed]
12. Balanzategui, J. 'Disturbing' Children's YouTube Genres and the Algorithmic Uncanny. *New Media Soc.* **2021**. [CrossRef]
13. Caldeiro-Pedreira, M.-C.; Renés-Arellano, P.; Castillo-Abdul, B.; Aguaded, I. YouTube Videos for Young Children: An Exploratory Study. *Digit. Educ. Rev.* **2022**, 32–43. [CrossRef]
14. Statista Research Department YouTube–Statistics & Facts. Available online: https://www.statista.com/topics/2019/youtube/#dossierKeyfigures (accessed on 13 March 2023).
15. Gkolemi, M.; Papadopoulos, P.; Markatos, E.; Kourtellis, N. YouTubers Not MadeForKids: Detecting Channels Sharing Inappropriate Videos Targeting Children. In Proceedings of the 14th ACM Web Science Conference, Barcelona, Spain, 26–29 June 2022; pp. 370–381.
16. Ben-Yair, S. Introducing the Newest Member of Our Family, the YouTube Kids App—Available on Google Play and the App Store. In *YouTube Official Blog*; YouTube: San Bruno, CA, USA, 23 February 2015.
17. Yousaf, K.; Nawaz, T. A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos. *IEEE Access* **2022**, *10*, 16283–16298. [CrossRef]
18. Papadamou, K.; Papasavva, A.; Zannettou, S.; Blackburn, J.; Kourtellis, N.; Leontiadis, I.; Stringhini, G.; Sirivianos, M. Disturbed YouTube for Kids: Characterizing and Detecting Disturbing Content on YouTube. In Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020, Virtually, 8–11 June 2020.
19. Wilson, H. Youtube Is Unsafe for Children: Youtube's Safeguards and the Current Legal Framework Are Inadequate to Protect Children from Disturbing Content. *Seattle J. Tech. Envtl. Innov. L.* **2020**, *10*, 237.
20. Weston, P. Youtube Kids App Is Still Showing Disturbing Videos. *Mail Online*, 6 February 2018.
21. Tahir, R.; Ahmed, F.; Saeed, H.; Ali, S.; Zaffar, F.; Wilson, C. Bringing the Kid Back into YouTube Kids: Detecting Inappropriate Content on Video Streaming Platforms. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019, Vancouver, BC, Canada, 27–30 August 2019.

22. Burroughs, B. Youtube Kids: The App Economy and Mobile Parenting. *Soc. Media Soc.* **2017**, *3*. [CrossRef]
23. Liu, J.; Tandon, R.; Durairaj, U.; Guo, J.; Zahabizadeh, S.; Ilango, S.; Tang, J.; Gupta, N.; Zhou, Z.; Mirkovic, J. Did Your Child Get Disturbed by an Inappropriate Advertisement on YouTube? *arXiv* **2022**, arXiv:2211.02356.
24. Reddy, S.; Srikanth, N.; Sharvani, G.S. Development of Kid-Friendly Youtube Access Model Using Deep Learning. In *Lecture Notes in Networks and Systems*; Springer: Berlin, Germany, 2021; Volume 132.
25. Ashraf, N.; Zubiaga, A.; Gelbukh, A. Abusive Language Detection in Youtube Comments Leveraging Replies as Conversational Context. *PeerJ Comput. Sci.* **2021**, *7*, e742. [CrossRef]
26. Hani, J.; Nashaat, M.; Ahmed, M.; Emad, Z.; Amer, E.; Mohammed, A. Social Media Cyberbullying Detection Using Machine Learning. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*. [CrossRef]
27. de Alcântara, C.S.; Feijó, D.; Moreira, V.P. Offensive Video Detection: Dataset and Baseline Results. In Proceedings of the LREC 2020–12th International Conference on Language Resources and Evaluation, Conference Proceedings, Palais du Pharo, Marseille, France, 11–16 May 2020.
28. Alsubait, T.; Alfageh, D. Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments. *Int. J. Comput. Sci. Netw. Secur.* **2021**, *21*. [CrossRef]
29. Mouheb, D.; Albarghash, R.; Mowakeh, M.F.; Aghbari, Z.A.; Kamel, I. Detection of Arabic Cyberbullying on Social Networks Using Machine Learning. In Proceedings of the Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA 2019, Abu Dhabi, United Arab Emirates, 3–7 November 2019; Volume 2019.
30. Ahmed, M.T.; Rahman, M.; Nur, S.; Islam, A.Z.M.T.; Das, D. Natural Language Processing and Machine Learning Based Cyberbullying Detection for Bangla and Romanized Bangla Texts. *Telkomnika (Telecommun. Comput. Electron. Control)* **2022**, *20*, 89–97. [CrossRef]
31. Awal, M.A.; Rahman, M.S.; Rabbi, J. Detecting Abusive Comments in Discussion Threads Using Naïve Bayes. In Proceedings of the 2018 International Conference on Innovations in Science, Engineering and Technology, ICISET 2018, Chittagong, Bangladesh, 27–28 October 2018; IEEE: New York, NY, USA, 2018.
32. Balakrishnan, V.; Ng, S.K. Personality and Emotion Based Cyberbullying Detection on YouTube Using Ensemble Classifiers. *Behav. Inf. Technol.* **2022**, 1–12. [CrossRef]
33. Alshamrani, S. Detecting and Measuring the Exposure of Children and Adolescents to Inappropriate Comments in YouTube. In Proceedings of the International Conference on Information and Knowledge Management, online, 19–23 October 2020.
34. Alshamrani, S.; Abusnaina, A.; Abuhamad, M.; Nyang, D.; Mohaisen, D. Hate, Obscenity, and Insults: Measuring the Exposure of Children to Inappropriate Comments in YouTube. In Proceedings of the The Web Conference 2021–Companion of the World Wide Web Conference, WWW 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 508–515.
35. Alshamrani, S.; Abuhamad, M.; Abusnaina, A.; Mohaisen, D. Investigating Online Toxicity in Users Interactions with the Mainstream Media Channels on YouTube. In Proceedings of the CEUR Workshop Proceedings, Galway, Ireland, 19–23 October 2020; Volume 2699.
36. AlHarbi, B.Y.; AlHarbi, M.S.; AlZahrani, N.J.; Alsheail, M.M.; Alshobaili, J.F.; Ibrahim, D.M. Automatic Cyber Bullying Detection in Arabic Social Media. *Int. J. Eng. Res. Technol.* **2019**, *12*, 2330–2335.
37. Yenala, H.; Jhanwar, A.; Chinnakotla, M.K.; Goyal, J. Deep Learning for Detecting Inappropriate Content in Text. *Int. J. Data Sci. Anal.* **2018**, *6*, 273–286. [CrossRef]
38. Cunha, A.A.L.; Costa, M.C.; Pacheco, M.A.C. Sentiment Analysis of YouTube Video Comments Using Deep Neural Networks. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11508 LNAI.
39. Martinez, V.R.; Somandepalli, K.; Singla, K.; Ramakrishna, A.; Uhls, Y.T.; Narayanan, S. Violence Rating Prediction from Movie Scripts. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019.
40. Chen, J.; Wei, K.; Hao, X. Detect Profane Language in Streaming Services to Protect Young Audiences. In Proceedings of the ECNLP 2021–4th Workshop on e-Commerce and NLP, Bangkok, Thailand, 1–6 August 2021.
41. Anand, V.; Shukla, R.; Gupta, A.; Kumar, A. Customized Video Filtering on YouTube. *arXiv* **2019**, arXiv:1911.04013.
42. Dadvar, M.; Eckert, K. Cyberbullying Detection in Social Networks Using Deep Learning Based Models. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin, Germany, 2020; Volume 12393 LNCS.
43. Mollas, I.; Chrysopoulou, Z.; Karlos, S.; Tsoumakas, G. ETHOS: A Multi-Label Hate Speech Detection Dataset. *Complex Intell. Syst.* **2022**, *8*. [CrossRef]
44. Vasantharajan, C.; Thayasivam, U. Towards Offensive Language Identification for Tamil Code-Mixed YouTube Comments and Posts. *SN Comput. Sci.* **2022**, *3*, 94. [CrossRef]
45. Haidar, B.; Chamoun, M.; Serhrouchni, A. Multilingual Cyberbullying Detection System: Detecting Cyberbullying in Arabic Content. In Proceedings of the 2017 1st Cyber Security in Networking Conference, CSNet 2017, Janeiro, Brazil, 18–20 October 2017; Volume 2017-January.
46. Rahman, M.H.U.; Divya, M.; Reddy, B.R.; Kumar, K.S.; Vani, P.R. Cyberbullying Detection Using Natural Language Processing. *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)* **2022**, *10*. [CrossRef]

47. Al-Makhadmeh, Z.; Tolba, A. Automatic Hate Speech Detection Using Killer Natural Language Processing Optimizing Ensemble Deep Learning Approach. *Computing* **2020**, *102*, 501–522. [CrossRef]
48. Stepanova, N.; Muthemba, W.; Todrzak, R.; Cross, M.; Ames, N.; Raiti, J. Natural Language Processing and Sentiment Analysis for Verbal Aggression Detection; A Solution for Cyberbullying during Live Video Gaming. In *Proceedings of the ACM International Conference Proceeding Series*; ACM: New York, NY, USA, 2021.
49. Haidar, B.; Chamoun, M.; Serhrouchni, A. Arabic Cyberbullying Detection: Enhancing Performance by Using Ensemble Machine Learning. In Proceedings of the 2019 International Conference on Internet of Things (Ithings) and IEEE Green Computing and Communications (Greencom) and IEEE Cyber, Physical and Social Computing (Cpscom) and IEEE Smart Data (Smartdata), Atlanta, GA, USA, 14–17 July 2019; pp. 323–327.
50. Moreno, M.A.; Gower, A.D.; Brittain, H.; Vaillancourt, T. Applying Natural Language Processing to Evaluate News Media Coverage of Bullying and Cyberbullying. *Prev. Sci.* **2019**, *20*, 1274–1283. [CrossRef]
51. Lucky, E.A.E.; Sany, M.M.H.; Keya, M.; Khushbu, S.A.; Noori, S.R.H. An attention on sentiment analysis of child abusive public comments towards bangla text and ml. In Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021, Kharagpur, India, 6–8 July 2021.
52. Mozafari, M.; Farahbakhsh, R.; Crespi, N. A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. In *Proceedings of the Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 928–940.
53. Biere, S. *Hate Speech Detection Using Natural Language Processing Techniques*; Vrije Universiteit Amsterdam: Amsterdam, The Netherlands, 2018.
54. Sigurbergsson, G.I.; Derczynski, L. Offensive Language and Hate Speech Detection for Danish. In Proceedings of the LREC 2020–12th International Conference on Language Resources and Evaluation, Conference Proceedings, Marseille, France, 11–16 May 2020.
55. Bhuiyan, H.; Ara, J.; Bardhan, R.; Islam, M.R. Retrieving YouTube Video by Sentiment Analysis on User Comment. In Proceedings of the 2017 IEEE International Conference on Signal and Image Processing Applications, ICSIPA 2017, Kuching, Malaysia, 12–14 September 2017; IEEE: New York, NY, USA, 2017.
56. Kavya Agarwal Should Your Young Child Watch YouTube? *Pros and Cons.* Available online: https://www.blackboardradio.com/post/should-your-young-child-watch-youtube-pros-and-cons (accessed on 10 November 2022).
57. Is YouTube For Kids: The Pros And Cons Of Kids On YouTube. Available online: https://tiptopbrain.com/blog/is-youtube-for-kids-the-pros-and-cons-of-kids-on-youtube/ (accessed on 10 November 2022).
58. Obadimu, A.; Mead, E.; Hussain, M.N.; Agarwal, N. Identifying Toxicity within Youtube Video Comment. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11549 LNCS.
59. Shekhar, S.; Akanksha; Saini, A. Utilizing Topic Modelling to Identify Abusive Comments on YouTube. In Proceedings of the 2021 International Conference on Intelligent Technologies, CONIT 2021, Hubli, India, 25–27 June 2021; IEEE: New York, NY, USA, 2021.
60. Ribeiro, M.H.; Ottoni, R.; West, R.; Almeida, V.A.F.; Wagner Meira, W.M. Auditing Radicalization Pathways on YouTube. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency FAT* '20, Barcelona, Spain, 27–30 January 2020; pp. 131–141.
61. Nicoll, B.; Nansen, B. Mimetic Production in YouTube Toy Unboxing Videos. *Soc. Media Soc.* **2018**, *4*, 2056305118790761. [CrossRef]
62. Ferreira, M.R.; Agante, L. The Use of Algorithms to Target Children While Advertising on YouTube Kids Platform: A Reflection and Analysis of the Existing Regulation. *Int. J. Mark. Commun. New Media* **2020**, 29–53.
63. Feller, G.; Burroughs, B. Branding Kidfluencers: Regulating Content and Advertising on YouTube. *Telev. New Media* **2022**, *23*, 575–592. [CrossRef]
64. Tan, L.; Ng, S.H.; Omar, A.; Karupaiah, T. What's on YouTube? A Case Study on Food and Beverage Advertising in Videos Targeted at Children on Social Media. *Child. Obes.* **2018**, *14*, 280–290. [CrossRef] [PubMed]
65. Castelló-Martínez, A.; Tur-Viñes, V. Obesity and Food-related Content Aimed at Children on YouTube. *Clin Obes* **2020**, *1*, e12389. [CrossRef]
66. Pattier, D. Science on Youtube: Successful Edutubers. *Rev. Int. Tecnol. Cienc. Soc.* **2021**, *10*, 1–15. [CrossRef]
67. Yeo, S.L.; Schaller, A.; Robb, M.B.; Radesky, J.S. Frequency and Duration of Advertising on Popular Child-Directed Channels on a Video-Sharing Platform. *JAMA Netw. Open* **2021**, *4*, e219890. [CrossRef] [PubMed]
68. Araújo, C.S.; Magno, G.; Meira, W.; Almeida, V.; Hartung, P.; Doneda, D. Characterizing Videos, Audience and Advertising in Youtube Channels for Kids. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 10539 LNCS.
69. Baldwin, H.J.; Freeman, B.; Kelly, B. Like and Share: Associations between Social Media Engagement and Dietary Choices in Children. *Public Health Nutr.* **2018**, *21*, 3210–3215. [CrossRef]
70. Tur-Viñes, V.; Núñez-Gómez, P.; González-Río, M.J. Kid Influencers on YouTube. A Space for Responsibility. *Rev. Lat. Comun. Soc.* **2018**, *73*, 1211–1230. [CrossRef]
71. Ishikawa, A.; Bollis, E.; Avila, S. Combating the Elsagate Phenomenon: Deep Learning Architectures for Disturbing Cartoons. In Proceedings of the 2019 7th International Workshop on Biometrics and Forensics, IWBF 2019, Cancun, Mexico, 2 May 2019.

72. Borg, M.; Tabone, A.; Bonnici, A.; Cristina, S.; Farrugia, R.A.; Camilleri, K.P. Detecting and Ranking Pornographic Content in Videos. *Forensic Sci. Int. Digit. Investig.* **2022**, *42*, 301436. [CrossRef]
73. Garcia, M.B.; Revano, T.F.; Habal, B.G.M.; Contreras, J.O.; Enriquez, J.B.R. A Pornographic Image and Video Filtering Application Using Optimized Nudity Recognition and Detection Algorithm. In Proceedings of the 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, HNICEM 2018, Baguio City, Philippines, 29 November 2019; IEEE: New York, NY, USA, 2019.
74. Singh, S.; Buduru, A.B.; Kaushal, R.; Kumaraguru, P. KidsGUARD: Fine Grained Approach for Child Unsafe Video Representation and Detection. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing SAC '19, Limassol, Cyprus, 8 April 2019; Volume Part F147772, pp. 2104–2111.
75. Kaushal, R.; Saha, S.; Bajaj, P.; Kumaraguru, P. KidsTube: Detection, Characterization and Analysis of Child Unsafe Content and Promoters on YouTube. In Proceedings of the 2016 14th Annual Conference on Privacy, Security and Trust (PST), 12–14 December 2016; IEEE: New York, NY, USA, 2016; pp. 157–164.
76. Qamar Bhatti, A.; Umer, M.; Adil, S.H.; Ebrahim, M.; Nawaz, D.; Ahmed, F. Explicit Content Detection System: An Approach towards a Safe and Ethical Environment. *Appl. Comput. Intell. Soft Comput.* **2018**, *2018*, 1463546. [CrossRef]
77. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; Volume 2016.
78. Chuttur, M.Y.; Nazurally, A. A Multi-Modal Approach to Detect Inappropriate Cartoon Video Contents Using Deep Learning Networks. *Multimed. Tools Appl.* **2022**, *81*, 16881–16900. [CrossRef]
79. de Freitas, P.V.A.; Mendes, P.R.C.; dos Santos, G.N.P.; Busson, A.J.G.; Guedes, Á.L.; Colcher, S.; Milidiú, R.L. A Multimodal CNN-Based Tool to Censure Inappropriate Video Scenes. *arXiv* **2019**, arXiv:1911.03974.
80. Khan, M.; Tahir, M.A.; Ahmed, Z. Detection of Violent Content in Cartoon Videos Using Multimedia Content Detection Techniques. In Proceedings of the 21st International Multi Topic Conference, INMIC 2018, Karachi, Pakistan, 31 July 2018.
81. Ali, A.; Senan, N. Violence Video Classification Performance Using Deep Neural Networks. In *Advances in Intelligent Systems and Computing*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 700.
82. Hou, C.; Wu, X.; Wang, G. End-to-End Bloody Video Recognition by Audio-Visual Feature Fusion. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11256 LNCS.
83. Chaudhari, A.; Davda, P.; Dand, M.; Dholay, S. Profanity Detection and Removal in Videos Using Machine Learning. In Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, Coimbatore, India, 20–22 January 2021.
84. Jansen, A.; Gemmeke, J.F.; Ellis, D.P.W.; Liu, X.; Lawrence, W.; Freedman, D. Large-Scale Audio Event Discovery in One Million YouTube Videos. In Proceedings of the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 March 2017.
85. Ramesh, K.; KhudaBukhsh, A.R.; Kumar, S. 'Beach'to "Bitch": Inadvertent Unsafe Transcription of Kids' Content on YouTube. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 12108–12118.
86. Ivie, E.J.; Pettitt, A.; Moses, L.J.; Allen, N.B. A Meta-Analysis of the Association between Adolescent Social Media Use and Depressive Symptoms. *J. Affect. Disord.* **2020**, *275*, 165–174. [CrossRef]
87. Barry, C.T.; Sidoti, C.L.; Briggs, S.M.; Reiter, S.R.; Lindsey, R.A. Adolescent Social Media Use and Mental Health from Adolescent and Parent Perspectives. *J. Adolesc.* **2017**, *61*, 1–11. [CrossRef]
88. O'Reilly, M.; Dogra, N.; Whiteman, N.; Hughes, J.; Eruyar, S.; Reilly, P. Is Social Media Bad for Mental Health and Wellbeing? Exploring the Perspectives of Adolescents. *Clin. Child Psychol. Psychiatry* **2018**, *23*, 601–613. [CrossRef]
89. Primack, B.A.; Escobar-Viera, C.G. Social Media as It Interfaces with Psychosocial Development and Mental Illness in Transitional Age Youth. *Child Adolesc. Psychiatr. Clin. N. Am.* **2017**, *26*, 217–233. [CrossRef]
90. Beeres, D.T.; Andersson, F.; Vossen, H.G.M.; Galanti, M.R. Social Media and Mental Health Among Early Adolescents in Sweden: A Longitudinal Study with 2-Year Follow-Up (KUPOL Study). *J. Adolesc. Health* **2021**, *68*, 953–960. [CrossRef]
91. Abi-Jaoude, E.; Naylor, K.T.; Pignatiello, A. Smartphones, Social Media Use and Youth Mental Health. *CMAJ* **2020**, *192*, E136–E141. [CrossRef] [PubMed]
92. Berryman, C.; Ferguson, C.J.; Negy, C. Social Media Use and Mental Health among Young Adults. *Psychiatr. Q.* **2018**, *89*, 307–314. [CrossRef]
93. Hoge, E.; Bickham, D.; Cantor, J. Digital Media, Anxiety, and Depression in Children. *Pediatrics* **2017**, *140*, S76–S80. [CrossRef] [PubMed]
94. O'Reilly, M. Social Media and Adolescent Mental Health: The Good, the Bad and the Ugly. *J. Ment. Health* **2020**, *29*, 200–206. [CrossRef]
95. Frith, E. *Social Media and Children's Mental Health: A Review of the Evidence*; Education Policy Institute: London, UK, 2017.
96. Chancellor, S.; de Choudhury, M. Methods in Predictive Techniques for Mental Health Status on Social Media: A Critical Review. *NPJ Digit. Med.* **2020**, *3*, 43. [CrossRef]
97. Coates, A.E.; Hardman, C.A.; Halford, J.C.G.; Christiansen, P.; Boyland, E.J. Social Media Influencer Marketing and Children's Food Intake: A Randomized Trial. *Pediatrics* **2019**, *143*, e20182554. [CrossRef] [PubMed]

98. Janavi, E.; Soleimani, M.; Gholampour, A.; Friedrichsen, M.; Ebrahimi, P. Effect of Social Media Adoption and Media Needs on Online Purchase Behavior: The Moderator Roles of Media Type, Gender, Age. *J. Inf. Technol. Manag.* **2021**, *13*, 1–24. [CrossRef]
99. Potvin Kent, M.; Pauzé, E.; Roy, E.A.; de Billy, N.; Czoli, C. Children and Adolescents' Exposure to Food and Beverage Marketing in Social Media Apps. *Pediatr. Obes.* **2019**, *14*, e12508. [CrossRef]
100. de Veirman, M.; Hudders, L.; Nelson, M.R. What Is Influencer Marketing and How Does It Target Children? A Review and Direction for Future Research. *Front. Psychol.* **2019**, *10*, 2685. [CrossRef]
101. Bragg, M.A.; Pageot, Y.K.; Amico, A.; Miller, A.N.; Gasbarre, A.; Rummo, P.E.; Elbel, B. Fast Food, Beverage, and Snack Brands on Social Media in the United States: An Examination of Marketing Techniques Utilized in 2000 Brand Posts. *Pediatr. Obes.* **2020**, *15*, e12606. [CrossRef]
102. Boerman, S.C.; van Reijmersdal, E.A. Disclosing Influencer Marketing on YouTube to Children: The Moderating Role of Para-Social Relationship. *Front. Psychol.* **2020**, *10*, 3042. [CrossRef] [PubMed]
103. Fleming-Milici, F.; Harris, J.L. Adolescents' Engagement with Unhealthy Food and Beverage Brands on Social Media. *Appetite* **2020**, *146*, 104501. [CrossRef] [PubMed]
104. John, A.; Glendenning, A.C.; Marchant, A.; Montgomery, P.; Stewart, A.; Wood, S.; Lloyd, K.; Hawton, K. Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review. *J. Med. Internet Res.* **2018**, *20*, e129. [CrossRef]
105. Zhu, C.; Huang, S.; Evans, R.; Zhang, W. Cyberbullying Among Adolescents and Children: A Comprehensive Review of the Global Situation, Risk Factors, and Preventive Measures. *Front. Public Health* **2021**, *9*, 634909. [CrossRef]
106. Jadambaa, A.; Thomas, H.J.; Scott, J.G.; Graves, N.; Brain, D.; Pacella, R. Prevalence of Traditional Bullying and Cyberbullying among Children and Adolescents in Australia: A Systematic Review and Meta-Analysis. *Aust. New Zealand J. Psychiatry* **2019**, *53*, 878–888. [CrossRef]
107. McInroy, L.B.; Mishna, F. Cyberbullying on Online Gaming Platforms for Children and Youth. *Child Adolesc. Soc. Work J.* **2017**, *34*, 597–607. [CrossRef]
108. Thun, L.J.; Teh, P.L.; Cheng, C.B. CyberAid: Are Your Children Safe from Cyberbullying? *J. King Saud Univ.–Comput. Inf. Sci.* **2022**, *34*, 4099–4108. [CrossRef]
109. Mesch, G.S. Parent–Child Connections on Social Networking Sites and Cyberbullying. *Youth Soc.* **2018**, *50*, 1145–1162. [CrossRef]
110. Camerini, A.L.; Marciano, L.; Carrara, A.; Schulz, P.J. Cyberbullying Perpetration and Victimization among Children and Adolescents: A Systematic Review of Longitudinal Studies. *Telemat. Inform.* **2020**, *49*, 101362. [CrossRef]
111. Gómez-Ortiz, O.; Romera, E.M.; Ortega-Ruiz, R.; del Rey, R. Parenting Practices as Risk or Preventive Factors for Adolescent Involvement in Cyberbullying: Contribution of Children and Parent Gender. *Int. J. Environ. Res. Public Health* **2018**, *15*, 2664. [CrossRef]
112. Khajeheian, D.; Colabi, A.M.; Shah, N.B.A.K.; Radzi, C.W.J.B.W.M.; Jenatabadi, H.S. Effect of Social Media on Child Obesity: Application of Structural Equation Modeling with the Taguchi Method. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1343. [CrossRef]
113. Parks, E.P.; Moore, R.H.; Li, Z.; Bishop-Gilyard, C.T.; Garrett, A.R.; Hill, D.L.; Bruton, Y.P.; Sarwer, D.B. Assessing the Feasibility of a Social Media to Promote Weight Management Engagement in Adolescents with Severe Obesity: Pilot Study. *JMIR Res. Protoc.* **2018**, *7*, e8229. [CrossRef]
114. Radzi, C.W.J.M.; Jenatabadi, H.S.; Alanzi, A.R.A.; Mokhtar, M.I.; Mamat, M.Z.; Abdullah, N.A. Analysis of Obesity among Malaysian University Students: A Combination Study with the Application of Bayesian Structural Equation Modelling and Pearson Correlation. *Int. J. Environ. Res. Public Health* **2019**, *16*, 492. [CrossRef]
115. Mazur, A.; Caroli, M.; Radziewicz-Winnicki, I.; Nowicka, P.; Weghuber, D.; Neubauer, D.; Dembiński, Ł.; Crawley, F.P.; White, M.; Hadjipanayis, A. Reviewing and Addressing the Link between Mass Media and the Increase in Obesity among European Children: The European Academy of Paediatrics (EAP) and The European Childhood Obesity Group (ECOG) Consensus Statement. *Acta Paediatr. Int. J. Paediatr.* **2018**, *107*, 568–576. [CrossRef]
116. European Data Protection Law. Available online: https://fra.europa.eu/en/publication/2017/mapping-minimum-age-requirements-concerning-rights-child-eu/consent-use-data-children (accessed on 31 December 2022).
117. Children's Online Privacy Protection Act. Available online: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa (accessed on 31 December 2022).
118. Blumenau, J. Children's Media Regulations: A Report into State Provisions for the Protection and Promotion of Home-Grown Children's Media. *Save Kid's TV*, 30 March 2011.
119. Garlen, J.C.; Hembruff, S.L. Unboxing Childhood: Risk and Responsibility in the Age of YouTube. *J. Child. Stud.* **2021**. [CrossRef]
120. Charlop, M.H.; Schreibman, L.; Tryon, A.S. Learning through Observation: The Effects of Peer Modeling on Acquisition and Generalization in Autistic Children. *J. Abnorm. Child Psychol.* **1983**, *11*, 355–366. [CrossRef] [PubMed]
121. Fawcett, M.; Watson, D. *Learning through Child Observation*; Jessica Kingsley Publishers: London, UK, 2016; ISBN 1784501417.
122. Legare, C.H. The Development of Cumulative Cultural Learning. *Annu. Rev. Dev. Psychol.* **2019**, *1*, 119–147. [CrossRef]
123. Keles, B.; McCrae, N.; Grealish, A. A Systematic Review: The Influence of Social Media on Depression, Anxiety and Psychological Distress in Adolescents. *Int. J. Adolesc. Youth* **2020**, *25*, 1–15. [CrossRef]
124. Valakunde, N.; Ravikumar, S. Prediction of Addiction to Social Media. In Proceedings of the 2019 3rd IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT 2019, Coimbatore, Tamil Nadu, India, 20–22 February 2019.

125. Budzinski, O.; Gaenssle, S.; Lindstädt-Dreusicke, N. The Battle of YouTube, TV and Netflix: An Empirical Analysis of Competition in Audiovisual Media Markets. *SN Bus. Econ.* **2021**, *1*, 116. [CrossRef]
126. Alrehaili, A.; Alsaeedi, A.; Yafooz, W. Sentiment Analysis on YouTube Videos for Kids: Review. In Proceedings of the 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2021, Noida, India, 7–9 September 2021.
127. Hussain, M.N.; Tokdemir, S.; Agarwal, N.; Al-Khateeb, S. Analyzing Disinformation and Crowd Manipulation Tactics on Youtube. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018, Virtual, 10–13 November 2018.
128. Baghel, N.; Kumar, Y.; Nanda, P.; Shah, R.R.; Mahata, D.; Zimmermann, R. Kiki Kills: Identifying Dangerous Challenge Videos from Social Media. *arXiv* **2018**, arXiv:1812.00399.
129. Le, T.; Huang, D.Y.; Apthorpe, N.; Tian, Y. Skillbot: Identifying Risky Content for Children in Alexa Skills. *ACM Trans. Internet Technol. (TOIT)* **2022**, *22*, 1–31. [CrossRef]
130. Yafooz, W.M.S.; Al-Dhaqm, A.; Alsaeedi, A. Detecting Kids Cyberbullying Using Transfer Learning Approach: Transformer Fine-Tuning Models. In *Kids Cybersecurity Using Computational Intelligence Techniques*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 255–267.
131. Han, W.; Ansingkar, M. Discovery of Elsagate: Detection of Sparse Inappropriate Content from Kids Videos. In Proceedings of the 2020 Zooming Innovation in Consumer Technologies Conference, ZINC 2020, Novi Sad, Serbia, 26–27 May 2020.
132. Sjöbergh, J.; Araki, K. A Multi-Lingual Dictionary of Dirty Words. In Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008, Marrakech, Morocco, 26 May–1 June 2008.
133. Jevremovic, A.; Veinovic, M.; Cabarkapa, M.; Krstic, M.; Chorbev, I.; Dimitrovski, I.; Garcia, N.; Pombo, N.; Stojmenovic, M. Keeping Children Safe Online with Limited Resources: Analyzing What Is Seen and Heard. *IEEE Access* **2021**, *9*, 32723–132732. [CrossRef]
134. Halevy, A.; Canton-Ferrer, C.; Ma, H.; Ozertem, U.; Pantel, P.; Saeidi, M.; Silvestri, F.; Stoyanov, V. Preserving Integrity in Online Social Networks. *Commun. ACM* **2022**, *65*, 92–98. [CrossRef]
135. Kumar, A.; Sachdeva, N. Multi-Input Integrative Learning Using Deep Neural Networks and Transfer Learning for Cyberbullying Detection in Real-Time Code-Mix Data. *Multimedia Syst.* **2020**, *28*, 2027–2041. [CrossRef]
136. Alberto, T.C.; Lochter, J.V.; Almeida, T.A. TubeSpam: Comment Spam Filtering on YouTube. In Proceedings of the Proceedings–2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015, Miami, FL, USA, 9–11 December 2015.
137. Mironov, V.V.; Gusarenko, A.S.; Yusupova, N.I. Monitoring YouTube Video Views in the Educational Environment Based on Situation-Oriented Database and RESTful Web Services. *Системная инженерия и информационные технологии* **2021**, *3*, 39–49.